Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών
Σχολή Θετικών Επιστημών
Τμήμα Βιολογίας
Τομέας Βιοχημείας και Μοριακής Βιολογίας

Συστημική Αποκωδικοποίηση του ανθρώπινου Uniquome: Εξελικτικές, Μηχανιστικές και Θεραπευτικές Προσεγγίσεις

Διδακτορική Διατριβή

Ευάγγελος Μ. Κοντοπόδης
Μηχανικός Η/Υ και Πληροφορικής

ΑΘΗΝΑ 2023

Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών
Σχολή Θετικών Επιστημών
Τμήμα Βιολογίας
Τομέας Βιοχημείας και Μοριακής Βιολογίας

Συστημική Αποκωδικοποίηση του ανθρώπινου Uniquome: Εξελικτικές, Μηχανιστικές και Θεραπευτικές Προσεγγίσεις

Διδακτορική Διατριβή

Ευάγγελος Μ. Κοντοπόδης
Μηχανικός Η/Υ και Πληροφορικής

ΑΘΗΝΑ 2023

«Η έγκριση της Διδακτορικής Διατριβής από το Τμήμα Βιολογίας της Σχολής Θετικών Επιστημών του ΕΚΠΑ δεν υποδηλώνει αποδοχή των απόψεων του συγγραφέα »
(ν. 5343/1932, άρθρο 202)

«Το κείμενο της Διδακτορικής Διατριβής δεν αποτελεί προϊόν λογοκλοπής.»

Μέλη της Τριμελούς Συμβουλευτικής Επιτροπής

Επιβλέπων καθηγητής

Κωνσταντίνος Βοργιάς	Καθηγητής, Τμήμα Βιολογίας, ΕΚΠΑ

Μέλη

Ισιδώρα Παπασιδέρη	Ομότιμη Καθηγήτρια, Τμήμα Βιολογίας, ΕΚΠΑ
Δημήτριος Στραβοπόδης	Αναπληρωτής Καθηγητής, Τμήμα Βιολογίας, ΕΚΠΑ

Μέλη της Επταμελούς Εξεταστικής Επιτροπής

Επιβλέπων καθηγητής

Κωνσταντίνος Βοργιάς	Καθηγητής, Τμήμα Βιολογίας, ΕΚΠΑ

Μέλη

Ισιδώρα Παπασιδέρη	Ομότιμη Καθηγήτρια, Τμήμα Βιολογίας, ΕΚΠΑ
Δημήτριος Στραβοπόδης	Αναπληρωτής Καθηγητής, Τμήμα Βιολογίας, ΕΚΠΑ
Ουρανία Τσιτσιλώνη	Καθηγήτρια, Τμήμα Βιολογίας ΕΚΠΑ
Γεώργιος Παυλόπουλος	Ερευνητής Β', Ινστιτούτο Α. Φλέμινγκ
Ηλίας Ηλιόπουλος	Καθηγητής, Τμήμα Βιοτεχνολογίας Γεωπονικό Πανεπιστήμιο Αθηνών
Αγγελική Κατσαφάδου	Επίκουρη Καθηγήτρια, Τμήμα Δημόσιας και Ενιαίας Υγείας, Πανεπιστήμιο Θεσσαλίας

Περίληψη

Η Πρωτεωμική είναι ένα σύνολο πολύπλοκων μεθόδων και τεχνολογιών που αποσκοπεί στην ταυτοποίηση, καταγραφή και μελέτη του ολικού πρωτεϊνικού περιεχομένου ενός βιολογικού υλικού. Περιλαμβάνει το διαχωρισμό των πρωτεϊνών ενός βιολογικού δείγματος, την ανάλυση τους με φασματομετρία μάζας, την ταυτοποίησή τους με τη χρήση εργαλείων βιοπληροφορικής, τη συστηματική εισαγωγή των αποτελεσμάτων σε βάσεις δεδομένων και, τέλος την επεξεργασία τους. Οι πλέον εύχρηστες μέθοδοι για την ταυτοποίηση των πρωτεϊνών είναι αυτές που αξιοποιούν το πεπτιδικό αποτύπωμά τους (peptide finger-print) και αναλύουν την αμινοξική αλληλουχία των πεπτιδίων τους. Τα σημαντικότερα μειονεκτήματα αυτών των μεθόδων είναι πως για την ασφαλή ταυτοποίηση μίας πρωτεΐνης, απαιτείται η ανάλυση τουλάχιστον δύο πεπτιδίων ανά πρωτεΐνη καθώς και ότι πολλά από τα πεπτίδια που ταυτοποιούνται από το φασματογράφο μάζας δεν οδηγούν τελικά σε ασφαλή χαρακτηρισμό μίας πρωτεΐνης και απορρίπτονται κατά τη βιοπληροφορική επεξεργασία. Οι παραπάνω αδυναμίες των ήδη υπαρχόντων μεθόδων, οδήγησαν στην ανάγκη ανάπτυξης μιας νέας προσέγγισης για την ταυτοποίηση των πρωτεϊνών ενός οργανισμού. Η προσέγγιση αυτή βασίστηκε στην υπόθεση ότι η αμινοξική αλληλουχία κάθε πρωτεΐνης θα πρέπει να περιλαμβάνει τουλάχιστον ένα πεπτίδιο που η αμινοξική του αλληλουχία είναι απόλυτα μοναδική (Unique) ως προς το πρωτέωμα του οργανισμού που ανήκει, με αποτέλεσμα να χαρακτηρίζει την πρωτεΐνη διαφορικά και μονοσήμαντα. Έτσι, σαν αποτέλεσμα αυτής της προσέγγισης, στην παρούσα διατριβή καταγράφηκαν τα μοναδικά πεπτίδια του συνόλου των θεωρημένων (reviewed) πρωτεϊνών του ανθρώπου και εντός αυτών αναδείχθηκαν δύο νέες οντότητες μοναδικών πεπτιδίων, τα μοναδικά πεπτίδια ελαχίστου μήκους (core unique peptide - CrUP) και τα σύνθετα μοναδικά πεπτίδια (composite unique peptide - CmUP). Τέλος, εισήχθη για πρώτη φορά ο όρος του Uniquome που περιλαμβάνει το σύνολο των μοναδικών πεπτιδίων (CrUPs και CmUPs) ενός οργανισμού. Τα αντικείμενα της παρούσας διατριβής περιλαμβάνουν: **α)** Την ανάπτυξη μεθοδολογίας για την ανάλυση μεγάλων δεδομένων (big data analysis) με σκοπό την δημιουργία του ανθρώπινου Uniquome, **β)** την κατάρτιση και πλήρη καταγραφή του ανθρώπινου Uniquome που περιλαμβάνει τόσο τα CrUPs όσο και τα CmUPs, **γ)** την ανάλυση και την διερεύνηση των χαρακτηριστικών των μοναδικών πεπτιδίων σε ένα υψηλά συστημικό και συνθετικό επίπεδο και **δ)** την διερεύνηση εφαρμογών του ανθρώπινου Uniquome σε φυσιολογικές και παθολογικές καταστάσεις. Για την δημιουργία του ανθρώπινου Uniquome αναπτύχθηκε ένα νέο λογισμικό ανάλυσης που έχει την δυνατότητα να επεξεργαστεί μεγάλο όγκο δεδομένων (big data analysis) χρησιμοποιώντας κυρίως ως

γλώσσα προγραμματισμού τη C#, με παράλληλη χρήση μεθόδων που βασίζονται τόσο σε παράλληλα όσο και σε κατανεμημένα συστήματα.

Στο ανθρώπινο πρωτέωμα έως σήμερα έχουν περιληφθεί 20.430 θεωρημένες πρωτεΐνες, που περιλαμβάνουν 7.263.888 CrUPs και 77.697 CmUPs και απαρτίζουν το ανθρώπινο Uniquome, ενώ διαπιστώθηκε ότι 148 πρωτεΐνες (0,7%) δεν περιλαμβάνουν μοναδικά πεπτίδια καθώς φαίνεται να είναι ισομορφές με ομολογία μεγαλύτερη του 99%. Περαιτέρω ανάλυση των μοναδικών πεπτιδίων ως προς το μήκος τους, έδειξε ότι η πλειοψηφία των CrUPs και των CmUPs αποτελείται από πεπτίδια 6 και 11 αμινοξέων αντίστοιχα, ενώ η ανάλυση τους ως προς την σχετική θέση εμφάνισής τους μέσα στην πρωτεΐνη έδειξε πως τα CrUPs εντοπίζονται με το ίδιο ποσοστό σε όλες τις πιθανές θέσεις μέσα στις πρωτεΐνες, εν αντιθέσει με τα CmUPs που εντοπίζονται κυρίως στις αρχικές θέσεις των πρωτεϊνών. Η συνολική πυκνότητα από μοναδικά πεπτίδια για το ανθρώπινο πρωτέωμα υπολογίστηκε στο 64% για τα CrUPs, και στο 0,68% για τα CmUPs, ενώ η συνολική κάλυψη στο 93%. Αναφορικά με τον αριθμό από CrUPs που συνθέτουν ένα CmUP, η ομάδα των σύνθετων μοναδικών πεπτιδίων (6.103 πεπτίδια) που συνθέτονται από 5 μοναδικά πεπτίδια ελαχίστου μήκους είναι αυτή που εντοπίζεται με το μεγαλύτερο ποσοστό (7,85%). Η ανάλυση των χρωμοσωμάτων ως προς τα μοναδικά πεπτίδια που εμπεριέχονται σε αυτά, ανέδειξε πως χρωμοσώματα που εντοπίζονται με χαμηλά χαρακτηριστικά μοναδικότητας ενοχοποιούνται για χρωμοσωμικές ανωμαλίες που έχουν καταγραφεί στον άνθρωπο. Για την καλύτερη κατανόηση του ανθρώπινου Uniquome και των χαρακτηριστικών του, η μελέτη επεκτάθηκε στην κατάρτιση του Uniquome άλλων 19 πρότυπων οργανισμών που χρησιμοποιούνται σαν βιολογικά μοντέλα. Περεταίρω, αναφορικά με την εφαρμογή του Uniquome για την κατανόηση της βιολογικής δράσης του, αναλύθηκαν διάφορες οικογένειες πρωτεϊνών όπως η οικογένεια RAS, η οικογένεια Major histocompatibility complex class I (MHC I), η οικογένεια Peptidase C19 και η οικογένεια Peptidase S1. Τέλος, δύο ομάδες πεπτιδίων με ιδιαίτερη βιολογική σημασία σε ανθρώπινες παθήσεις είναι τα ανοσοπεπτίδια και τα αντιγονικά καρκινικά πεπτίδια. Διαπιστώθηκε ότι από τα υπάρχοντα ανοσοπεπτίδια το 87% είναι unique πεπτίδια, ενώ το 89% των υπαρχόντων αντιγονικών πεπτιδίων είναι επίσης unique.

Η κατάρτιση και η ανάλυση του ανθρώπινου Uniquome οδήγησε για πρώτη φορά στην αποκάλυψη δύο νέων οντοτήτων πεπτιδίων στο ανθρώπινο πρωτέωμα τα οποία όπως διαπιστώθηκε έχουν τεράστια βιολογική σημασία. Η ένταξη των μοναδικών πεπτιδίων στις ήδη υπάρχουσες εφαρμογές της φασματομετρίας μάζας μπορεί να αυξήσει σημαντικά τα ποσοστά ταυτοποίησης πρωτεϊνών στα υπό μελέτη δείγματα και να αποκαλύψει νέες πρωτεΐνες, καθόσον μια πρωτεΐνη δύναται να ταυτοποιηθεί από ένα και μόνο πεπτίδιο. Έτσι η χρήση των CrUPs και CmUPs θα οδηγήσει στην

αποτελεσματική, ασφαλή και ταχεία ταυτοποίηση πρωτεϊνικών βιοδεικτών παθολογικών καταστάσεων. Επιπλέον, από τα ευρήματα της παρούσας διατριβής διαπιστώνεται ότι η χρήση του Uniquome στην αντιμετώπιση παθολογικών καταστάσεων είναι δυνατόν να οδηγήσει στον σχεδιασμό νέων και πιο εξατομικευμένων θεραπευτικών προσεγγίσεων τόσο σε επίπεδο φαρμάκων όσο και σε επίπεδο εμβολίων.

Abstract

Proteomics are comprised of a setoff complex methods and technologies that aim to identify, register, and study the total protein content of a biological sample. It includes protein separation, mass spectrometry analysis, protein identification using bioinformatics tools, systematic introduction of the results in databases and analysis of the results. The most easy-to-use methods for protein identification are those that utilize the peptide fingerprint and analyze the amino acid sequence of their peptides. The biggest disadvantages of these methods are that they require the analysis of at least two peptides for every protein to allow for safe identification as well as that many of the peptides that are identified by mass spectrometry do not eventually lead to safe identification of a protein and are rejected throughout the bioinformatics process. The weaknesses mentioned above lead to the need to develop a new approach to identify the proteins of an organism. This approach was based on the hypothesis that every protein's amino acid sequence must include at least one peptide with an entirely unique amino acid sequence in a given organism. As a result, it would characterize this protein and distinguish it from all other proteins of this specific organism. Thus, as a result of this approach, in the present study, all the unique peptides of the total of review proteins of human are registered and two new entities of unique peptides emerge: core unique peptide (CrUP) and composite unique peptide (CmUP). Finally, the term "Uniquome" is introduced for the first time, a term that includes the ensemble of unique peptides (core and composite) of an organism. Objects of the present study include **a.** the development of a method to analyze big data for the creation of the human Uniquome, **b.** the setting-up of a full registration of the human Uniquome that includes CrUPs as well as CmUP, **c.** the analysis and expansion of their characteristics to a high systemic and synthetic level and **d.** the translation of the human Uniquome applications to physiologic and pathologic conditions. For the creation of the human Uniquome, a new analysis software was developed that is capable of big data analysis, using mainly C# and using methods that are based in parallel as well as distributed computing systems.

There are 20.430 reviewed proteins in the human proteome so far, that include 7.263.888 CrUPs and 77.697 CmUPs and comprise the human Uniquome, while 148 proteins (0.7%) were found not to include unique peptides because they are isoforms with over 99% homology. Further analysis of unique peptides length showed that the majority of CrUPs and CmUPs are comprised of 6 and 11 amino acids respectively while in-protein location analysis showed that CrUPs are found in all possible locations within a protein in contrast to CmUPs that are mostly found in the beginning of the

protein. The total density of unique peptides in the human proteome was calculated to 64% for CrUPs and 0.68% for CmUPs, while their total coverage was 93%. Regarding the number of CrUPs that comprise a CmUP, the group of composite peptides (6.103 peptides) that are made of 5 unique peptides of minimum length are the majority with a percentage of 7.85%. Analysis of unique peptides in chromosomes showed that chromosomes with low characteristics of uniqueness are found in chromosomal abnormalities in humans. To better understand Uniquome and its characteristics, this study went further to the creation of the Uniquome in another 19 model organisms used in research. Furthermore, several protein families were analyzed, such as the Ras protein family, proteins of the Major histocompatibility Complex class I (MHC-I), the family of Peptidase C19 and the family of Peptidase S1. Finally, two groups of peptides with biological significance in human disease are immune peptides and cancer antigenic peptides. We found that 87% of the existing immune peptides are unique peptides and 89% of cancer antigenic peptides are also unique peptides.

Creating and analyzing the human Uniquome led for the first time in the introduction of two new peptide entities of the human proteome, that are of paramount biological significance. Integrating unique peptides in the current applications of mass spectrometry can dramatically increase accurate protein identification and reveal new proteins, since each protein can be uniquely identified by one and only peptide. The use of CrUPs and CmUPs will result in effective, safe and fast identification of protein biomarkers in pathologic conditions. Furthermore, the findings in the present study show that using Uniquome in the treatment of pathologic conditions can lead to the design of new, personalized therapeutic approaches in the context of drug or vaccine development.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω ιδιαίτερα τα μέλη της τριμελούς επιτροπής του Τμήματος Βιολογίας του Εθνικού και Καποδιστριακού Πανεπιστημίου Αθηνών: τον καθηγητή κ. Κωνσταντίνο Βοργιά, την ομότιμη καθηγήτρια κ. Ισιδώρα Παπασιδέρη καθώς και τον αναπληρωτή καθηγητή κ. Δημήτρη Στραβοπόδη για τη στενή τους συνεργασία, τη μεγάλη τους υπομονή καθώς και τη στήριξή τους κατά τη διάρκεια της εκπόνησης της συγκεκριμένης Διδακτορικής Διατριβής.

Ένα ξεχωριστό ευχαριστώ θέλω να το αποδώσω στον κ. Δημήτρη Στραβοπόδη, με τον οποίον είχα την τιμή, την τύχη και την ευτυχία να συνεργαστώ, για τις γόνιμες επιστημονικές συμβουλές που μου προσέφερε και συντέλεσε στο μέγιστο στην ολοκλήρωση της παρούσας Διδακτορικής Διατριβής.

Ένα μεγάλο ευχαριστώ στον κ. Γιώργο Τσάγκαρη, Ειδικό Λειτουργικό Επιστήμονα Α' Βαθμίδας, τους Ιδρύματος Ιατροβιολογικών Ερευνών της Ακαδημίας Αθηνών, για την συνεχή προσφορά του και το αληθινό ενδιαφέρον του καθ' όλη τη διάρκεια της Διδακτορικής μου Διατριβής. Για όλες τις στιγμές που βοήθησε, αποδέχθηκε, προσπάθησε, συμφώνησε και στάθηκε σε μικρότερα ή μεγαλύτερα θέματα που προέκυψαν.

Ευχαριστώ θερμά τα υπόλοιπα μέλη της Επταμελούς Εξεταστικής Επιτροπής, την κ. Ουρανία Τσιτσιλώνη, Καθηγήτρια του τμήματος Βιολογίας του Εθνικού και Καποδιστριακού Πανεπιστημίου Αθηνών, τον Ερευνητή Β' του Ινστιτούτου Α. Φλέμινγκ, κ. Γεώργιο Παυλόπουλο, τον Καθηγητή του τμήματος Βιοτεχνολογίας, του Γεωπονικού πανεπιστημίου Αθηνών, κ. Ηλία Ηλιόπουλο και την κ. Αγγελική Κατσαφάδου, Επίκουρη καθηγήτρια του τμήματος Δημόσιας και Ενιαίας Υγείας, του Πανεπιστημίου Θεσσαλίας, για την τιμή που μου έκαναν να συμμετέχουν στην Επταμελή Εξεταστική Επιτροπή, καθώς και για την πολύτιμη και απόλυτη στήριξή τους.

Ιδιαίτερα ευχαριστώ τον υποψήφιο διδάκτορα Πιέρρο Βασίλη για την εξαιρετική μας συνεργασία, σε όλα αυτά τα χρόνια της φιλίας μας.

Περιεχόμενα

Κατάλογος Πινάκων

Κατάλογος Εικόνων

Συντομογραφίες

Ala	Αλανίνη
API	Application Programmer Interface
Arg	Αργινίνη
ASCII	American Standard Code for Information Interchange
Asn	Ασπαραγίνη
Asp	Ασπαρτικό οξύ
AVX	Advanced Vector eXtensions
CmUP	Composite Unique Peptide, Σύνθετα μοναδικά πεπτίδια
CPU	Central Processing Unit
CrUP	Core Unique Peptide, Μοναδικά πεπτίδια ελαχίστου μήκους
CUDA	Compute Unified Device Architecture
Cys	Κυστεΐνη
DNA	Deoxyribonucleic acid, Δεοξυ-ριβο(ζο)νουκλεϊ(νι)κό οξύ
GB	Gigabyte
Gln	Γλουταμίνη
Glu	Γλουταμινικό οξύ
Gly	Γλυκίνη
GPCR	G-protein coupled receptor 1 Family
His	Ιστιδίνη
HTTP	Hyper Text Transfer Protocol
Ile	Ισολευκίνη
kDa	Kilo Dalton
Leu	Λευκίνη
Lys	Λυσίνη
Met	Μεθειονίνη
MHC	Major histocompatibility complex
MS	Mass spectrometry
Phe	Φαινυλαλανίνη
Pro	Προλίνη
Pyr	Πυρρολυσίνη
RAM	Random Access Memory
RDBMS	Relational Database Management Systems
REST	Representational State Transfer
RNA	Ribonucleic acid, Ριβο(ζο)νουκλεϊ(νι)κό οξύ
Sec	Σεληνοκυστεΐνη
Ser	Σερίνη
SIMD	Single Instruction Multiple Dataset
SoC	System on a Chip
SSD	Solid State Drive
TCP/IP	Transmission Control Protocol/ Internet Protocol
Thr	Θρεονίνη
Trp	Τρυπτοφάνη
Tyr	Τυροσίνη
Val	Βαλίνη
VPU	Virtual Processing Unit
UP	Unique Peptide (μοναδικό πεπτίδιο)

Ορισμοί

- **Μοναδικά Πεπτίδια (Unique Peptides)**: Είναι τα πεπτίδια των οποίων η αμινοξική αλληλουχία εμφανίζεται μόνο σε μία πρωτεΐνη.

- **Μοναδικά Πεπτίδια Ελαχίστου Μήκους (Core Unique Peptides, CrUPs)**: Είναι τα ελαχίστου μήκους πεπτίδια των οποίων η αμινοξική αλληλουχία εμφανίζεται σε μία μόνο πρωτεΐνη.

- **Σύνθετα Μοναδικά Πεπτίδια (Composite Unique Peptides, (CmUPs)**: Είναι τα πεπτίδια τα οποία συνθέτονται από την ένωση δύο ή περισσότερων μοναδικών πεπτιδίων ελαχίστου μήκους, όταν το ένα επικαλύπτει το άλλο.

- **Μοναδίωμα (Uniquome):** Είναι το σύνολο των μοναδικών πεπτιδίων ενός οργανισμού (CrUP και CmUP).

- **Πυκνότητα Μοναδικών Πεπτιδίων Ελαχίστου Μήκους (Density of Core Unique Peptides):** Είναι ο λόγος του συνολικού αριθμού Μοναδικών Πεπτιδίων Ελαχίστου Μήκους ενός οργανισμού προς το σύνολο των αμινοξέων των πρωτεϊνών του οργανισμού.

- **Πυκνότητα Σύνθετων Μοναδικών Πεπτιδίων (Density of Composite Unique Peptides):** Είναι ο λόγος του συνολικού αριθμού Σύνθετων Μοναδικών Πεπτιδίων ενός οργανισμού προς το σύνολο των αμινοξέων των πρωτεϊνών του οργανισμού.

- **Μοναδική Κάλυψη (Unique Coverage):** Είναι ο λόγος του συνολικού αριθμού των αμινοξέων ενός οργανισμού που εμπεριέχονται έστω μία φορά στον σχηματισμό μοναδικών πεπτιδίων προς το σύνολο των αμινοξέων των πρωτεϊνών του οργανισμού.

Στο παρακάτω παράδειγμα παρουσιάζεται ο τρόπος με τον οποίο τα CrUPs πεπτίδια συνθέτουν ένα CmUP.

Το 16πεπτίδιο

..CMIVEFSRYLSQMRNL..

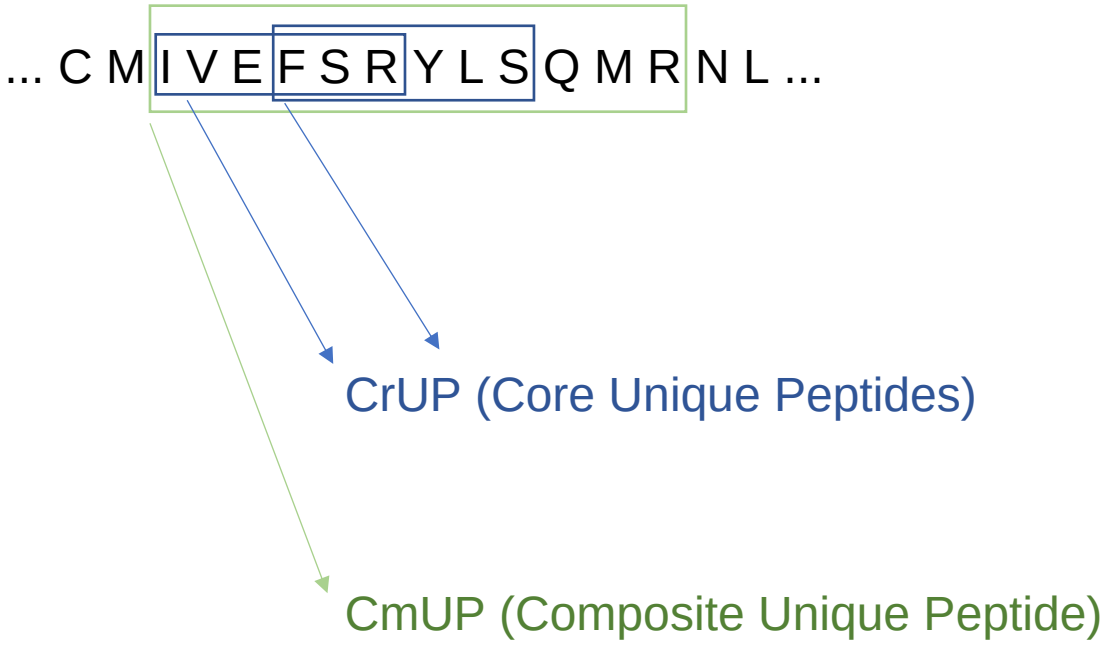
Αποτελείται από τα μοναδικά πεπτίδια ελαχίστου μήκους (CrUP):

- IVEFSR
- VEFSRY
- FSRYLS
- SRYLSQMR

Τα οποία συνθέτουν το σύνθετο μοναδικό πεπτίδιο:

- IVEFSRYLSQMR

Για το 16πεπτίδιο του παραδείγματος έχουμε:

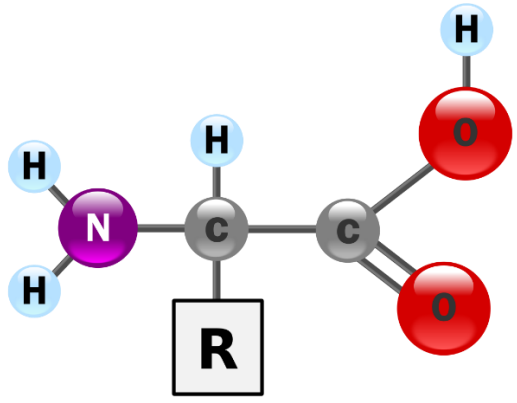


1. Εισαγωγή

1.1 Αμινοξέα

Τα αμινοξέα είναι μικρά οργανικά μόρια τα οποία περιέχουν μια αμινομάδα (-NH2), μια καρβοξυλομάδα (-COOH) καθώς και μια πλευρική αλυσίδα [R] που διαφοροποιεί κάθε αμινοξύ από τα υπόλοιπα. Τα βασικά στοιχεία των αμινοξέων είναι ο άνθρακας (C), το υδρογόνο (H) το οξυγόνο (O) και το άζωτο (N) αν και σε κάποια αμινοξέα μπορούν να βρεθούν και άλλα στοιχεία στην πλευρική τους αλυσίδα [1].

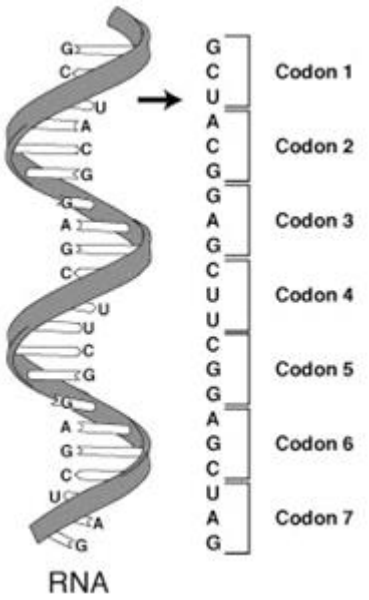
Σήμερα γνωρίζουμε περίπου 500 αμινοξέα τα οποία μπορούν να σχηματιστούν με φυσικό τρόπο. Από αυτά μόνο μια μικρή μειοψηφία έχει ιδιαίτερο βάρος για την βιοχημεία. Τα συγκεκριμένα αμινοξέα έχουν ένα κεντρικό άτομο άνθρακα ενωμένο με μια αμινομάδα, μια καρβοξυλομάδα, ένα άτομο υδρογόνου και μια πλευρική αλυσίδα (-R) (Εικόνα 1). Ονομάζονται α-αμινοξέα και αποτελούν τις δομικές μονάδες των πρωτεϊνών, γι' αυτό και καλούνται πρωτεϊνογενετικά (proteinogenic). Το όνομά τους (άλφα) προέρχεται από ένα παλαιότερο σύστημα ονοματολογίας της οργανικής χημείας όπου τα άτομα μιας υδρογονανθρακικής αλυσίδας που είναι συνδεδεμένη ομοιοπολικά με μια καρβοξυλομάδα, συμβολίζονται με γράμματα του ελληνικού αλφαβήτου (α, β, ...).



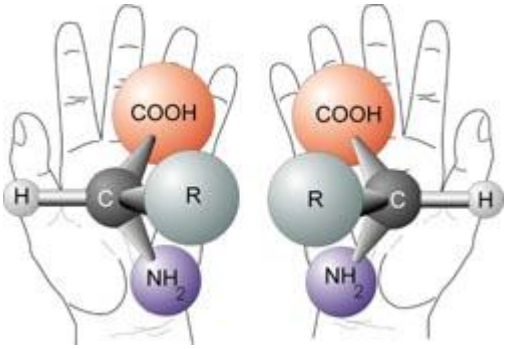
Εικόνα 1 Γενικός συντακτικός τύπος α-αμινοξέων

Υπάρχουν 22 αμινοξέα από τα οποία μόνο τα 20 μπορούν να εξαχθούν κατευθείαν από τριπλέτες κωδικονίων του γενετικού κώδικα (Εικόνα 2 και Πίνακας 2). Τα 20 αυτά αμινοξέα συμβολίζονται με 1 ή με 3 γράμματα (Πίνακας 1). Με εξαίρεση την γλυκίνη, όλα τα αμινοξέα έχουν τέσσερα διαφορετικά είδη λειτουργικών ομάδων συνδεδεμένα στο άτομο άνθρακα. Αυτές μπορούν να διαταχθούν στον χώρο με δύο διαφορετικούς τρόπους σχηματίζοντας δύο διαφορετικά στερεοϊσομερή τα D-(Dextrorotatory) και L-(levorotatory)

που είναι μεταξύ τους κατοπτρικά είδωλα (χειρόμορφα). Από τις δύο αυτές μορφές μόνο η L- χρησιμοποιείται από τα κύτταρα για την παραγωγή των πρωτεϊνών (Εικόνα 3). Η γλυκίνη είναι το μόνο αμινοξύ που δεν είναι χειρόμορφο αφού δεν διαθέτει κατοπτρικό είδωλο [2].



Εικόνα 2 Μετάφραση κωδικονίων σε Αμινοξέα



Εικόνα 3 Κατοπτρική μορφή D- / L- αμινοξέων

Αμινοξύ	Χημικός Τύπος	κωδικός ενός γράμματος	Κωδικός τριών Γραμμάτων	Μονοϊστοτροπική μάζα
Αλανίνη	C3H5ON	A	Ala	71,03711
Αργινίνη	C6H12ON4	R	Arg	156,10111
Ασπαραγίνη	C4H6O2N2	N	Asn	114,04293
Ασπαρτικό οξύ	C4H5O3N	D	Asp	115,02694
Κυστεΐνη	C3H5ONS	C	Cys	103,00919
Γλουταμινικό οξύ	C5H7O3N	E	Glu	129,04259
Γλουταμίνη	C5H8O2N2	Q	Gln	128,05858
Γλυκίνη	C2H3ON	G	Gly	57,02146
Ιστιδίνη	C6H7ON3	H	His	137,05891
Ισολευκίνη	C6H11ON	I	Ile	113,08406
Λευκίνη	C6H11ON	L	Leu	113,08406
Λυσίνη	C6H12ON2	K	Lys	128,09496
Μεθειονίνη	C5H9ONS	M	Met	131,4049
Φενυλαλανίνη	C9H9ON	F	Phe	147,06841
Προλίνη	C5H7ON	P	Pro	97,05276
Σερίνη	C3H5O2N	S	Ser	87,03203
Θρεονίνη	C4H7O2N	T	Thr	101,04768
Τρυπτοφάνη	C11H10ON2	W	Trp	186,07931
Τυροσίνη	C9H9O2N	Y	Tyr	163,06333
Βαλίνη	C5H9ONS	V	Val	99,06841

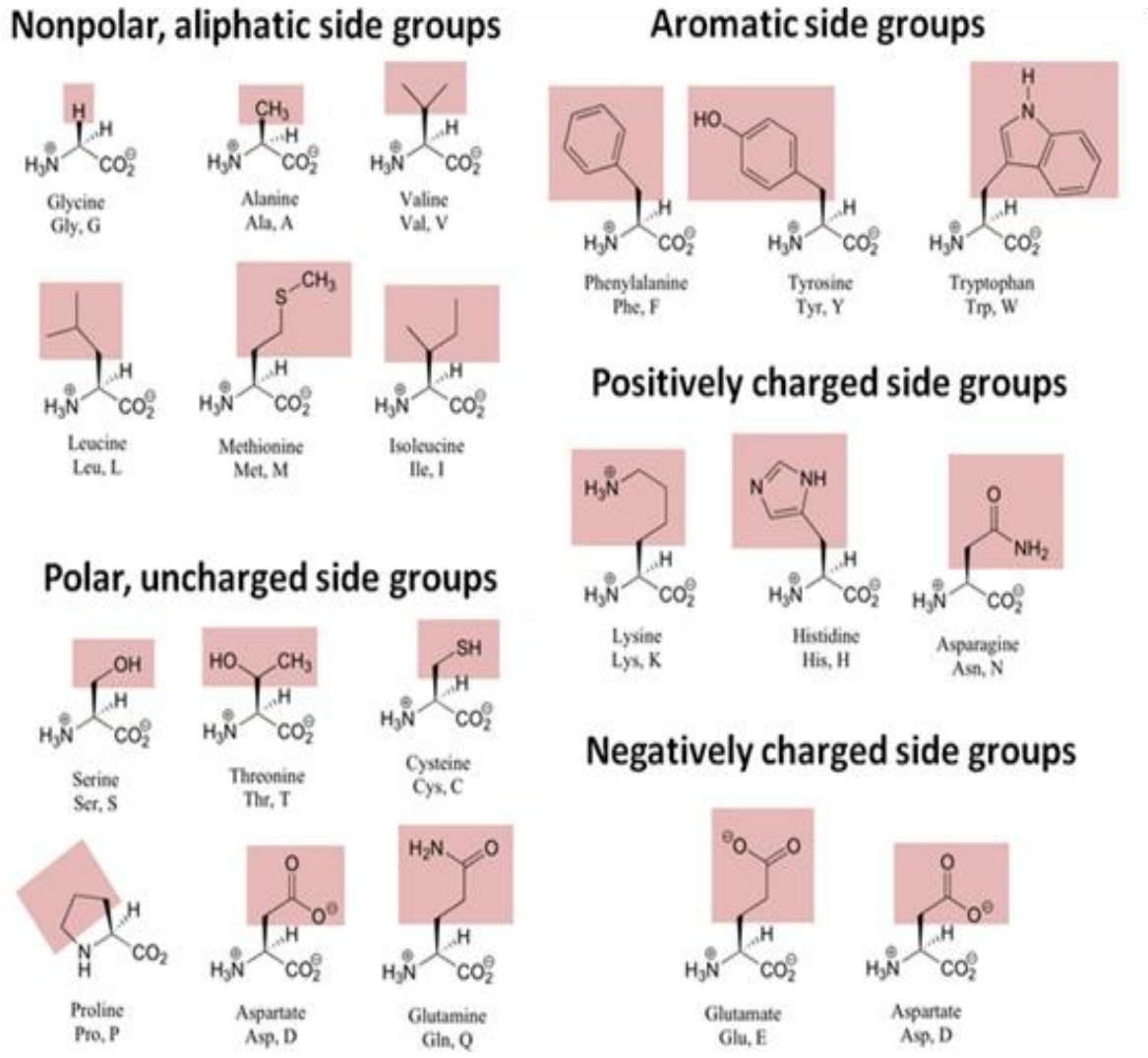
Πίνακας 1 Πίνακας Αμινοξέων

Πρώτη θέση	Δεύτερη θέση: U	Δεύτερη θέση: C	Δεύτερη θέση: A	Δεύτερη θέση: G	Τρίτη θέση
U	Phe	Ser	Tyr	Cys	UCAG
	Leu		*ochre*	*Opal/Sec*	
			Amber/Pyr	Trp	
C	Leu	Pro	His	Arg	UCAG
			Gln		
A	Ile	Thr	Asn	Ser	UCAG
	Met		Lys	Arg	
G	Val	Ala	Asp	Gly	UCAG
			Glu		

Πίνακας 2 Ο Καθολικός γενετικός κώδικας. Τα αμινοξέα ταυτοποιούνται από τον κώδικα τριών γραμμάτων. Μη συμπληρωματικά κωδικόνια ταυτοποιούνται με όνομα. Παρατηρείστε την διπλή λειτουργία των κωδικονίων UAG και UGA

Οι φυσικές και χημικές ιδιότητες κάθε αμινοξέος καθορίζονται από την χαρακτηριστική πλευρική του αλυσίδα (-R). Αμινοξέα με παρόμοιες πλευρικές αλυσίδες έχουν και παρόμοιες ιδιότητες λόγω της πολικότητας που αυτή τους προσδίδει. Ανάλογα με αυτές τις ιδιότητες τα 20 αμινοξέα διακρίνονται σε 4 ομάδες (Εικόνα 4) [3] :

1. Βασικό χαρακτήρα: Οι πλευρικές ομάδες είναι δέκτες πρωτονίων. Σε αυτή την ομάδα ανήκουν τα αμινοξέα αργινίνη (arginine), λυσίνη (lysine), ιστιδίνη (histidine).
2. Όξινο χαρακτήρα: Έχουν στην πλευρική ομάδα μια ομάδα καρβοξυλίου. Αυτή είναι μη πρωτονιωμένη και έχει αρνητικό φορτίο σε pH 7. Σε αυτή την ομάδα ανήκουν το ασπαρτικό οξύ (aspartic acid) και το γλουταμικό οξύ (glutamic acid).
3. Πολικό χαρακτήρα αλλά χωρίς φορτίο σε pH 7: Σε αυτή την ομάδα ανήκουν τα αμινοξέα ασπαριγίνη (asparagine), γλουταμίνη (glutamine), σερίνη (serine), θρεονίνη (threonine) και τυροσίνη (tyrosine).
4. Μη πολικό χαρακτήρα: Σε αυτή την ομάδα ανήκουν τα αμινοξέα γλυκίνη (glycine), αλανίνη (alanine), ισολευκίνη (isoleucine), λευκίνη (leucine), βαλίνη (valine), φαινυλαλανίνη (phenylalanine), τρυπτοφάνη (tryptophan), μεθειονίνη (methionine) και η προλίνη (proline).



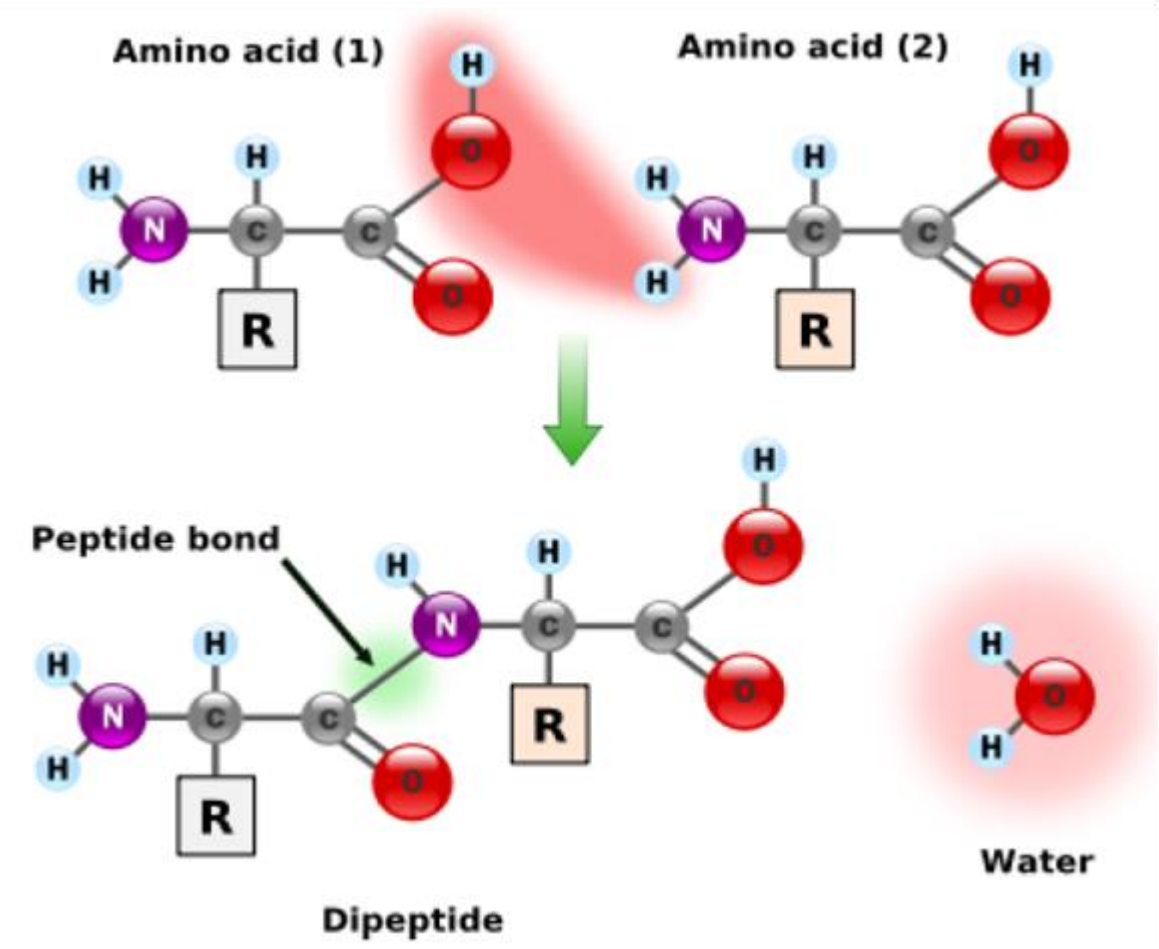
Εικόνα 4 Ομαδοποίηση αμινοξέων βάσει της πολικότητάς τους

Ένα σημαντικό χαρακτηριστικό με ιδιαίτερη σημασία για την κατανομή στον χώρο των αμινοξέων σε μια πεπτιδική αλυσίδα είναι η υδροφοβικότητά τους. Ο όρος υδροφοβικότητα χρησιμοποιείται για να προσδιορίσει την προδιάθεση ενός μορίου να απωθηθεί από τα περιβάλλοντα μόρια του νερού. Πιο συγκεκριμένα η υδροφοβικότητα προσδιορίζει την απουσία ελκτικών δυνάμεων των αμινοξέων. Τα υδροφοβικά μόρια τείνουν να μην παρουσιάζουν πολικότητα και γι' αυτό έχουν προτίμηση στα μη πολικά διαλύματα καθώς και στα ουδέτερα μόρια. Παραδείγματα υδροφοβικών μορίων είναι τα λίπη και τα έλαια. Η υδροφοβικότητα των αμινοξέων επηρεάζει την αναδίπλωση των πρωτεϊνών στον τρισδιάστατο χώρο, αφού κάποια αμινοξέα έλκονται από τα μόρια του νερού που βρίσκονται στον περιβάλλοντα χώρο αυτής οδηγώντας στην τελική της διάταξη [4-6].

1.2 Πεπτίδια

Τα αμινοξέα συνδέονται μεταξύ τους μέσω του πεπτιδικού δεσμού σχηματίζοντας μεγαλύτερες αλυσίδες, τα πεπτίδια (Εικόνα 5). Συγκεκριμένα, η σύνδεση αυτή επιτυγχάνεται μέσω της αντίδρασης της καρβοξυλομάδας ενός αμινοξέος με την αμινομάδα ενός άλλου. Αυτός ο δεσμός είναι ομοιοπολικός και κατά τον σχηματισμό του οδηγεί στην αποβολή ενός μορίου νερού [7].

Σε αντιστοιχία με τις πολυνουκλεοτιδικές αλυσίδες, οι πεπτιδικές αλυσίδες εμφανίζουν επίσης πολικότητα, αφού στο ένα άκρο της πεπτιδικής αλυσίδας υπάρχει μια ελεύθερη αμινομάδα (θετικά φορτισμένη) ενώ στο άλλο μια ελεύθερη καρβοξυλομάδα (αρνητικά φορτισμένη). Τα πεπτίδια χωρίζονται κατά σύμβαση ως προς το μέγεθος τους σε 2 ομάδες, τα ολιγοπεπτίδια που έχουν 50 ή λιγότερα αμινοξέα και τα πολυπεπτίδια με περισσότερα από 50 αμινοξέα. Το πλήθος των πιθανών πεπτιδίων όπως εύκολα γίνεται κατανοητό είναι απεριόριστο αφού μόνο για τα ολιγοπεπτίδια μήκους 50 έχουμε 20^{50} διαφορετικά πεπτίδια που μπορούν να σχηματιστούν [8].

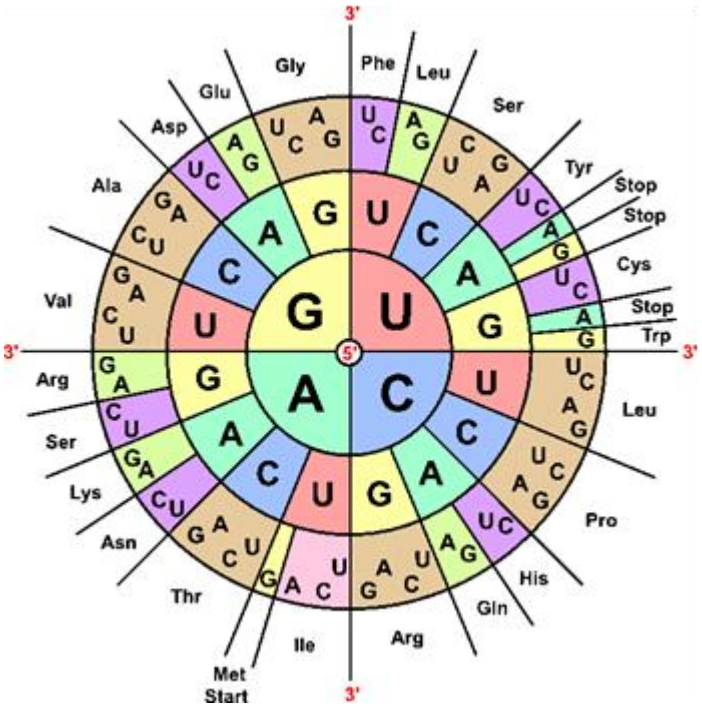


Εικόνα 5 Σχηματισμός διπεπτιδίου με πεπτιδικό δεσμό

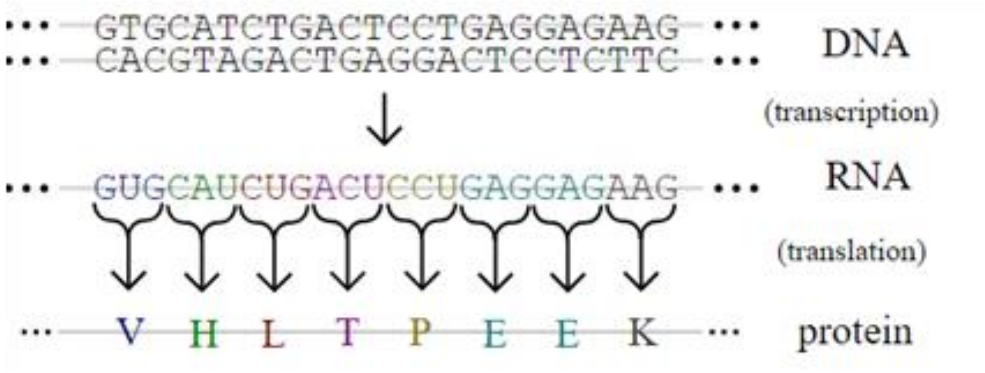
1.3 Πρωτεΐνες

Ο όρος πρωτεΐνη επινοήθηκε το 1838 από τον Σουηδό χημικό Jons Jacob Berzelius για να περιγράψει μια ιδιαίτερη κατηγορία μακρομορίων, πλούσια στους ζωντανούς οργανισμούς, που αποτελούνται από γραμμικές αλυσίδες αμινοξέων. Ο όρος προέρχεται από την ελληνική λέξη proteios που σημαίνει «της πρώτης τάξης» και επιλέχθηκε για να αποδώσει την κεντρική σημασία των πρωτεϊνών στον ανθρώπινο οργανισμό. Καθώς η γνώση μας για αυτή την κατηγορία των μακρομορίων όλο και αυξάνεται, ο ορισμός αυτός φαίνεται ότι είναι ο πιο κατάλληλος. Έχουμε ανακαλύψει ότι οι πρωτεΐνες είναι ζωτικής σημασίας συστατικά σχεδόν σε κάθε βιολογικό σύστημα και σε κάθε ζωντανό οργανισμό. Υπάρχουν χιλιάδες πρωτεΐνες ακόμα και στο απλούστερο τύπο κυττάρου και αποτελούν τη βάση της κάθε πιθανής βιολογικής λειτουργίας [9].

Οι πρωτεΐνες είναι μακρομόρια που αποτελούνται από ένα ή περισσότερα πολυπεπτίδια, το καθένα από τα οποία είναι μια γραμμική αλυσίδα αμινοξέων. Υπάρχουν 20 αμινοξέα που αναγνωρίζονται από τον γενετικό κώδικα και ακόμα άλλα δύο τροποποιημένα παράγωγα, σεληνοκυστεΐνη (Sec) και πυρρολυσίνη (Pyr) που ενσωματώνονται με τρόπο που εξαρτάται από το περιβάλλον (Πίνακας 2, Εικόνα 6). Οι πρωτεΐνες συντίθενται με βάση τις οδηγίες που είναι κωδικοποιημένες στο DNA κάθε οργανισμού μέσω της διαδικασίας της μεταγραφής / μετάφρασης (Εικόνα 7) [9 - 11].



Εικόνα 6 Γενετικός κώδικας / πίνακας αντιστοίχισης τριάδων βάσεων με αμινοξέα



Εικόνα 7 Παραγωγή πρωτεΐνης από DNA

Ανάλογα με τη λειτουργία τους, οι πρωτεΐνες, μπορούν να χωριστούν κυρίως σε δύο κατηγορίες: τις δομικές πρωτεΐνες και τις πρωτεΐνες με βιολογική δράση [12]. Οι δομικές πρωτεΐνες αποτελούν τα δομικά υλικά του κυττάρου, είναι ινώδεις πρωτεΐνες όπως η κερατίνη, το κολλαγόνο, η ελαστίνη κ.α. και υπάρχουν σε όλους τους ιστούς (τους μύες, τα οστά, το δέρμα, τα εσωτερικά όργανα, τις κυτταρικές μεμβράνες και τα ενδοκυτταρικά οργανίδια). Οι πρωτεΐνες με βιολογική δράση είναι τα ένζυμα (βιολογικοί καταλύτες με μεγάλη επιλεκτικότητα), οι ορμόνες (π.χ. ινσουλίνη) που ρυθμίζουν μεταβολικές αντιδράσεις, οι συσταλτικές πρωτεΐνες (π.χ. μυοσίνη), οι πρωτεΐνες μεταφοράς (π.χ. αιμοσφαιρίνη), οι πρωτεΐνες με προστατευτική δράση στο αίμα (π.χ. ανοσοσφαιρίνη), οι αποθηκευτικές πρωτεΐνες (π.χ. γλιαδίνη), οι τοξικές πρωτεΐνες (π.χ. τοξίνη), οι αντιβιοτικές πρωτεΐνες, τα αντιγόνα και οι αντιθρεπτικές πρωτεΐνες.

Δομές Πρωτεΐνης

Πρωτοταγής Δομή

Η αλληλουχία των αμινοξέων σε ένα πολυπεπτίδιο είναι γνωστή ως πρωτοταγής δομή. Τα πεπτίδια συνδέονται με πεπτιδικούς δεσμούς που συνήθως υιοθετούν την trans στερεοδιάταξη, έτσι ώστε το καρβονυλικό οξυγόνο και το αμιδικό υδρογόνο γειτονικών αμινοξέων να είναι απομακρυσμένα το ένα από το άλλο. Ο πεπτιδικός δεσμός είναι άκαμπτος αλλά οι άλλοι δεσμοί είναι αρκετά ευκίνητοι και επιτρέπουν στον σκελετό του πολυπεπτιδίου να διπλώνει στο χώρο [9,13].

Δευτεροταγής Δομή

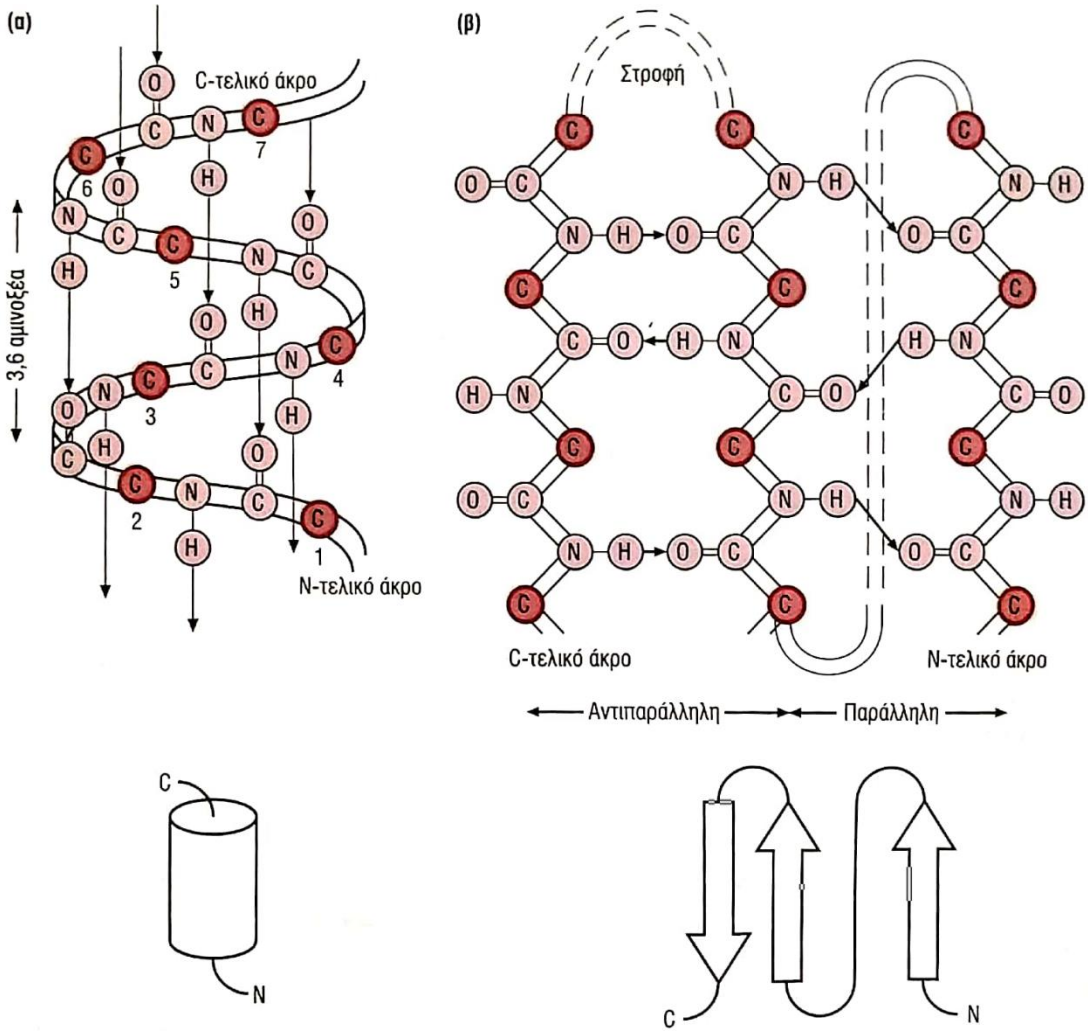
Οι δευτεροταγείς δομές των πρωτεϊνών είναι κανονικές επαναλαμβανόμενες τοπικές διαμορφώσεις που παράγονται από ενδομοριακούς δεσμούς υδρογόνου. Αυτές μερικές φορές περιέχουν πολικές πλευρικές ομάδες (όπως αυτές των αμινοξέων σερίνη και θρεονίνη), αλλά ο σκελετός του πολυπεπτιδίου πολλές φορές από μόνος του είναι πολικός

γιατί η NH αμινομάδα μπορεί να δρα ως δότης υδρογόνου, ενώ η C=0 καρβοξυλομάδα μπορεί να δρα ως δέκτης υδρογόνου. Η κανονική απόσταση των πεπτιδικών δεσμών σε όλη την πολυπεπτιδική αλυσίδα επιτρέπει τον σχηματισμό κανονικών δομών. Οι δύο πιο κοινές δομές είναι η α-έλικα και το β-πτυχωτό φύλλο, οι οποίες σχηματίζονται από ένα κανονικό σχήμα δεσμών υδρογόνου μεταξύ των πεπτιδικών ομάδων N-H και C=O εκείνων των αμινοξέων που είναι *κοντά το ένα στο άλλο στη γραμμική αλληλουχία* τους. Πολύ συχνά, οι δομές αυτές απεικονίζονται σαν οι μοναδικές δευτεροταγείς δομές πρωτεϊνών με διάσπαρτες μη δομημένες περιοχές γνωστές ως θηλιές (Εικόνα 8). Τυπικά οι α-έλικες είναι δεξιόστροφες και περιέχουν 4-40 κατάλοιπα αμινοξέων που αντιστοιχούν σε 1-12 στροφές της έλικας. Εμφανίζονται δε όταν δεσμοί υδρογόνου σχηματίζονται μεταξύ πεπτιδικών μονάδων τεσσάρων καταλοίπων από αμινοξέα μεταξύ τους, ευθυγραμμίζοντάς τα και δίδοντας σε ολόκληρη τη δομή μια σημαντική διπολική ροπή, αν και οι γωνίες των δεσμών είναι οξείες. Άλλες ελικοειδείς δευτεροταγείς δομές (η 3_{10}-έλικα και η π-έλικα) συναντώνται πιο σπάνια στο τέλος των πιο τυπικών α-ελίκων, εξαιτίας των περιορισμών του διπλώματος. Σε αντίθεση με τις έλικες, τα β-πτυχωτά φύλλα σχηματίζονται από περιοχές της πολυπεπτιδικής αλυσίδας όπου οι γωνίες των δεσμών είναι πλήρως εκτεταμένες (αυτές είναι γνωστές ως β-κλώνοι). Αρκετές β-πτυχωτές επιφάνειες μπορούν να ευθυγραμμιστούν παράλληλα, αντιπαράλληλα ή σε μικτές συστοιχίες σχηματίζοντας πληθώρα δεσμών υδρογόνου μεταξύ πεπτιδικών μονάδων γειτονικών β-κλώνων. Τόσο η α-έλικα όσο και οι β-πτυχωτές επιφάνειες μπορούν να ενωθούν μεταξύ τους με συνδετικές δομές οι οποίες υιοθετούν τις δικές τους δευτερογενείς διαμορφώσεις που μπορούν να οριστούν ως στροφές. Για παράδειγμα, σχηματίζεται μια β-στροφή όταν ένας δεσμός υδρογόνου σχηματίζεται μεταξύ πεπτιδικών μονάδων τριών καταλοίπων από αμινοξέα. Όπου δεν υπάρχουν δεσμοί υδρογόνου, οι περιοχές σύνδεσης είναι γνωστές ως βρόγχοι. Ο πυρήνας μιας πρωτεΐνης είναι συχνά πλούσιος σε δευτεροταγείς δομές, επειδή αυτό επιτρέπει ενεργειακά αποδοτικό πακετάρισμα, ενώ οι βρόγχοι γενικά βρίσκονται στην επιφάνεια όπου μπορούν να συμβούν αλληλεπιδράσεις με το διαλύτη. Οι βρόγχοι είναι γενικά πολύ πιο μεταλλάξιμοι από τις περιοχές του πυρήνα επειδή δεν παρεμβάλλονται στον τρόπο με τον οποίο η πρωτεΐνη συσκευάζεται και ενεργούν περισσότερο σαν «διακοσμήσεις» στην επιφάνεια της πρωτεΐνης για τον έλεγχο των αλληλεπιδράσεων [9,14].

Τριτοταγής Δομή

Η τριτοταγής δομή ή αναδιπλωμένη δομή ενός πολυπεπτιδίου είναι η τελική εικόνα του, αντανακλώντας τον τρόπο με τον οποίο οι δευτεροταγείς δομές και τα μοτίβα συνδυάζονται για να σχηματίσουν συγκροτημένες δομικές λειτουργικές περιοχές (domains). Μια δομική λειτουργική περιοχή μπορεί να θεωρηθεί ως τμήμα μια πολυπεπτιδικής αλυσίδας που μπορεί να αναδιπλώνεται ανεξάρτητα από την υπόλοιπη αλυσίδα, σε μια

σταθερή τριτοταγή δομή, οπότε οι δομικές λειτουργικές περιοχές είναι οι μονάδες πρωτεϊνικής λειτουργίας. Μια πρωτεΐνη μπορεί να περιέχει μια μόνο ή πολλαπλές δομικές λειτουργικές περιοχές και στην τελευταία περίπτωση οι διαφορετικές δομικές λειτουργικές περιοχές μπορούν να διεξάγουν επιμέρους λειτουργίες στα πλαίσια της συνολικής βιολογικής λειτουργίας της πρωτεΐνης [9,15].



Εικόνα 8 Κοινές δευτεροταγείς δομές σε πρωτεΐνες, (α) η α-έλικα, (β) η β-πτυχωτή επιφάνεια [9]

Τεταρτοταγής Δομής

Πολλές πρωτεΐνες είναι απλά πολυπεπτίδια, ενώ άλλες αποτελούνται από περισσότερα του ενός πολυπεπτίδια. Ο τρόπος που αυτάς τα πολυπεπτίδια συναρμολογούνται καθορίζει την τεταρτοταγή δομή των πρωτεϊνών. Δεν υπάρχει λειτουργική διαφορά μεταξύ μιας πρωτεΐνης με πολλές λειτουργικές επικράτειες και μίας πρωτεΐνης με πολλές διαφορετικές πολυπεπτιδικές υπομονάδες και έτσι πολλές πρωτεΐνες μπορούν να υπάρξουν και στις δύο μορφές. Για παράδειγμα, οι περισσότεροι μεταγραφικοί παράγοντες είναι απλά πολυπεπτίδια με λειτουργικές επικράτειες με μεταγραφική

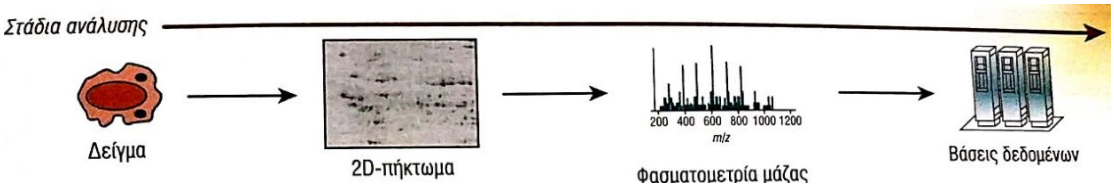
ενεργότητα και αλληλεπιδρούν με DNA, ενώ άλλοι αποτελούνται από πολλές μονάδες [9,16,17].

1.4 Πρωτεωμική

Ο όρος πρωτέωμα [18] παράγεται από την σύντμηση της λέξεως πρωτεΐνη (πρώτε-) με την κατάληξη "-ωμα" που στα ελληνικά δηλώνει σύνολο. Έτσι ο όρος πρωτέωμα, σε αντιπαραβολή με τον όρο γένωμα [19], υποδηλώνει το σύνολο των πρωτεϊνών ενός βιολογικού υλικού. Πρωτεωμική είναι η συστηματική, μεγάλης κλίμακας ανάλυση των πρωτεϊνών. Βασίζεται στην έννοια του πρωτεώματος σαν ένα ολοκληρωμένο σύνολο πρωτεϊνών που παράγονται από ένα συγκεκριμένο κύτταρο, ιστό ή οργανισμό είτε σαν έναν ολοκληρωμένο κατάλογο πρωτεϊνών ή σαν μια λίστα πρωτεϊνών που παράγονται υπό ορισμένες συνθήκες. Πρωταρχικός στόχος τη πρωτεωμικής ανάλυσης είναι να διαχωρίσει, να χαρακτηρίσει και να ταυτοποιήσει τις πρωτεΐνες ενός βιολογικού υλικού και στη συνέχεια να διευκρινίσει τις μεταξύ τους αλληλεπιδράσεις. Η πρωτεωμική μπορεί να εφαρμοστεί σε βιολογικά υγρά τα οποία ως γνωστόν δεν περιέχουν DNA ή RNA, όπως ο ορός και το πλάσμα αίματος, τα ούρα, τα πτύελα κ.λπ. και έτσι η χρήση της πρωτεωμικής σήμερα έχει γενικευθεί σε όλους σχεδόν τους τομείς της βιολογίας [9].

1.4.1 Σκοπός της πρωτεωμικής

Η ταυτοποίηση και ποσοτικοποίηση πρωτεϊνών είναι οι πιο θεμελιώδεις πτυχές της πρωτεωμικής ανάλυσης. Μια τυπική πρωτεωμική ανάλυση περιλαμβάνει το διαχωρισμό πολύπλοκων πρωτεϊνικών μιγμάτων, την αναγνώριση των επιμέρους συστατικών και τη συστηματική ποσοτική ανάλυσή τους (Εικόνα 9). Οι κύριες μορφές δεδομένων που συλλέγονται με αυτή την προσέγγιση είναι η πρωτεϊνική ταυτοποίηση, η παρουσία/απουσία συγκεκριμένων πρωτεϊνών σε συγκεκριμένα δείγματα καθώς και η πρωτεϊνική αφθονία, δηλαδή η ποσότητα των ταυτοποιημένων πρωτεϊνών (πρότυπα έκφρασης πρωτεϊνών ή πρωτεωμική έκφραση) [20].



Εικόνα 9 Τα στάδια της πρωτεωμικής

Στρατηγικές Διαχωρισμού Πρωτεϊνών

Πολλές τεχνικές μπορούν να χρησιμοποιηθούν για να διαχωρίσουν σύνθετα μίγματα πρωτεϊνών. Οι τεχνικές αυτές ανάλογα τις ιδιότητες (φυσικές και χημικές) που αξιοποιούν σαν βάση για τον διαχωρισμό χωρίζονται σε μονοδιάστατες (όσων δηλαδή αξιοποιούν μία μόνο χημική η φυσική ιδιότητα) και πολυδιάστατες. Στην πρωτεωμική λόγω της ανάγκης για

την υψηλή διαχωριστική ικανότητα που απαιτείται καθώς και την υψηλή αποδοτικότητα (να διαχωρίζονται όλες οι πρωτεΐνες με ένα πείραμα) χρησιμοποιούνται πολυδιάστατες τεχνικές διαχωρισμού (χρησιμοποιούν δηλαδή διαδοχικά δύο ή περισσότερες αρχές διαχωρισμού). Οι δύο ομάδες τεχνικών που επικράτησαν στην πρωτεωμική είναι η **ηλεκτροφόρηση πηκτώματος δύο διαστάσεων (2DGE)** και η **πολυδιάστατη υγρή χρωματογραφία (MDLC),** με την τελευταία να συνδυάζεται κατά περίπτωση με περαιτέρω τεχνικές διαχωρισμού, όπως η μονοδιάστατη ηλεκτροφόρηση πηκτώματος, η τριχοειδής ηλεκτροφόρηση ή η χρωματογραφία ισοηλεκτρικής εστίασης [21,22].

Τεχνικές Ταυτοποίησης πρωτεϊνών

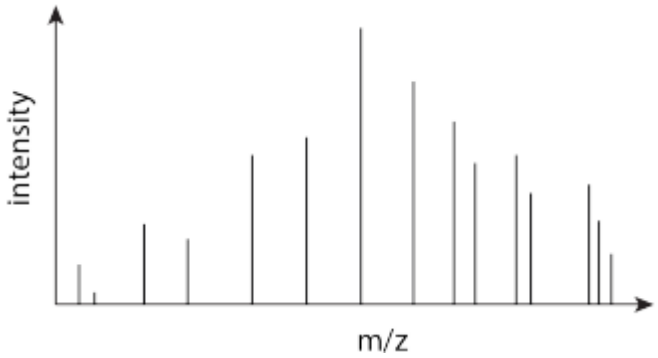
Στις αρχές της πρωτεωμικής ανάλυσης, η **ταυτοποίηση των πρωτεϊνών γινόταν με την χρήση αντισωμάτων**. Με την τεχνική αυτή οι κηλίδες (διαχωρισμένες πρωτεΐνες) στα δισδιάστατα πηκτώματα μπορούσαν να ταυτοποιηθούν χρησιμοποιώντας μόνο δύο μεθόδους. Η πρώτη ήταν να τρέξουν παράλληλα πηκτώματα, το ένα ως πειραματικό και το άλλο με καθορισμένα πρωτεϊνικά πρότυπα. Η ταυτοποίηση των πρωτεϊνών γινόταν συγκρίνοντας την κίνηση των πρωτεϊνών στο κύριο πήκτωμα με αυτή του πυκνώματος με τα γνωστά πρότυπα. Η δεύτερη μέθοδος ήταν η μεταφορά των πρωτεϊνών από το δισδιάστατο πήκτωμα σε μία κατάλληλη μεμβράνη ή υπόστρωμα και η ταυτοποίηση πρωτεϊνών *in situ* με ανιχνευτές, τυπικά αντισώματα [23,24].

Μία άλλη τεχνική που χρησιμοποιήθηκε για την ταυτοποίηση των πρωτεϊνών ήταν ο **προσδιορισμός των αλληλουχιών τους μέσω χημικής αποικοδόμησης**. Οι πρωτεΐνες μπορούν να διασπαστούν πλήρως στα συστατικά τους αμινοξέα με βρασμό σε υψηλής συγκέντρωσης υδροχλωρικό οξύ για 24-72 ώρες. Τα αμινοξέα μπορούν έπειτα να σημανθούν με ένα παράγοντα όπως η νινυδρίνη ή η φλορεσκαμίνη, να διαχωριστούν, και να εντοπιστούν καθώς έλκονται από τη στήλη χρησιμοποιώντας ένα σύνολο αμινοξέων ως πρότυπο. Παρόλο που με την τεχνική αυτή προσδιορίζουμε την αμινοξική σύσταση μιας πρωτεΐνης δεν προσδιορίζουμε την αλληλουχία γιατί όλοι πεπτιδικοί δεσμοί στην πρωτεΐνη είναι σπασμένοι και έτσι τα διαδοχικά αμινοξικά κατάλοιπα δεν μπορούν να αναγνωριστούν άμεσα [9].

Η πιο διαδεδομένη τεχνική που χρησιμοποιείται για την ταυτοποίηση των πρωτεϊνών είναι η **φασματομετρία μάζας [25]**, κατά την οποία τα μόρια των πρωτεϊνών διαχωρίζονται με βάση τη σχέση μάζας/φορτίου τους. Κύριο όργανο για την μέθοδο της φασματομετρίας μάζας είναι ο φασματογράφος μάζας. Ο φασματογράφος μάζας είναι ένα όργανο, το οποίο μπορεί να προσδιορίσει το λόγο μάζα/φορτίο (m/z) των ιόντων στο κενό (Εικόνα 10). Από τα δεδομένα αυτά, μπορεί να προσδιοριστεί η μάζα των μορίων με υψηλό βαθμό ακρίβειας, καθιστώντας εφικτό τον προσδιορισμό της μοριακής σύνθεσης μιας αναλυόμενης ουσίας. Στην πρωτεωμική, η αναλυόμενη ουσία είναι ένα σύνολο από πεπτίδια, που προέρχονται

από ένα δείγμα πρωτεΐνης ύστερα από πέψη του με θρυψίνη ή άλλο παρόμοιο αντιδραστήριο. Υπάρχουν τρείς τύποι ανάλυσης που μπορούν να πραγματοποιηθούν:

- Ανάλυση ακέραιων πεπτιδικών ιόντων. Αυτή η ανάλυση επιτρέπει την μέτρηση της μάζας των ακέραιων πεπτιδίων και οι μάζες αυτές μπορούν να χρησιμοποιηθούν για την ταυτοποίηση πρωτεϊνών στο δείγμα με αναζήτηση συσχετίσεων σε βάσεις δεδομένων.
- Ανάλυση θραυσμάτων πεπτιδικών ιόντων. Σε αυτή την ανάλυση είναι εφικτός ο προσδιορισμός της μάζας πεπτιδικών θραυσμάτων, τα οποία μπορούν να χρησιμοποιηθούν για αναζήτηση συσχετίσεων σε βάσεις δεδομένων ή προσδιορισμό αλληλουχιών *de novo* ή με υβριδικές.
- Ανάλυση κατακερματισμένων πρωτεϊνών. Η τεχνική είναι γνωστή ως προσέγγιση από πάνω προς τα κάτω γιατί αρχίζει με την ακέραιη πρωτεΐνη. Η προσέγγιση αυτή χρησιμοποιείται κυρίως για πρωτεΐνες των οποίων η μάζα ξεπερνά σε μέγεθος τα 50 KDa (περίπου 500 αμινοξέα) [9].



Εικόνα 10 προσδιορισμός λόγου m/z

Οι φασματογράφοι μάζας έχουν τρία βασικά λειτουργικά στοιχεία: μια πηγή ιόντων, έναν αναλυτή μάζας και έναν ανιχνευτή ιόντων. Η λειτουργία της πηγής ιόντων είναι να μετατρέπει το αναλυόμενο δείγμα σε ιόντα στην αέρια φάση σε συνθήκες κενού. Τα ιόντα στη συνέχεια επιταχύνονται σε ένα ηλεκτρικό πεδίο προς τον αναλυτή, ο οποίος τα διαχωρίζει ανάλογα με το λόγο m/z που έχει το καθένα καθώς κατευθύνονται στον ανιχνευτή. Η λειτουργία του ανιχνευτή είναι να καταγράφει το σήμα κάθε ιόντος ξεχωριστά [26,27].

Ένα από τα προβλήματα που αντιμετώπισε αυτή η τεχνική ήταν η δυσκολία να παραχθούν ακέραια ιόντα σε αέρια μορφή. Αυτό επιλύθηκε με την ανάπτυξη των αποκαλούμενων **ήπιων μεθόδων ιονισμού** [28] οι οποίες επιτυγχάνουν τον ιονισμό πεπτιδίων και μεγαλύτερων μορίων χωρίς σημαντικό κατακερματισμό. Οι πρώτες ήπιες μέθοδοι ιονισμού στην πρωτεωμική ήταν οι μέθοδοι **MALDI** [29] και **ESI** [30]. Η MALDI είναι μια διαδικασία κατά την οποία το αναλυόμενο δείγμα είναι αρχικά αναμειγμένο με μια μήτρα

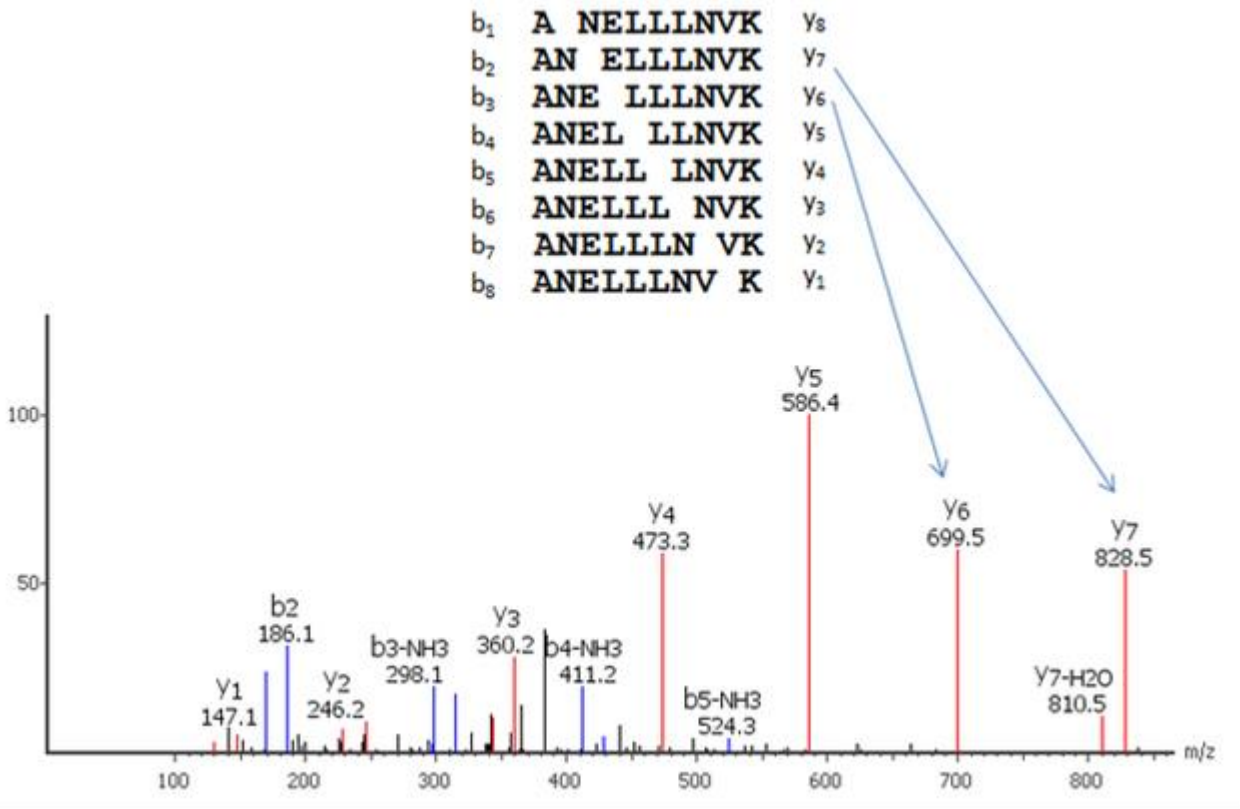
αρωματικής χημικής ένωσης, η οποία μπορεί να απορροφήσει την ενέργεια ενός laser. Το αναλυόμενο δείγμα και η μήτρα διαλύονται σε ένα οργανικό διαλυτικό μέσο και τοποθετούνται σε ένα μεταλλικό δειγματολήπτη. Ο διαλύτης εξατμίζεται, αφήνοντας το δείγμα εμποτισμένο στους κρυστάλλους της μήτρας, οι οποίοι τοποθετούνται στον θάλαμο κενού του φασματογράφου μάζας και εκτίθενται σε υψηλή τάση. Την ίδια στιγμή οι κρύσταλλοι στοχεύονται από ένα στιγμιαίο παλμό laser. Η ενέργεια του laser απορροφάται από τους κρυστάλλους και εκπέμπεται από αυτούς με τη μορφή θερμότητας, με αποτέλεσμα την ταχεία εξάχνωση, που μετατρέπει το αναλυόμενο δείγμα σε ιόντα αέριας φάσης. Αυτά επιταχύνονται μακριά από το στόχο μέσω του αναλυτή προς τον ανιχνευτή. Η μέθοδος ESI αναφέρεται στον ιονισμό με ηλεκτροψεκασμό στον οποίο το αναλυόμενο δείγμα διαλύεται και προωθείται μέσα από μια λεπτή βελόνα που διατηρείται σε υψηλή απόσταση. Από τη βελόνα προκύπτει ένας ψεκασμός φορτισμένων σταγονιδίων, ο οποίος κατευθύνεται στο θάλαμο κενού του φασματογράφου μάζας μέσω ενός λεπτού στομίου. Όταν τα σταγονίδια εισέρχονται στον φασματογράφο μάζας στεγνώνονται από ένα ρεύμα αδρανούς αερίου, με αποτέλεσμα τα ιόντα στην αέρια να επιταχύνονται μέσω του αναλυτή προς τον ανιχνευτή [31].

Η πρώτη ευρέως χρησιμοποιούμενη μέθοδος για την ταυτοποίηση πρωτεϊνών με τη χρήση φασματομετρίας μάζας ήταν το πεπτιδικό δακτυλικό αποτύπωμα μάζας (peptide mass fingerprinting. PMF) η οποία αφορά την ταυτοποίηση πρωτεϊνών μέσω της χρήσης δεδομένων από τις μάζες των ακέραιων πεπτιδίων [32]. Αυτή η μέθοδος είναι συμβατή με την 2DGE και την φασματομετρία μάζας MALDI-TOF, στην οποία οι πρωτεΐνες κόβονται πριν την πέψη τους σε πεπτίδια. Η αρχή της μεθόδου βασίζεται στο ότι κάθε πρωτεΐνη μπορεί να ταυτοποιηθεί μοναδικά από τις μάζες των συστατικών πεπτιδίων της, με μια μοναδική υπογραφή γνωστή και ως πεπτιδικό αποτύπωμα μάζας. Αλγόριθμοι που επιτρέπουν την αναζήτηση της πρωτεΐνης, βάσει του πεπτιδικού αποτυπώματος σε βάσεις δεδομένων εφαρμόστηκαν σε πολλά προγράμματα υπολογιστών και μερικά από τα πιο κοινώς χρησιμοποιούμενα είναι τα Mascot, MS-Fit kai Profound [20,33-36].

Τα στάδια που ακολουθούνται με την μέθοδο της φασματομετρίας μάζας είναι τα εξής:

- Το δείγμα που μας ενδιαφέρει πρέπει να συνιστά μία πρωτεΐνη ή ένα απλό μίγμα, για παράδειγμα μια μοναδική κηλίδα ενός δισδιάστατού πηκτώματος. Το δείγμα πέπτεται με ένα ειδικό αντιδραστήριο πέψης, συνήθως θρυψίνη [37,38].
- Οι μάζες πεπτιδίων προσδιορίζονται, για παράδειγμα με φασματομετρία μάζας maldi-tof (Εικόνα 11).
- Ο ερευνητής επιλέγει το πρόγραμμα και μία η περισσότερες βάσεις πρωτεϊνικών αλληλουχιών που θα χρησιμοποιήσει για αναζήτηση μέσω συσχέτισης [39].

- Ο αλγόριθμος εκτελεί εικονική πέψη σε κάθε αλληλουχία των πρωτεϊνών στην βάση χρησιμοποιώντας το ίδιο ένζυμο που χρησιμοποιήθηκε πειραματικά (θρυψίνη) και στη συνέχεια υπολογίζει τις θεωρητικές πεπτιδικές μάζες για κάθε πρωτεΐνη.
- Ο αλγόριθμος προσπαθεί να συσχετίσει τις θεωρητικές πεπτιδικές μάζες με τις αντίστοιχες πειραματικές.
- Οι πρωτεΐνες της βάσης δεδομένων ταξινομούνται κατά σειρά μεγαλύτερης συσχέτισης και συνήθως χρησιμοποιείται ένα κατώφλι σημαντικότητας (όριο) που βασίζεται σε έναν ελάχιστο αριθμό πεπτιδίων που ταυτίζονται [40] (Εικόνα 12).



Εικόνα 11 Προσδιορισμός των μαζών των πεπτιδίων

Οι μάζες των ακέραιων πεπτιδίων είναι υπερβολικά διακριτές, καθιστώντας την PMF μια πολύ αξιόπιστη μέθοδο ταυτοποίησης πρωτεϊνών. Παρόλα αυτά, επειδή η PMF στηρίζεται σε αναζήτηση συσχετίσεων, η πιθανότητα εύρεσης μιας ταυτόσημης πρωτεΐνης εξαρτάται τόσο από την ποιότητα των πειραματικών δεδομένων, όσο και από τη διαθεσιμότητα των πληροφοριών για τον οργανισμό από τον οποίο συλλέχθηκε το δείγμα. Τα χαρακτηριστικά των δεδομένων που πρέπει να ληφθούν υπόψη για την αξιόπιστη ταυτοποίηση των πρωτεϊνών περιλαμβάνουν την ποιότητα και τη σχετική ένταση των κορυφών του φάσματος μάζας, της ακρίβειας του οργάνου μέτρησης, την περιοχή κάλυψης της πρωτεΐνης και πιθανούς παρεμβαλλόμενους παράγοντες όπως μεταφραστικές τροποποιήσεις και μη σωστές διασπάσεις. Οι παραπάνω παράγοντες

επηρεάζουν την πιθανότητα μιας ταύτισης να είναι αληθής ή ψευδής, μια πιθανότητα που συνήθως εκφράζεται ως MOWSE score [41] (Εικόνα 12).

HN	Proteins description	MOWSE score	MW (Da)	Peptides matched
2	Glycinin	2436	54927	86/99
3	alpha subunit of beta conglycinin	1624	63184	52/74
4	Lipoxygenase	1001	97490	40/48
5	Sucrose-binding protein	878	60884	34/42
6	unnamed protein product	855	22972	24/25
7	Seed biotin-containing protein	654	67894	20/24
8	ribulose-1,5-bisphosphate carboxy.	652	53056	35/48
9	AtpB	406	51944	14/16
10	beta-amylase	399	56378	15/19
11	chloroplast protein	347	26530	8/9
12	unnamed protein product	341	47117	11/13
13	allergen Gly m Bd 28K	328	52780	9/9
14	AtpA	324	54044	10/12
15	seed maturation protein	312	17907	11/13
16	HSP 70 kDa protein 1	312	71420	10/13
17	protein disulfide isomerase	308	58963	12/19
18	unknown protein	299	43082	13/16
19	putative histone H2B	284	14338	2/2
20	Enolase	232	48127	7/11
21	alcohol dehydrogenase 1	127	20101	3/5

Εικόνα 12 Αποτελέσματα φασματογράφου μάζας έπειτα από χρήση βιοπληροφορικών προγραμμάτων

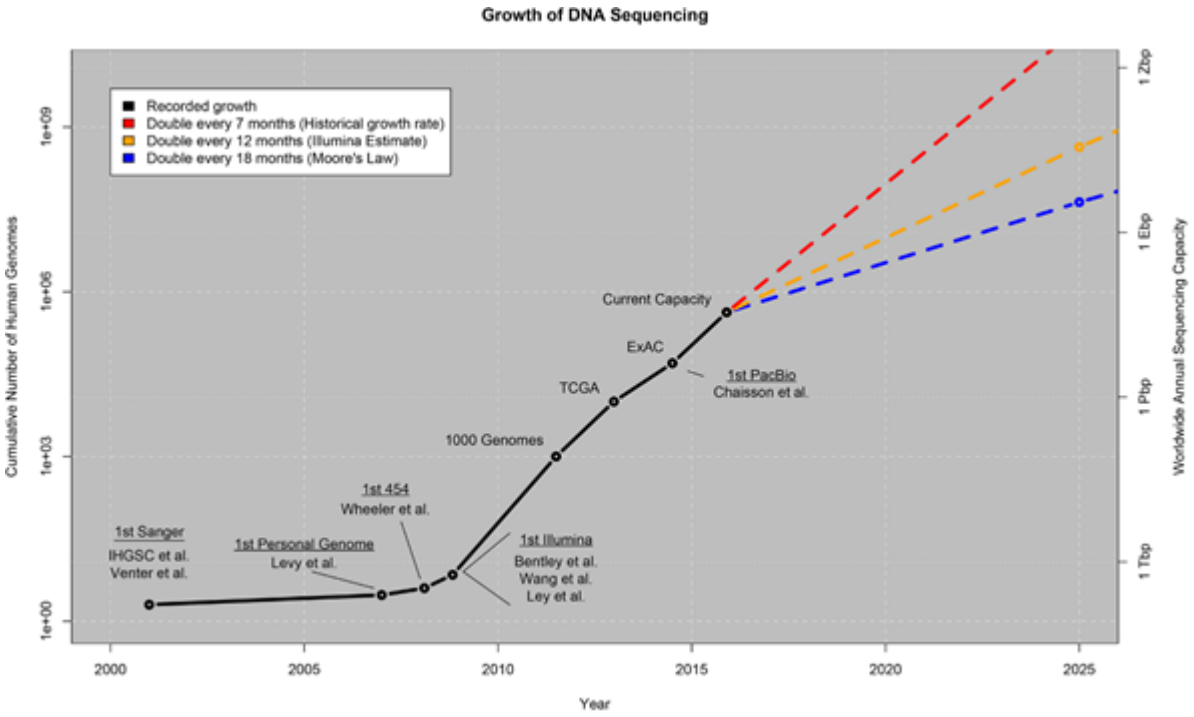
1.4.2 Βιολογικά δεδομένα και δημόσιες βάσεις δεδομένων

Οι τεχνολογίες που αναπτύχθηκαν τις τελευταίες δεκαετίες για την διερεύνηση του γονιδιώματος και του πρωτεώματος των οργανισμών, καθώς και τις αλληλεπιδράσεις και συσχετίσεις τους έχουν δημιουργήσει και συνεχίζουν να δημιουργούν τεράστια ποσά δεδομένων τα οποία είναι αδύνατο να αναλυθούν και να εκτιμηθούν χωρίς την βοήθεια της πληροφορικής [42,43]. Αποτέλεσμα είναι η δημιουργία διαφόρων ανοιχτών βάσεων δεδομένων για μια πληθώρα διαφορετικών δεδομένων σε πεδία έρευνας όπως αυτά της γενωμικής, της πρωτεωμικής, των παθήσεων, της μεταβολικής και άλλων.

Data Phase	Astronomy	Twitter	YouTube	Genomics
Acquisition	25 zetta-bytes/year	0.5–15 billion tweets/year	500–900 million hours/year	1 zetta-bases/year
Storage	1 EB/year	1–17 PB/year	1–2 EB/year	2–40 EB/year
Analysis	In situ data reduction	Topic and sentiment mining	Limited requirements	Heterogeneous data and analysis
	Real-time processing	Metadata analysis		Variant calling, ~2 trillion central processing unit (CPU) hours
	Massive volumes			All-pairs genome alignments, ~10,000 trillion CPU hours
Distribution	Dedicated lines from antennae to server (600 TB/s)	Small units of distribution	Major component of modern user's bandwidth (10 MB/s)	Many small (10 MB/s) and fewer massive (10 TB/s) data movement

doi:10.1371/journal.pbio.1002195.t001

Εικόνα 13 Σύγκριση δεδομένων αλληλούχισης DNA με άλλες πλατφόρμες περιεχομένου



Εικόνα 14 Ρυθμός αύξησης δεδομένων αλληλούχισης DNA

Στην πρωτεωμική, μια από τις πιο συχνές εφαρμογές της βιοπληροφορικής είναι η αναζήτηση σε βάσεις δεδομένων που περιλαμβάνουν τις αμινοξικές αλληλουχίες των πρωτεϊνών με σκοπό την συσχέτιση τους με κατακερματισμένες πρωτεΐνες που έχουν προέλθει από την ανάλυση τους με την τεχνική της φασματομετρίας μάζας. Αν τα αποτελέσματα είναι επιτυχημένα οδηγούν στην οριστική ταυτοποίηση μιας πρωτεΐνης, που μέχρι τη στιγμή εκείνη χαρακτηριζόταν αποκλειστικά από τη θέση της σε ένα δισδιάστατο πήκτωμα ηλεκτροφόρησης ή είχε περιγραφεί θεωρητικά από την μηχανιστική μετάφραση ενός γονιδίου. Συγκρίνοντας την αλληλουχία μίας πρωτεΐνης με άλλες που έχουν καταγραφεί στις βάσεις δεδομένων, ένας ερευνητής μπορεί να βρει πληροφορίες για τα λειτουργικά δομικά στοιχεία της πρωτεΐνης, τις φυσικοχημικές της ιδιότητες, τις αλληλεπιδράσεις με άλλα μόρια, την ύπαρξη τροποποιήσεων, τη βιοχημική δραστικότητα (μοριακή λειτουργία), το συνολικό της ρόλο στο κύτταρο ή τον οργανισμό, τις σχετικές πρωτεΐνες σε άλλους οργανισμούς, την εξέλιξη της οικογένειας της πρωτεΐνης και ακόμη τον πιθανό δυνητικό ρόλο της σε κάποια ασθένεια ή αλληλεπίδραση της με φάρμακα [9].

Πολλές χιλιάδες αλληλουχίες πρωτεϊνών έχουν κατατεθεί σε τρείς κυρίως βάσεις δεδομένων, είτε με *de novo* προσδιορισμό των αλληλουχιών είτε με μετάφραση της νουκλεοτιδικής αλληλουχίας. Η πιο περιεκτική βάση δεδομένων είναι η Uniprot [44 - 46], η οποία ξεκίνησε το 2003 για να συνδυάσει την αλληλοεπικάλυψη των πηγών της SwissProt [47], της TrEMBL [48] και της PIR-PDB [49]. Πριν από αυτή τη συγχώνευση, οι βάσεις δεδομένων συνυπήρχαν, αλλά διέφεραν στην κάλυψη των πρωτεϊνικών αλληλουχιών και του λειτουργικού σχολιασμού. Μερικές από τις σημαντικότερες βάσεις εμφανίζονται στον παρακάτω πίνακα (Πίνακας 3) [9,50]. Οι βάσεις που παρουσιάζονται συνοπτικά στον

πίνακα, επιλέχθηκαν είτε γιατί έγινε χρήση των δεδομένων τους, είτε γιατί η πληροφορία που περιέχουν θεωρείται πως μπορεί να ενταχθεί στην παρούσα μελέτη.

Ονομασία	Δεδομένα που περιέχονται – Παρατηρήσεις
UnitProtKB/Swiss-Prot	Περιέχει δεδομένα πρωτεϊνών που έχουν επισημανθεί και αξιολογηθεί χειροκίνητα από ερευνητές και αξιολογητές. Στη βάση αποθηκεύονται μόνο αναθεωρημένα (reviewed) δεδομένα.
UnitProtKB/TrEMBL	Περιέχει δεδομένα που έχουν αναλυθεί και εμπλουτιστεί με επισημάνσεις μέσω υπολογιστικών μεθόδων.
Protein Information Resource	Περιέχει πλήθος διαφορετικών δεδομένων που αφορούν τις πρωτεΐνες όπως πληροφορίες για την λειτουργικότητα ή την τοπολογία των πρωτεϊνών, βιβλιογραφία και άλλες.
Uniprot	Αποτελείται από την σύνθεση των τριών παραπάνω βάσεων καθώς και από άλλες πηγές που ομαδοποιούνται κάτω από μια κοινή βάση. Παράλληλα προσφέρει πληθώρα εργαλείων, όπως εργαλεία για την στοίχιση αλληλουχιών (alignment), αναζήτησης εντός των βάσεων και άλλα.
Human Protein Atlas	Αντιστοίχιση πρωτεϊνών του ανθρώπινου οργανισμού με τα όργανα χρησιμοποιώντας διάφορα δεδομένα (-omics data)
Protein Data Bank	Περιέχει τις τρισδιάστατες δομές πρωτεϊνών που έχουν προκύψει από διάφορες τεχνολογίες όπως Κρυσταλλογραφία ακτινών X (X-ray crystallography) και η NMR φασματοσκοπία (NMR spectroscopy). Αποτελείται από σύμπραξη των παρακάτω: Protein Data Bank in Europe Protein Data Bank in Japan Research Collaboratory for Structural Bioinformatics (RCSB)
Pfam	Περιέχει πληροφορίες για τον διαχωρισμό των πρωτεϊνών σε οικογένειες με βάση τις επισημάνσεις, την στοίχιση των αλληλουχιών τους καθώς και την χρήση Κρυφών Μαρκοβιανών Μοντέλων (Hidden Markov Models)
Reactome	Περιέχει πληροφορίες για βιολογικά μονοπάτια και διαδικασίες και την εμπλοκή των πρωτεϊνών σε αυτά.
CharProtDP	Είναι μια βάση που την επιμελούνται ειδικοί και περιλαμβάνει βιοχημικά χαρακτηρισμένες πρωτεΐνες που βασίζονται στη συλλογή πληροφοριών για την λειτουργία των πρωτεϊνών από την βιβλιογραφία και περαιτέρω επέκταση συμπεριλαμβάνοντας δεδομένα από άλλες ελεύθερα διαθέσιμες συλλογές πειραματικά χαρακτηρισμένων πρωτεϊνών.

Πίνακας 3 Πρωτεϊνικές βάσεις δεδομένων [2,5]

1.5 Υπολογιστικό Υπόβαθρο

1.5.1 Αλγόριθμος – Επίλυση προβλημάτων

Ο άνθρωπος από σχετικά νωρίς στην ιστορία του, έμαθε να χρησιμοποιεί επαναλαμβανόμενες διαδικασίες για να επιλύσει τα καθημερινά του προβλήματα. Με το πέρασμα των αιώνων και την έλευση των μαθηματικών κωδικοποίησε αυτές τις διαδικασίες με καλώς ορισμένους τρόπους.

Η κωδικοποίηση αυτή ονομάστηκε αλγόριθμος και ορίστηκε ως: «μια πεπερασμένη σειρά ενεργειών αυστηρά καθορισμένων και εκτελέσιμων σε πεπερασμένο χρονικό όριο που στοχεύει στην επίλυση ενός προβλήματος». Πήρε το όνομά της προς τιμήν του Πέρση μαθηματικού Μοχάμεντ Ιμπν Μουσά αλ-Χουαρίζμι που ήταν ο πρώτος που συστηματικά συγκέντρωσε και τυποποίησε λύσεις για την επίλυση αλγεβρικών προβλημάτων.

Όπως γίνεται εύκολα κατανοητό, ο όρος αλγόριθμος δεν περιορίζεται μόνο στην επίλυση μαθηματικών προβλημάτων αλλά περιλαμβάνει κάθε πρόβλημα το οποίο μπορεί να λυθεί σε εύλογο χρονικό διάστημα ακολουθώντας τα ίδια βήματα και προσδοκώντας πάντα το ίδιο αποτέλεσμα (ντετερμινιστική έκβαση της επίλυσης) [51,52].
Ένας αλγόριθμος πρέπει να ικανοποιεί τα παρακάτω κριτήρια:

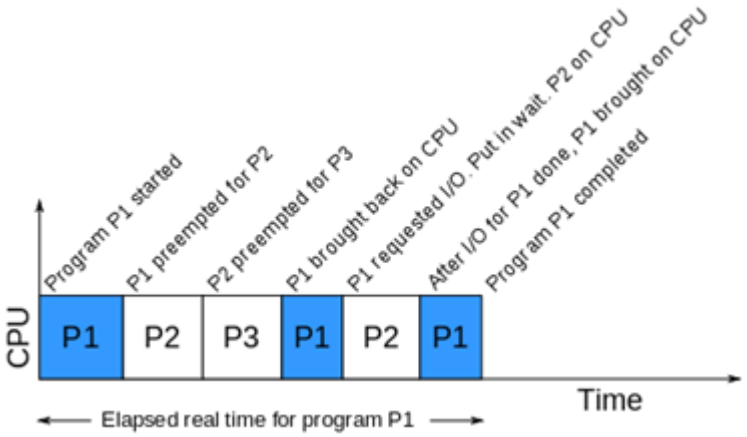
- **Περατότητα** – Finiteness: Κάθε εκτέλεση του αλγορίθμου πρέπει να ολοκληρώνεται σε πεπερασμένο χρονικό διάστημα.
- **Καθοριστικότητα** – Definiteness: Κάθε εντολή πρέπει να ορίζεται ακριβώς χωρίς να μπορεί να παρερμηνευτεί ο τρόπος με τον οποίο θα εκτελεστεί και να μπορεί να αντιμετωπίσει τυχόν ανωμαλίες που μπορεί να προκύψουν κατά την εκτέλεση αυτής.
- **Αποτελεσματικότητα** – Effectiveness: Κάθε βήμα / εντολή του αλγορίθμου πρέπει να έχει σαφώς οριστεί και να είναι εκτελέσιμο/η.
- **Είσοδος Δεδομένων** - Input Data: Ένας αλγόριθμος μπορεί λαμβάνει κάποια δεδομένα ως είσοδο χωρίς αυτό όμως να είναι υποχρεωτικό (π.χ. μπορεί απλά να παράγει τυχαίους αριθμούς).
- **Έξοδος Δεδομένων** – Output Data: Όλοι οι αλγόριθμοι πρέπει να έχουν τουλάχιστον μία έξοδο με κάποια μορφή (π.χ. αριθμητικά δεδομένα, ενεργοποίηση μιας λειτουργίας κ.α.).

1.5.2 Βασικές αρχές επεξεργασίας

Οι ηλεκτρονικοί υπολογιστές έχουν μπει στην ζωή μας από τα μέσα του 20ου αιώνα. Η τεράστια εξέλιξή τους έχει βοηθήσει στην επίλυση προβλημάτων για τα οποία οι άνθρωποι θα ήθελαν πολλά χρόνια να εκτελέσουν χωρίς την ταχύτητα που τους πρόσφεραν. «Bicycle for the mind» («Ποδήλατα για το μυαλό») όπως περιέγραφε ο Steve Jobs. Για να μπορέσει ο υπολογιστής να επιλύσει αυτά τα προβλήματα είναι απαραίτητο να γνωρίζει τα βήματα που πρέπει να ακολουθήσει για την επίλυσή τους, δηλαδή τον αλγόριθμο.

Οι σημερινοί ηλεκτρονικοί υπολογιστές μπορούν να αντιληφθούν/αναγνωρίσουν μόνο το δυαδικό αριθμητικό σύστημα (0 / 1) [53]. Ο τρόπος με τον οποίο η περιγραφή ενός αλγορίθμου από την φυσική γλώσσα που αντιλαμβάνεται ο άνθρωπος μεταφέρεται στην δυαδική που αναγνωρίζει ο υπολογιστής είναι μέσα από τις διάφορες γλώσσες προγραμματισμού που αναλαμβάνουν με την βοήθεια και του λειτουργικού συστήματος να γεφυρώσουν το παραπάνω χάσμα [54,55]. Ο κώδικας μεταφράζεται από την γλώσσα προγραμματισμού που έχει γραφτεί (π.χ. C, C++, Java) μέσω κατάλληλου λογισμικού σε εντολές γλώσσας μηχανής και εκτελείται από την Κεντρική Μονάδα Επεξεργασίας (CPU). Για να δρομολογήσει την εκτέλεση των εντολών στην CPU το λειτουργικό σύστημα χρησιμοποιεί την έννοια του νήματος (thread). Ως νήμα ορίζεται η μικρότερη ανεξάρτητη ακολουθία εντολών που μπορεί να δρομολογήσει το λειτουργικό σύστημα. Το νήμα είναι αυτό το οποίο αναλαμβάνει την πραγματική εκτέλεση των εντολών στον πυρήνα της CPU.

Οι πρώιμες CPU (μέχρι τα τέλη της δεκαετίας του 1990) αποτελούνταν από ένα πυρήνα οπότε σε κάθε χρονική στιγμή μπορούσαν να εξυπηρετήσουν ένα μόνο νήμα. Παρόλα αυτά είχαν την δυνατότητα να μπορούν να τρέξουν ταυτόχρονα παραπάνω από ένα πρόγραμμα (παράλληλα). Ο τρόπος με τον οποίο μονοπύρηνα συστήματα φαίνεται να εκτελούν παράλληλα περισσότερες διεργασίες, είναι μέσω της κατάτμησης του χρόνου σε μικρές υπομονάδες και της εκτέλεσης κάθε νήματος για ένα μικρό χρονικό διάστημα μετά από το οποίο το λειτουργικό σύστημα αλλάζει σε άλλο νήμα εκτέλεσης για άλλο ένα χρονικό διάστημα (Εικόνα 15). Μέσω αυτής της διαδικασίας ο χρήστης έχει την ψευδαίσθηση πως τα προγράμματά του εκτελούνται παράλληλα. Η εναλλαγή μεταξύ των διαφορετικών προγραμμάτων ώστε να φαίνεται πως εκτελούνται παράλληλα, εισάγει επιπλέον καθυστερήσεις στο σύστημα. Σε παλιότερα μονοπύρηνα συστήματα συχνές ήταν οι περιπτώσεις όπου οι εντολές κάποιου νήματος οδηγούσαν σε κάποια κατάσταση ανταγωνισμού με άλλο (race condition) με αποτέλεσμα ακόμη και την πλήρη παύση λειτουργίας του λειτουργικού συστήματος και την ανάγκη επανεκκίνησης του υπολογιστή [56].



Εικόνα 15 Παράδειγμα εκτέλεσης πολλών προγραμμάτων σε μονοπύρηνο επεξεργαστή με ένα νήμα μέσω διαμοιρασμού χρόνο

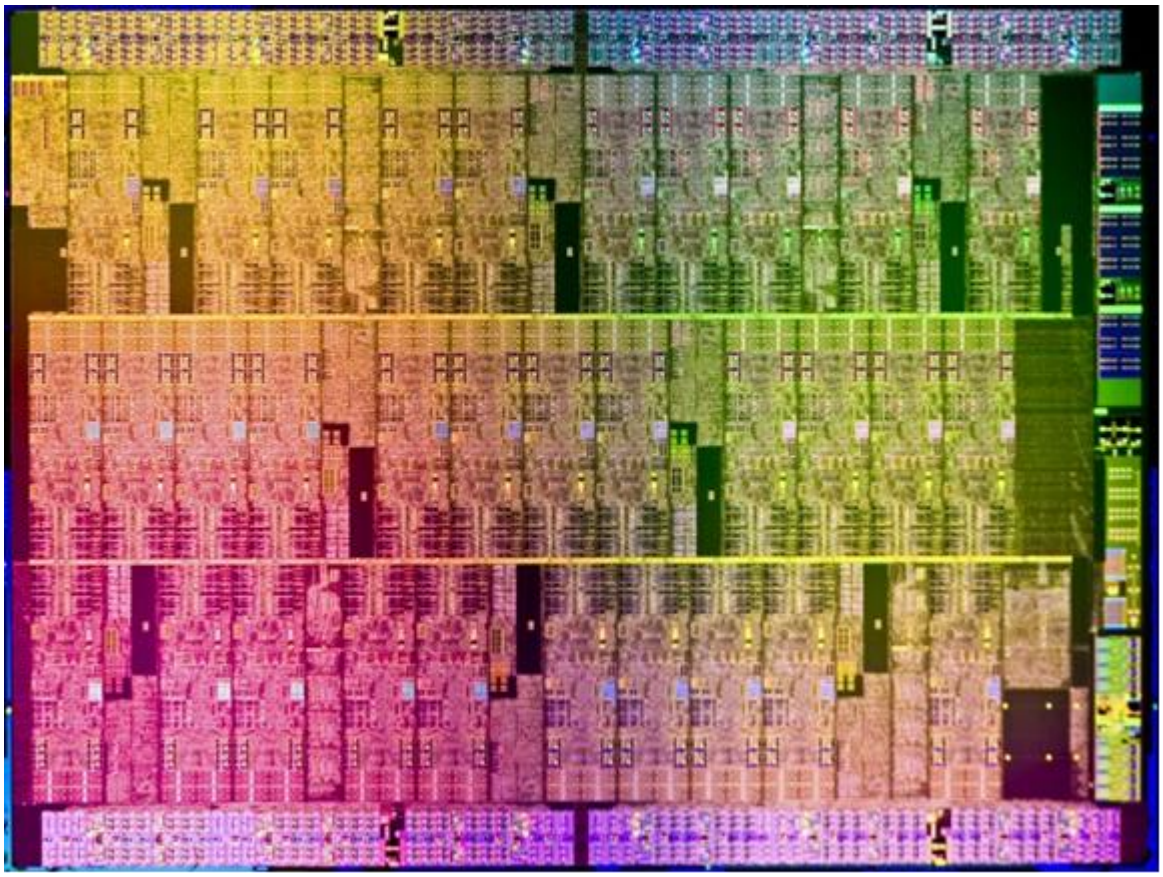
1.5.3 Παράλληλη επεξεργασία

Ο παραλληλισμός των υπολογιστικών διεργασιών, παρότι έχει γίνει ευρέως διαδεδομένος τα τελευταία χρόνια, δεν είναι μια καινούρια ιδέα. Υπάρχει τουλάχιστον από τα μέσα του 1950.

Τις τελευταίες δεκαετίες οι επεξεργαστές γινόντουσαν όλο και γρηγορότεροι και πιο δυνατοί ακολουθώντας τον νόμο του Moore που όριζε πως κάθε 18 μήνες περίπου το ρολόι τους θα διπλασιαζόταν σε απόδοση. Η συρρίκνωση των τρανζίστορ εντός των chip των επεξεργαστών καθώς και οι πολύ υψηλές ταχύτητες αυτών, οδήγησαν στα τέλη της δεκαετίας του 1990 τους σχεδιαστές να αγγίξουν τα φυσικά όρια στα οποία μπορούσε να φτάσει ένας επεξεργαστής πριν την καταστροφή του ή την αλλοίωση των αποτελεσμάτων, μέσω της εμφάνισης κβαντικών φαινομένων ανάμεσα στα τρανζίστορ του.

Μπροστά στο αδιέξοδο οι κατασκευαστές των επεξεργαστών υιοθέτησαν μια διαφορετική προσέγγιση, ενσωματώνοντας περισσότερους από έναν πυρήνες εντός του chip της CPU. Σήμερα ο νόμος του Moore συνεχίζει να ισχύει, μόνο που πλέον αυτό που διπλασιάζεται ανά 18 μήνες περίπου είναι το πλήθος των επεξεργαστών σε μια μονάδα CPU.

Είναι πολύ σύνηθες πλέον σε οικονομικά συστήματα μερικών εκατοντάδων ευρώ να συναντώνται επεξεργαστές με πολλούς πυρήνες / νήματα επεξεργασίας (4, 8 ή και παραπάνω). Αυτή η αύξηση στην επεξεργαστική ισχύ μέσω της αύξησης των πυρήνων, ανάγκασε και τους δημιουργούς λογισμικού να προσεγγίζουν διαφορετικά τις υλοποιήσεις τους.



Εικόνα 16 Intel Knights Landing Many core CPU (72 cores x 4 threads)

Ο παραλληλισμός μπορεί να επιτευχθεί με αρκετούς τρόπους. Μερικοί από αυτούς είναι [57-59]:

- **Κατανεμημένος Παραλληλισμός:** Ολόκληρες εφαρμογές ή κομμάτια υπολογισμών μεταφέρονται για εκτέλεση σε άλλα συστήματα, ανταλλάσσοντας μεταξύ τους στοιχεία με πρωτόκολλα όπως η Διεπαφή Μετάδοσης Μηνύματος (MPI: Message Passage Interface)
- **Εικονικοποίηση** (Virtualization): Εκτέλεση πολλών λειτουργικών συστημάτων στον ίδιο υπολογιστή. Είναι απαραίτητη η ύπαρξη κατάλληλου λογισμικού (hypervisor) που αναλαμβάνει την διαχείριση των πόρων.
- **Παραλληλισμός σε επίπεδο διεργασίας:** Εστιάζει στην παραλληλοποίηση ολόκληρων διεργασιών. Συνήθως υπάρχουν περισσότερες διεργασίες απ' ότι νήματα εκτέλεσης (threads) οπότε υπάρχουν κατάλληλοι μηχανισμοί χρονοπρογραμματισμού ώστε κάθε διεργασία να έχει δίκαιη χρονική πρόσβαση σε κάποιο νήμα εκτέλεσης.
- **Παραλληλισμός σε επίπεδο νήματος:** Μια εφαρμογή μπορεί να επιλέξει να χρησιμοποιήσει πολλαπλά νήματα για την εκτέλεση των υπολογιστικών τις μερών.
- **Παραλληλισμός σε επίπεδο εντολής:** Για να μπορέσει να υλοποιηθεί θα πρέπει να υποστηρίζεται από τον επεξεργαστή και συνήθως γίνεται αυτόματα από αυτόν. Ο

επεξεργαστής έχει την ικανότητα να τρέξει παράλληλα διαφορετικές εντολές εντός του.

- **Παραλληλισμός σε επίπεδο δεδομένων:** Για να μπορέσει να υλοποιηθεί ο συγκεκριμένος τύπος παραλληλοποίησης θα πρέπει ο επεξεργαστής να υποστηρίζει λειτουργίες Single Instruction Multiple Dataset. Βασίζεται στην ύπαρξη μεγάλων καταχωρητών πχ 128 bit οι οποίοι μπορούν να διατηρήσουν 4 αριθμούς 32 bit και να εκτελέσουν μια κοινή λειτουργία στον ίδιο κύκλο του ρολογιού και στους 4. Παραδείγματα τέτοιων τεχνολογιών είναι το AVX της Intel, το CUDA της NVIDIA και άλλα.

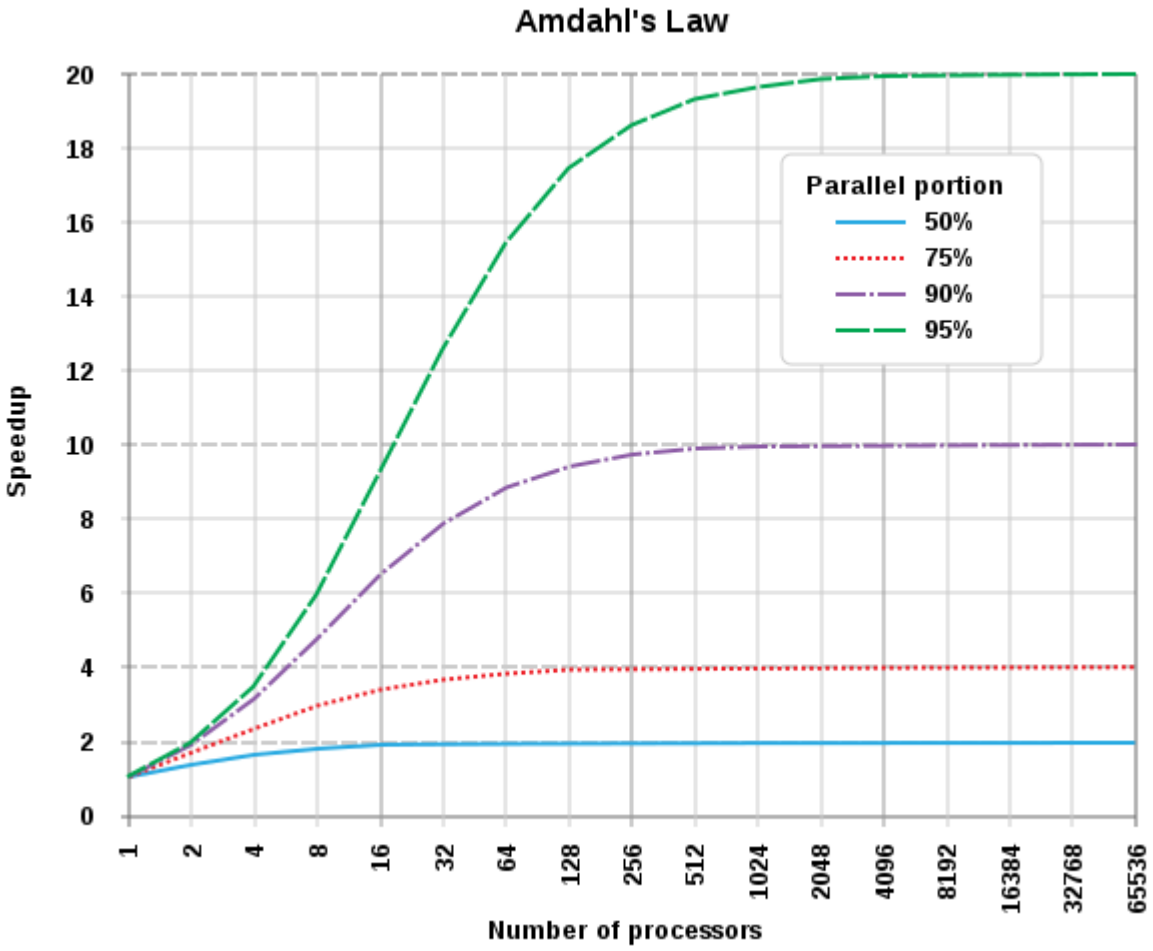
Όπως αναφέρθηκε παραπάνω για την επίλυση ενός προβλήματος, ο υπολογιστής ακολουθεί καλά καθορισμένα βήματα που περιγράφουν τον τρόπο επίλυσης της συγκεκριμένης κλάσης προβλημάτων (σειριακή επίλυση). Αν το σύστημα έχει περισσότερες από μια υπολογιστικές μονάδες (VPU) τότε αυτές μπορούν να χρησιμοποιηθούν ώστε να εκτελέσουν τμήματα των υπολογισμών παράλληλα ώστε να ολοκληρωθεί ταχύτερα η επίλυση. Θεωρητικά αν είχαμε N υπολογιστικές μονάδες θα μπορούσαμε να μειώσουμε τον χρόνο εκτέλεσης στο 1/N.

Εικόνα 17 Παραδείγματα 2-πύρηνων επεξεργαστών με 1 (N=2) και 2 (N=4) νήματα ανά πυρήνα αντίστοιχα

Στην πράξη όμως η πραγματική απόδοση που θα πάρουμε είναι πολύ κατώτερη όπως είχε από νωρίς αποδείξει ο Amdahl στον ομώνυμο νόμο του και θα εξαρτάται πλήρως από την αναλογία χρόνου του μέρους του υπολογισμού που δεν μπορούμε να παραλληλοποιήσουμε (Εικόνα 18 και Εικόνα 19) .

$$S_{\text{latency}}(s) = \frac{1}{(1-p) + \frac{p}{s}} \qquad \begin{cases} S_{\text{latency}}(s) \leq \dfrac{1}{1-p} \\ \lim\limits_{s \to \infty} S_{\text{latency}}(s) = \dfrac{1}{1-p}. \end{cases}$$

Εικόνα 18 Νόμος του Amdahl



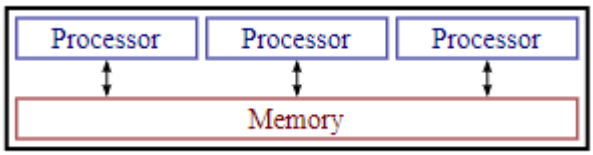
Εικόνα 19 Εξέλιξη της αύξησης απόδοσης βάσει του νόμου του Amdahl

Επιπλέον της προηγούμενης παρατήρησης, ο συγχρονισμός των νημάτων, ειδικά όταν αυτά θα πρέπει να ανταλλάξουν μεταξύ τους μηνύματα μπορεί να υποβαθμίσει κατά πολύ την απόδοση ενός παράλληλου αλγορίθμου. Η μετατροπή λοιπόν των ακολουθιακών αλγορίθμων σε παράλληλους θα πρέπει να γίνεται με γνώμονα την μεγαλύτερη δυνατή ανεξαρτησία μεταξύ της εκτέλεσης σε κάθε νήμα, καθώς και της ελάχιστης ανάγκης ανταλλαγής δεδομένων μεταξύ των νημάτων. Η πιο αργή διαδικασία είναι αυτή που πάντα θα προσδιορίζει τον απαιτούμενο χρόνο [60].

1.5.4 Κατανεμημένα συστήματα

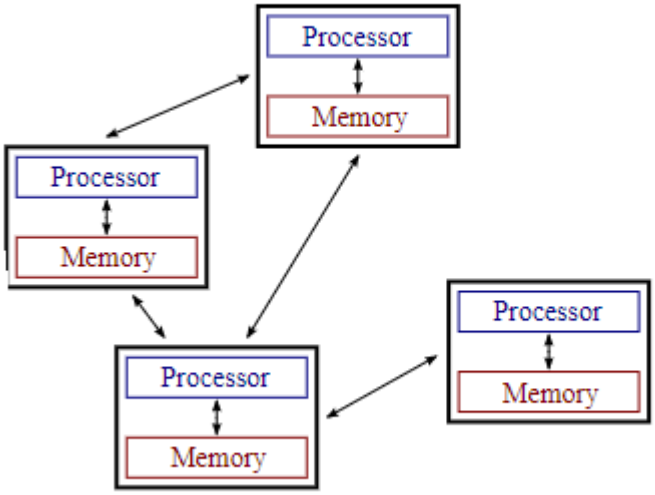
Στην προηγούμενη ενότητα αναφέρθηκε ως τύπος παραλληλισμού ο κατανεμημένος παραλληλισμός (distributed parallelism). Ως κατανεμημένο ορίζεται ένα σύστημα τα συστατικά του οποίου βρίσκονται σε διαφορετικές διαδικτυακές τοποθεσίες και επικοινωνούν συγχρονίζοντας την λειτουργίας τους μέσα από την ανταλλαγή μηνυμάτων μεταξύ τους αποβλέποντας στην ολοκλήρωση ενός κοινού στόχου.

Μια σημαντική διαφορά των κατανεμημένων και των παράλληλων συστημάτων είναι η πρόσβαση που έχουν οι μονάδες επεξεργασίας κάθε συστήματος στην μνήμη RAM. Στα παράλληλα συστήματα που αποτελούνται από διαφορετικές CPU εντός του ίδιου φυσικού μηχανήματος, οι πυρήνες έχουν πρόσβαση σε μια κοινή μνήμη μέσω της οποίας μπορούν να ανταλλάξουν γρήγορα μηνύματα.



Εικόνα 20 Ανταλλαγή μηνυμάτων σε παράλληλο σύστημα

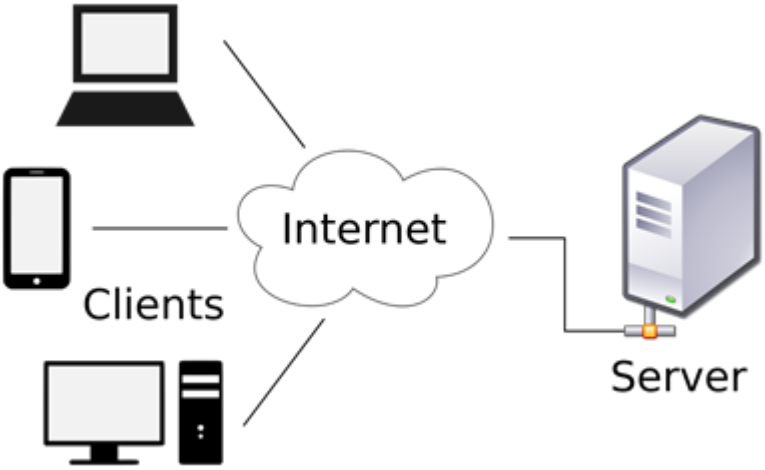
Αντίθετα στα κατανεμημένα συστήματα κάθε επεξεργαστής έχει πρόσβαση μόνο στην δική του μνήμη RAM και μπορεί να ανταλλάξει μηνύματα μόνο μέσω του πολύ πιο αργού δικτύου που συνδέει τα συστατικά του συστήματος μεταξύ τους. Φυσικά ακόμα και σε ένα κατανεμημένο σύστημα, τα επιμέρους συστήματα έχουν την δυνατότητα να τρέξουν εσωτερικά κάποιες διεργασίες παράλληλα με σκοπό την επιτάχυνσή τους [61].



Εικόνα 21 Ανταλλαγή μηνυμάτων σε κατανεμημένο σύστημα

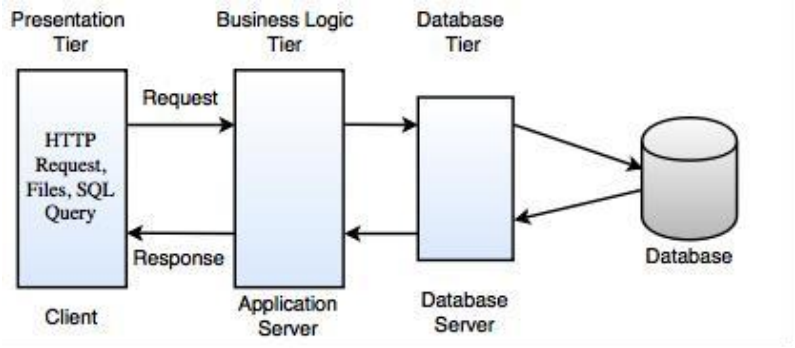
Υπάρχουν αρκετές διαφορετικές αρχιτεκτονικές για τον σχεδιασμό ενός κατανεμημένου συστήματος. Μερικές από τις πιο διαδεδομένες είναι οι:

- **Διακομιστής – Πελάτης (Client – Server):** Στην συγκεκριμένη αρχιτεκτονική γίνεται διαχωρισμός μεταξύ των διεργασιών που παρέχει ο προμηθευτής (server) μιας υπηρεσίας και αυτού ο οποίος την ζητάει (client) [62]



Εικόνα 22 Παράδειγμα αρχιτεκτονικής Client – Server

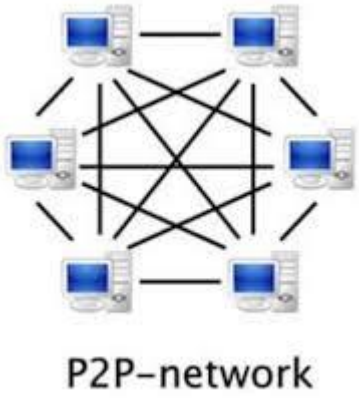
- **Πολυεπίπεδη αρχιτεκτονική (N-Tier):** Βασίζεται στην προαναφερθείσα αρχιτεκτονική πελάτη – διακομιστή όπου κάθε επίπεδο ορίζεται να επιτελεί ένα συγκεκριμένο μέρος των εντολών που μπορεί να έχει ένας πελάτης. Τα επίπεδα επικοινωνούν μεταξύ τους με σαφώς καθορισμένες διεπαφές. Κάθε επίπεδο μπορεί να επικοινωνήσει μόνο με τους άμεσους γείτονές του. Συνήθως χρησιμοποιείται για να απομονώσει μεταξύ τους λειτουργίες όπως η Διεπαφή χρήστη (Presentation Tier), η πρόσβαση στην Βάση Δεδομένων (Database Tier), η πρόσβαση στους επιχειρησιακούς κανόνες (Business Tier) [63]



Εικόνα 23 Παράδειγμα αρχιτεκτονικής Πολλαπλών Επιπέδων (N-Tier)

- **Δίκτυο ομότιμων κόμβων (Peer to Peer):** Κάθε κόμβος σε ένα τέτοιο κατανεμημένο σύστημα μπορεί να λειτουργήσει και να διαμοιραστεί αρχεία ή άλλους

πόρους χωρίς την ανάγκη ύπαρξης ενός κεντρικού διακομιστή. Ένας κόμβος εντός του συστήματος μπορεί να συνδεθεί με περισσότερους από έναν κόμβους, δημιουργώντας έτσι ένα πολύπλοκο δίκτυο κόμβων που εξασφαλίζει την βιωσιμότητα του δικτύου απέναντι στην κατάρρευση αρκετών από αυτούς [64].



Εικόνα 24 Παράδειγμα αρχιτεκτονικής ομότιμων λ κόμβων (Peer to Peer)

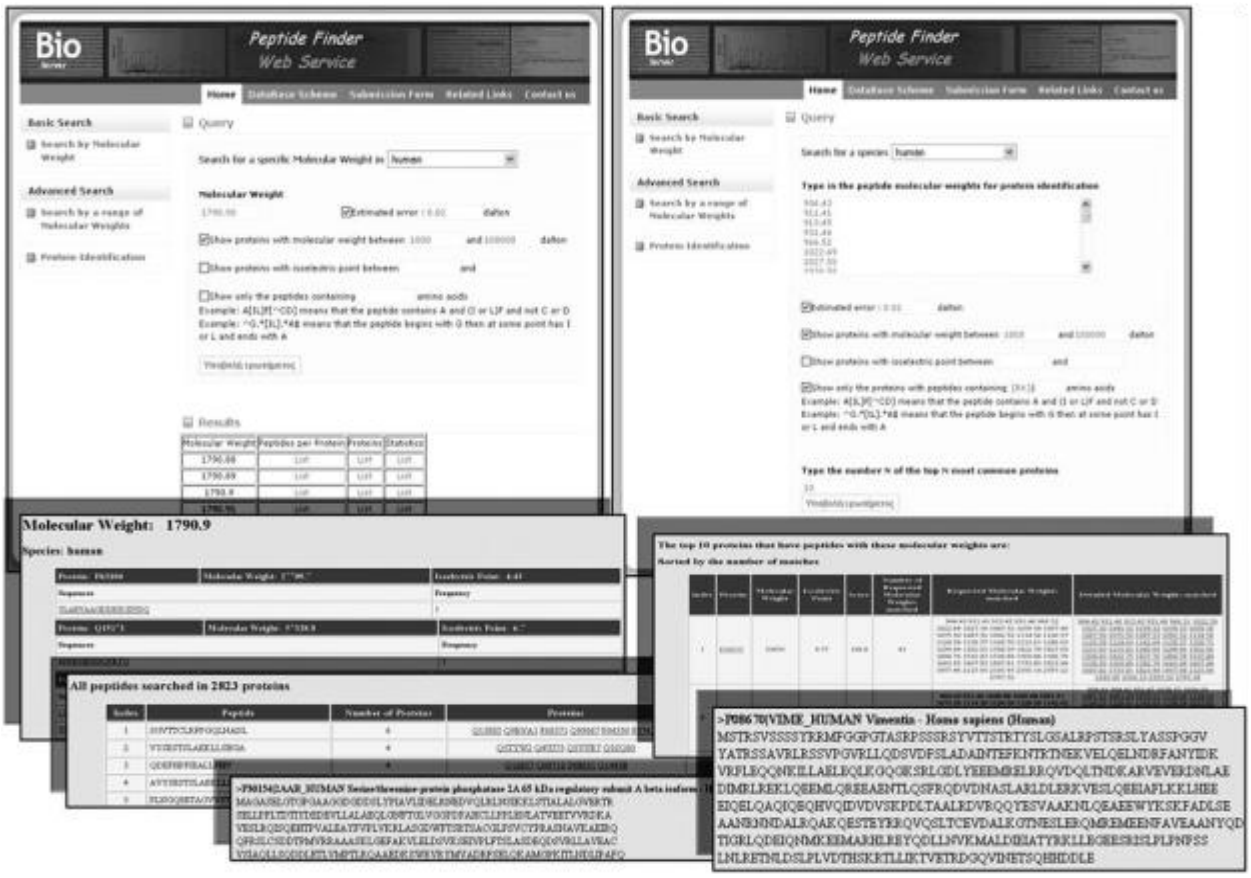
Ανεξάρτητα με την επιλεχθείσα αρχιτεκτονική ο σχεδιασμός ενός κατανεμημένου συστήματος πρέπει να γίνεται με γνώμονα την μεγαλύτερη δυνατή ανεξαρτησία κάθε συστατικού μέσα στο σύστημα, καθώς και την ελάχιστη δυνατή επικοινωνία μεταξύ τους. Ένα κατανεμημένο σύστημα θα πρέπει να έχει δικλείδες ώστε να μπορεί να συνεχίσει την επίλυση του προβλήματος που του έχει ανατεθεί ακόμη και αν κάποιο από τα συστατικά του βγει εκτός λειτουργίας

Τα κατανεμημένα συστήματα σήμερα είναι ευρέως διαδεδομένα λόγω της πληθώρας δεδομένων που συγκεντρώνονται στους σύγχρονους διακομιστές και πρέπει να αναλυθούν. Η ημερήσια παραγωγή δεδομένων το 2017 υπολογίζεται σε περίπου 2.5 exabytes [65]. Η μεταφορά δεδομένων πολλών Terabyte (ίσως και Petabyte) σε κοινά μηχανήματα για την επεξεργασία τους είναι απαγορευτική και αυτό οδήγησε στην μεγάλη αποδοχή των κατανεμημένων συστημάτων από τον κόσμο της πληροφορικής. Η μεταφορά του αλγορίθμου υπολογισμού μεταξύ των συστατικών είναι πολύ φτηνότερη από την μεταφορά των δεδομένων. Ο πολύ πετυχημένος αλγόριθμος Map / Reduce είναι ένα εξαιρετικό παράδειγμα [66].

Μεγάλο προτέρημα των κατανεμημένων συστημάτων είναι επίσης η σχετικά εύκολη επεκτασιμότητα που παρέχουν. Αρκεί η προσθήκη νέων μηχανημάτων στο κατανεμημένο σύστημα για να επεκταθεί η συνολική επεξεργαστική και αποθηκευτική του ισχύς . Τέλος σημαντικό πλεονέκτημά τους είναι και η δυνατότητα να επεκταθούν από ετερόκλητα μεταξύ τους συστατικά μηχανήματα.

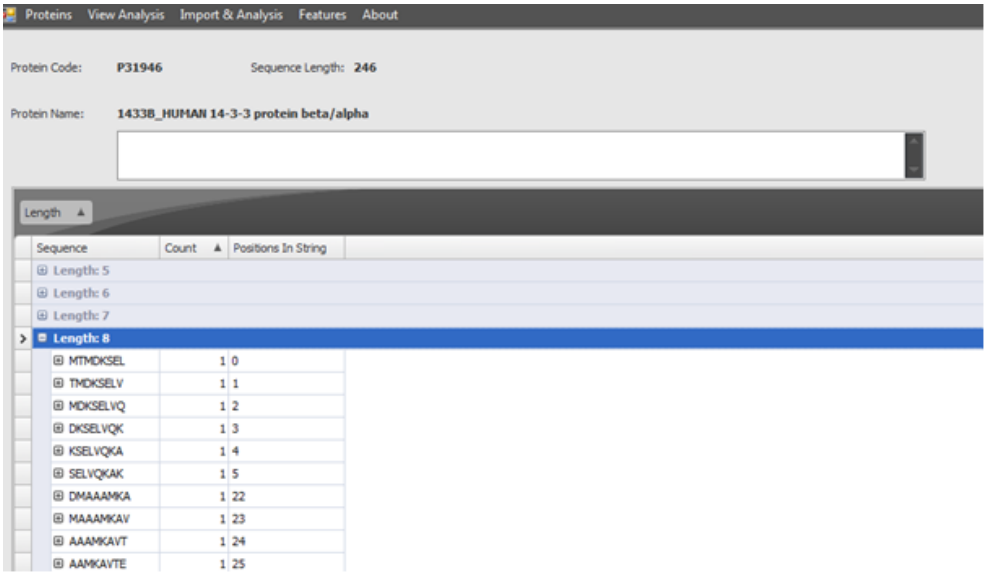
1.6 Εφαρμογές αλγορίθμων εύρεσης μοναδικών πεπτιδίων

Προηγούμενες έρευνες έχουν ορίσει τις έννοιες του Core Unique Peptide και του Composite Unique Peptide [67-69]. Η πρώτη προσπάθεια ανάπτυξης μιας εφαρμογής που θα μπορούσε να εντοπίσει τα Core Unique πεπτίδια ενός πρωτεώματος ήταν το UniMap το οποίο υλοποιήθηκε σε CGI Perl και MySQL και ολοκληρώθηκε το 2009 [68]. Το UniMap αποτελεί μέρος μια ομάδας εργαλείων και βάσεων δεδομένων που αναπτύχθηκαν στο IIBEAA και λειτουργούν σε ένα Linux διακομιστή με προσβάσιμες από το διαδίκτυο υπηρεσίες που ονομάζεται BioServer (Εικόνα 25) [67]. Το UniMap έχει στηριχθεί στην ανάλυση του ανθρώπινου πρωτεώματος και μπορεί να ενημερώνει την βάση του κάθε φορά που ανανεώνεται η Swiss-Prot από την οποία αντλεί τα δεδομένα του. Επίσης παρέχει ένα Web Interface το οποίο μπορούν να προσπελάσουν οι χρήστες και να αναζητήσουν πεπτίδια για την μοναδικότητά τους εντός του ανθρώπινου πρωτεώματος. Ο ορισμός των μοναδικών πεπτιδίων σε αυτές τις μελέτες αξιολογήθηκε βάση το μοριακό βάρος των πεπτιδίων και όχι της αμινοξικής τους αλληλουχίας. Αυτή όμως η προσέγγιση είχε προβλήματα στο διαχωρισμό των πεπτιδίων που είχαν στο σύνολο τα ίδια αμινοξέα (ίδιο μοριακό βάρος) σε διαφορετική θέση μέσα στο πεπτίδιο (π.χ. STELA και STEAL).

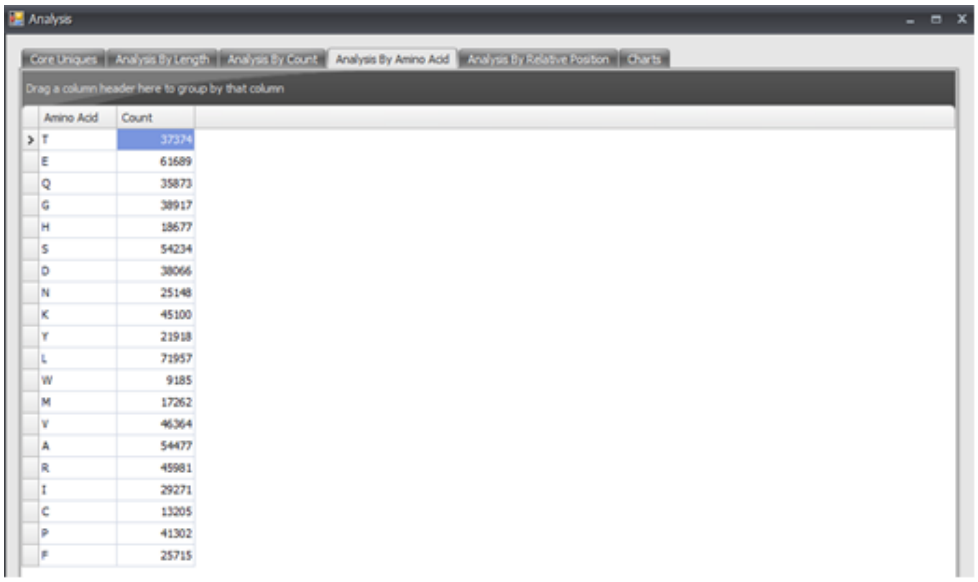


Εικόνα 25 Web Interface του BioServer

Το 2015 ο αλγόριθμος του UniMap υλοποιήθηκε εκ νέου στην προσπάθεια της ομάδας να εντοπίσει τα Composite Unique peptides αλλά και να αναλύσει τα ποιοτικά χαρακτηριστικά τόσο των Core όσο και τον Composite Unique peptides εντός του ανθρώπινου πρωτεώματος [70,71]. Η νέα εφαρμογή ονομάστηκε **Protein Analysis** και υλοποιήθηκε σε Microsoft .NET 4 με γλώσσα C# ενώ τα αποτελέσματα αποθηκεύτηκαν σε Microsoft SQL Server. Ο χρήστης μπορούσε να αποθηκεύσει διαφορετικές εκδόσεις του πρωτεώματος και κατ' επέκτασιν και του Uniquome. Επιπλέον η νέα εφαρμογή επέτρεπε την ελεύθερη επιλογή σχεσιακού συστήματος βάσεων δεδομένων (RDBMS). Τέλος μπορούσε να εκτελεστεί σε Microsoft Windows διευκολύνοντας έτσι την χρήση της από χρήστες μη εξοικειωμένους με το λειτουργικό σύστημα Linux.



Εικόνα 26 Windows Based Interface για την εφαρμογή Protein Analysis



Εικόνα 27 Ανάλυση CrUP ανά Αμινοξύ στην εφαρμογή Protein Analysis

2. Σκοπός Διδακτορικής Διατριβής

Οι πιο διαδεδομένες μέθοδοι για την ταυτοποίηση των πρωτεϊνών είναι αυτές που αξιοποιούν το πεπτιδικό αποτύπωμα των πρωτεϊνών (peptide finger-print) και αναλύουν την αμινοξική αλληλουχία των πεπτιδίων τους. Τα σημαντικότερα μειονεκτήματα αυτών των μεθόδων είναι πως για την ασφαλή ταυτοποίηση μίας πρωτεΐνης, απαιτείται η ανάλυση τουλάχιστον δύο πεπτιδίων ανά πρωτεΐνη καθώς και ότι πολλά από τα πεπτίδια που ταυτοποιούνται από το φασματογράφο μάζας δεν οδηγούν τελικά σε ασφαλή χαρακτηρισμό μίας πρωτεΐνης και απορρίπτονται κατά τη βιοπληροφορική επεξεργασία. Κύριος σκοπός της παρούσας διατριβής είναι η δημιουργία μιας νέας προσέγγισης για την ταυτοποίηση των πρωτεϊνών ενός οργανισμού. Η υπόθεση που ακολουθήθηκε για την επίτευξη αυτού του σκοπού είναι ότι η αμινοξική αλληλουχία κάθε πρωτεΐνης θα πρέπει να περιλαμβάνει τουλάχιστον ένα πεπτίδιο που η αμινοξική του αλληλουχία είναι απόλυτα μοναδική (Unique) ως προς τον οργανισμό που ανήκει, με αποτέλεσμα να την καθιστά την πρωτεΐνη διακριτή έναντι κάθε άλλης πρωτεΐνης σε ένα πρωτέωμα. Η προσέγγιση αυτή οδήγησε στην καταγραφή των μοναδικών πεπτιδίων του ανθρώπου στο σύνολο των θεωρημένων πρωτεϊνών αναδεικνύοντας δύο νέες οντότητες μοναδικών πεπτιδίων, τα μοναδικά πεπτίδια ελαχίστου μήκους (core unique peptide - CrUP) και τα σύνθετα μοναδικά πεπτίδια (composite unique peptide - CmUP). Τέλος, για τους σκοπούς της παρούσας διατριβής εισήχθη για πρώτη φορά ο όρος του Uniquome που περιλαμβάνει το σύνολο των μοναδικών πεπτιδίων (ελάχιστου μήκους και σύνθετων πεπτιδίων) ενός οργανισμού.

Η ανάλυση του Uniquome πέρα από την ανάγκη της δημιουργίας μιας νέας προσέγγισης με σκοπό την αύξηση του ποσοστού των ταυτοποιημένων πρωτεϊνών ενός υπό μελέτη δείγματος, επεκτάθηκε και σε άλλες εφαρμογές με σκοπό τόσο τη δημιουργία νέων θεραπευτικών προσεγγίσεων για την αντιμετώπιση παθολογικών καταστάσεων όσο και την δημιουργία προσεγγίσεων μέσω των οποίων θα μπορεί να προβλεφθεί η δράση των παθογόνων μικροοργανισμών. Για τους παραπάνω σκοπούς, η παρούσα διατριβή περιλαμβάνει **α)** Την ανάπτυξη μεθοδολογίας ανάλυσης μεγάλων δεδομένων (big data analysis) για την δημιουργία του ανθρώπινου Uniquome , **β)** την κατάρτιση και πλήρη καταγραφή του ανθρώπινου Uniquome που περιλαμβάνει τόσο τα CrUPs όσο και τα CmUPs, **γ)** την ανάλυση και την διερεύνηση των χαρακτηριστικών τους σε ένα υψηλά συστημικό και συνθετικό επίπεδο και **δ)** την διερεύνηση εφαρμογών του ανθρώπινου Uniquome σε φυσιολογικές και παθολογικές καταστάσεις.

3. Υλικά και μέθοδοι

3.1 Βάσεις Δεδομένων

Χρησιμοποιήθηκαν οι πληροφορίες και τα εργαλεία από τις βάσεις δεδομένων:

- Uniprot (www.uniprot.org) [44-46]
  - Απομόνωση των πρωτεωμάτων με τις θεωρημένες πρωτεΐνες (αρχεία fasta)
  - Αλγόριθμοι 'blast' για αναζήτηση αμινοξικών αλληλουχιών [72]
  - Αλγόριθμοι ' alignment' για την αμινοξική ευθυγράμμιση αλληλουχιών [73]
- Unipept (unipept.ugent.be) [74]
  - Χρήση αλγορίθμων για την προσομοίωση της επώασης των πρωτεϊνών με θρυψίνη [75]
- Iedb (www.iedb.org) [76]
  - Απομόνωση των επιτοπικών ανοσοπεπτιδίων [77]
- Caped (www.caped.icp.ucl.ac.be)
  - Απομόνωση των καρκινικών αντιγονικών πεπτιδίων [78]
- Ncbi (www.ncbi.nlm.nih.gov) [79]
  - Αλγόριθμοι 'blast' για αναζήτηση νουκλεοτιδικών αλληλουχιών [80]

3.2 Δημιουργία νέου αλγορίθμου για την κατασκευή των Uniquomes

Για τον επαναπροσδιορισμό της κατασκευής του Uniquome του ανθρώπινου πρωτεώματος, ορίζοντας την μοναδικότητα των πεπτιδίων βάση την αμινοξική τους αλληλουχία, καθώς και την επέκταση αυτού σε άλλους οργανισμούς, δημιουργήθηκε ένας νέος αλγόριθμος με τον οποίο ξεπεράσθηκαν τα προβλήματα και οι δυσκολίες των προηγούμενων υλοποιήσεων. Η δημιουργία του νέου και επεκτάσιμου αλγορίθμου, ο οποίος μπορεί να εκτελείται παράλληλα αλλά και κατανεμημένα, έδωσε τη δυνατότητα για την κατασκευή του Uniquome (χρησιμοποιώντας δεδομένα από την βάση δεδομένων Uniprot version 10/2019) ενός οργανισμού ή συστήματος. Επιπλέον, ο νέος αλγόριθμος έχει την δυνατότητα προαιρετικού καθορισμού υπό-ομάδων του δοθέντος πρωτεώματος ώστε να μπορεί να βρίσκει μοναδικότητες εντός αυτών (π.χ. οικογένειες πρωτεϊνών).

Η υλοποίηση του παραπάνω αλγορίθμου έγινε σε ένα σύγχρονο υπολογιστικό σύστημα ώστε να μπορεί να εκμεταλλεύεται με τον καλύτερο δυνατό τρόπο το υποκείμενο υλικό λογισμικό (software / hardware) του υπολογιστή στον οποίο εκτελείται. Το νέο αυτό σύστημα ονομάστηκε Uniquome Analysis.

Για το Uniquome Analysis ορίσθηκαν οι παρακάτω τεχνικές προδιαγραφές:

- Δυνατότητα να εκτελείται σε όλα τα σύγχρονα υπολογιστικά συστήματα από SoC όπως το RaspberryPI, μέχρι μεγάλα συστήματα διακομιστών με πολλαπλούς επεξεργαστές.
- Δυνατότητα να εκτελείται και στα τρία κύρια σύγχρονα λειτουργικά συστήματα Windows, Linux και MacOS με τον ίδιο τρόπο.
- Δυνατότητα επέκτασης της εκτέλεσης σε περισσότερα από ένα φυσικά ή λογικά μηχανήματα (scalability).
- Εύκολη εγκατάσταση και παραμετροποίηση και χρήση χωρίς την ανάγκη ύπαρξης τρίτων συστημάτων (π.χ. RDBMS).
- Δυνατότητα επέκτασης της λειτουργικότητας της εφαρμογής με χρήση πρόσθετων αρχείων χωρίς να είναι απαραίτητη η επανεγκατάσταση της εφαρμογής (add-ons).
- Δυνατότητα χρήσης της εφαρμογής από την γραμμή τερματικού (terminal) αλλά και ανάπτυξη γραφικής διεπαφής (User Interface).
- Αποθήκευση των αποτελεσμάτων σε αρχεία κειμένου για την εύκολη μεταφορά αλλά και χρήση τους από οποιοδήποτε υπολογιστή που μπορεί να επεξεργαστεί κείμενο.

Επιπλέον ορίσαμε τις παρακάτω λειτουργίες:

- Δυνατότητα στον χρήστη να ορίζει το σύστημα ή τον οργανισμό που θέλει να αναλύσει κατασκευάζοντας το Uniquome του.
- Δυνατότητα ορισμού υπό-ομάδων εντός ενός πρωτεώματος για την ανεύρεση του Uniquome με ελάχιστη μονάδα αναζήτησης την υπό-ομάδα και όχι την πρωτεΐνη.
- Δυνατότητα σύγκρισης μεταξύ δύο ή περισσοτέρων Uniquome διαφορετικών οργανισμών για την ανεύρεση κοινών πεπτιδίων.

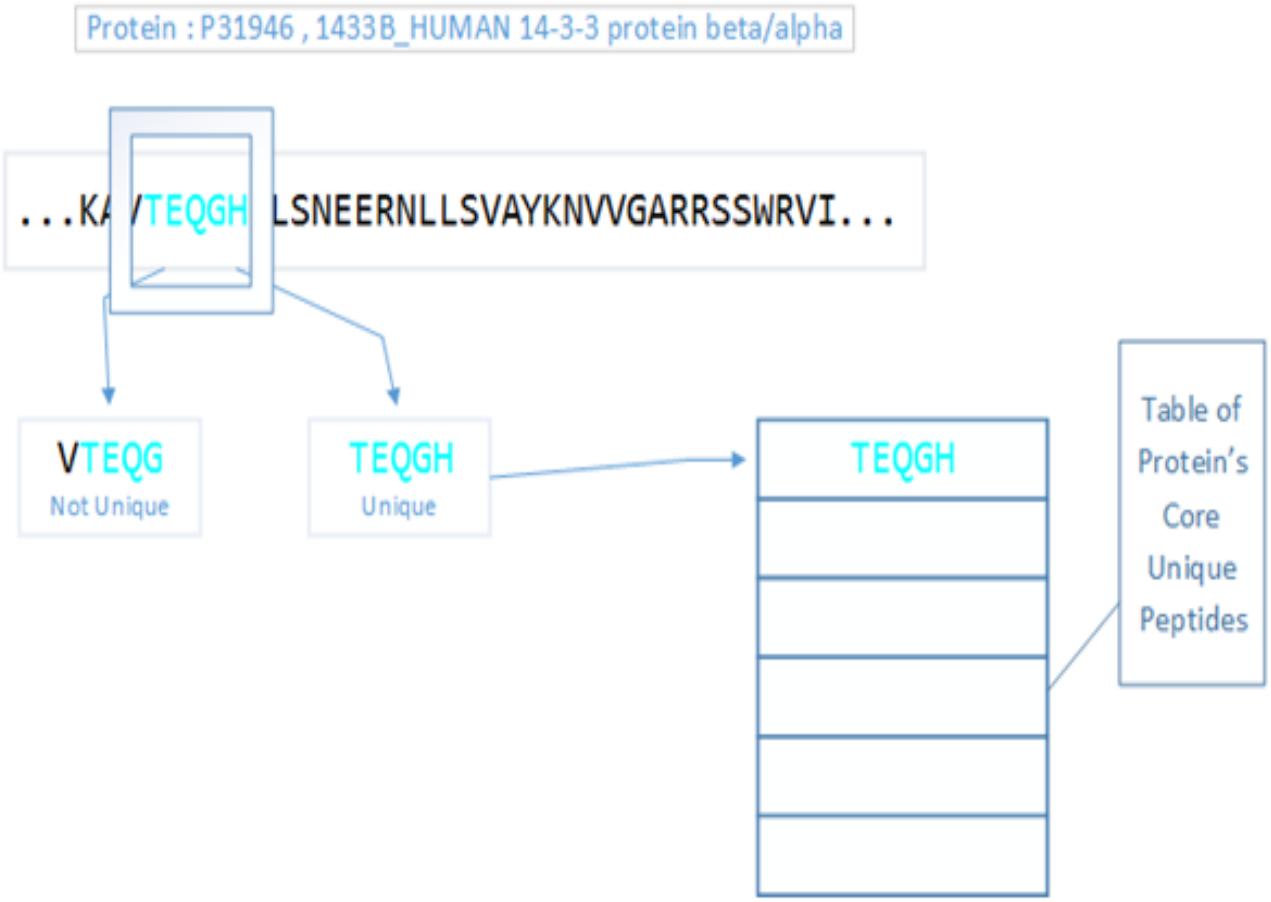
3.3 Ακολουθιακός Αλγόριθμος

Αρχικά ο αλγόριθμος που σχεδιάστηκε και υλοποιήθηκε στο UniMap και το Protein Analysis ήταν σειριακός και μπορεί να περιγράφει από το παρακάτω κομμάτι ψευδοκώδικα και τις εντολές:

- Για κάθε πρωτεΐνη P σε ένα σύνολο πρωτεϊνών
    - Για K από X μέχρι Y
        - Δημιούργησε ένα κυλιόμενο παράθυρο μεγέθους K και ολίσθησέ το στην αλληλουχία της πρωτεΐνης P
            - Πάρε το πεπτίδιο C κάτω από το παράθυρο. Αν το πεπτίδιο C περιέχει ήδη ένα Core Unique peptide τότε σταμάτα την αναζήτηση.
            - Για κάθε πρωτεΐνη M στο ίδιο σύνολο πρωτεϊνών εκτός της P
                - Αν το πεπτίδιο C υπάρχει στην πρωτεΐνη Mf τότε σταμάτα την αναζήτηση.
            - Αν το πεπτίδιο C δεν υπάρχει σε καμία άλλη πρωτεΐνη τότε είναι Core Unique peptide και αποθήκευσε το.

Στην Εικόνα 28 παρουσιάζεται μια οπτική εκτέλεση του αλγορίθμου όπως περιγράφηκε.

Ο αλγόριθμος δημιουργεί παράθυρα μεγέθους M μεταξύ δυο προκαθορισμένων μηκών X και Y όπου συνήθεις τιμές αυτών είναι το X= 4, Y=100. Στην συνέχεια για κάθε παράθυρο εξάγει ένα κατάλληλο πεπτίδιο και ελέγχει αν αυτό περιέχει μέσα του κάποιο άλλο Core Unique Peptide. Αν δεν περιέχει τότε αναζητεί στις υπόλοιπες πρωτεΐνες αν υπάρχει ή όχι το πεπτίδιο. Σε περίπτωση που το βρει σταματάει την αναζήτηση για το συγκεκριμένο πεπτίδιο. Αν αυτό δεν βρεθεί θεωρείται Core Unique και αποθηκεύεται.



Εικόνα 28 Προσπέλαση πρωτεΐνης για ανεύρεση Core Unique Peptide

3.4 Βελτιστοποίηση λειτουργιών εντός του ίδιου νήματος εκτέλεσης (single thread optimizations)

3.4.1 Βελτιστοποίηση αναζήτησης Πεπτιδίου σε πρωτεΐνη

Όπως προκύπτει από τον ακολουθιακό αλγόριθμο που περιγράφηκε πριν, μια από τις βασικότερες λειτουργίες που πρέπει να εκτελεστεί είναι η αναζήτηση ενός πεπτιδίου μέσα σε μία πρωτεΐνη. Ασφαλής απάντηση για την ύπαρξη του ή όχι μπορεί να δοθεί μόνο όταν αναζητηθεί κατά μήκος όλης της αλληλουχίας της πρωτεΐνης. Για να γίνει καλύτερα κατανοητό το μέγεθος του προβλήματος ας υποθέσουμε ότι αναζητούμε το 5-πεπτίδιο **ELLLK** στην TITIN που έχει μέγεθος 34.350 αμινοξέων και η οποία δεν το περιέχει. Για να μπορέσει ο αλγόριθμος αναζήτησης να μας επιστρέψει αυτή την απάντηση, θα πρέπει να συγκρίνει το ELLK κατά μήκος της αλληλουχίας ξεκινώντας από την θέση 1 μέχρι την θέση 34.347. Θα εκτελέσει λοιπόν τουλάχιστον 34.347 συγκρίσεις. Φυσικά το πλήθος των συγκρίσεων θα είναι πολύ μεγαλύτερο αφού για κάθε γράμμα (αμινοξύ) του πεπτιδίου που ταιριάζει με το σημείο της αλληλουχίας που ελέγχει θα εκτελεί μία ακόμη σύγκριση όπως φαίνεται στο παρακάτω παράδειγμα από την αναζήτηση σε τμήμα της συγκεκριμένης πρωτεΐνης:

VAADKAKEQELKSRTKEVITTKQEQMHVTHEQIRKETEKTFVPKVVISAAKAKEQET
ELLK

Για να αποκλειστούν τα αμινοξέα με την πορτοκαλί επισήμανση (VAADKAK) στην παραπάνω αλληλουχία ως σημεία πιθανής έναρξης του πεπτιδίου ELLK απαιτείται μία σύγκριση για το καθένα, άρα 7 συγκρίσεις. Στην συνέχεια το επόμενο τμήμα EQ απαιτούνται 3 συγκρίσεις μια για το E που ταιριάζει με την πρώτη θέση του μια για να αποκλειστεί το Q που δεν ταιριάζει με το δεύτερο αμινοξύ L και μία για να αποκλειστεί το Q μετά την ολίσθηση του ELLK κατά μία θέση δεξιά στην αλληλουχία ώστε να συγκριθεί το E με το Q. Κατ' αντιστοιχία για να αποκλειστεί το ELK απαιτούνται 4. Γίνεται εύκολα αντιληπτό το πολύ μεγάλο υπολογιστικό κόστος που απαιτείται για συγκρίσεις που ξέρουμε ότι δεν μπορούν να καταλήξουν σε ταίριασμα του πεπτιδίου με την αλληλουχία αφού δεν ξεκινούν με το ίδιο γράμμα με το προς αναζήτηση πεπτίδιο.

Το παραπάνω πρόβλημα αντιμετωπίσθηκε εν μέρει σαν πρόβλημα αναζήτησης λέξης σε λεξικό. Όταν αναζητάμε για μια λέξη η πρώτη κίνηση είναι να πάμε στο σημείο όπου ξεκινούν οι λέξεις με ίδιο πρώτο γράμμα όπως η προς αναζήτηση λέξη. Με βάση αυτή την παρατήρηση το πρώτο βήμα του νέου αλγορίθμου είναι να προ-επεξεργαστεί όλες τις πρωτεΐνες χαρτογραφώντας τις θέσεις που περιέχεται κάθε αμινοξύ. Δημιουργούμε έτσι ένα πίνακα όπως ο παρακάτω που αφορά την πρωτεΐνη P0AAW9 του οργανισμού *E. coli* με αλληλουχία: MLELLKSLVFAVIMVPVVMAIILGLIYGLGEVFNIFSGVGKKDQPGQNH

Στον πίνακα στην πρώτη στήλη τοποθετούνται όλα τα πιθανά αμινοξέα της αλληλουχίας, και σε κάθε γραμμή την θέση μέσα στην αμινοξική αλληλουχία που εμφανίζονται.

M	**1**	**14**	**19**				
L	2	4	5	8	23	25	29
E	3	31					
S	7	36					
K	6	41	42				
V	9	12	15	17	18		

Εικόνα 29 Τμήμα χαρτογράφησης αμινοξέων πρωτεΐνης P0AAW9

Με τον παραπάνω χάρτη τροποποιήθηκε το μέρος του αλγορίθμου που αναζητά το πεπτίδιο στην πρωτεΐνη ως εξής:

- Εντόπισε την λίστα των θέσεων που αντιστοιχούν στο πρώτο αμινοξύ του προς αναζήτηση πεπτιδίου.

- Για κάθε θέση από τη λίστα των θέσεων του αμινοξέος έλεγξε τα αμινοξέα που ξεκινούν από την θέση + 1 σε σχέση με τα αμινοξέα του πεπτιδίου μετά το πρώτο. Το πρώτο αμινοξύ δεν χρειάζεται να ελεγχθεί αφού η ταυτοποίησή του έχει επιτευχθεί μέσω του χάρτη. Επομένως για ένα πεπτίδιο μεγέθους L και για κάθε θέση του χάρτη θα γίνουν το πολύ L-1 συγκρίσεις.

Στο προηγούμενο παράδειγμα που παρατέθηκε για να αποκλειστεί η ύπαρξη του **ELLK** εντός του τμήματος της πρωτεΐνης αρκούν 10 συγκρίσεις. Μία για κάθε αμινοξύ Ε εντός της αλληλουχίας, εκτός του τμήματος EL όπου χρειάζονται 2 αφού το L είναι κοινό μεταξύ του τμήματος και του πεπτιδίου. Για το ίδιο τμήμα ο προηγούμενος αλγόριθμος χρειαζόταν να εκτελέσει 77 συγκρίσεις.

Για κάθε πρωτεΐνη ο εντοπισμός της θέσης του προς αναζήτηση αμινοξέος στον χάρτη έχει πολυπλοκότητα O (1). Αυτό επιτυγχάνεται καταναλώνοντας ένα μικρό ποσό μνήμης χρησιμοποιώντας όλα τα γράμματα της αγγλικής αλφαβήτου, ακόμη και αυτά που δεν αντιστοιχούν σε αμινοξέα, ενώ παράλληλα θεωρούμε ως συνάρτηση εύρεσης της θέσης στον πίνακα την f(γράμμα) = γράμμα – 64. Ο αριθμός 64 προκύπτει από την δεκαδική αναπαράσταση του χαρακτήρα πριν το γράμμα Α στον ASCII πίνακα. Έτσι όταν αναζητήσουμε τον χάρτη για το αμινοξύ Q αντί να εκτελέσουμε μια επανάληψη εντός του

χάρτη μέχρι να βρούμε το σημείο που περιέχει τις θέσεις του στην αλληλουχία εκτελούμε την συνάρτηση f(Q) = 81 – 64 = 17.

3.4.2 Αποκλεισμός πεπτιδίων από την αναζήτηση αν είναι ή όχι CrUP

Όπως περιγράφηκε στον ακολουθιακό αλγόριθμο για να μπορέσουμε να εντοπίσουμε τα CrUP ξεκινούμε από ένα παράθυρο μεγέθους X (συνήθως 4) και συνεχίζουμε την αναζήτησή μας αυξάνοντας κατά 1 αμινοξύ το μέγεθος του παραθύρου. Ο λόγος για τον οποίο η αναζήτηση γίνεται με το μικρότερο δυνατό πεπτίδιο είναι για να αποκλειστούν από το Uniquome, πεπτίδια τα οποία περιέχουν μικρότερα Core Unique Peptides εντός τους. Αυτό επιτυγχάνεται ελέγχοντας όλα τα Core Unique Peptides της υπό μελέτη πρωτεΐνης που έχουν μέγεθος μικρότερο ή ίσο με το πεπτίδιο που ελέγχεται, ως προς τη μοναδικότητα του, για το αν κάποιο από αυτά εμπεριέχεται μέσα του. Για την αναζήτηση κάθε CrUP εντός του νέου πεπτιδίου ισχύουν τα προβλήματα που αναφέρθηκαν στην προηγούμενη ενότητα, χωρίς όμως να είναι εφικτή η δημιουργία ενός μίνι χάρτη αμινοξέων για κάθε CrUP που εντοπίζεται αφού αυτό κρίνεται ασύμφορο τόσο σε υπολογιστικούς κύκλους όσο και στην μνήμη που απαιτείται. Η παραπάνω παρατήρηση μπορεί να λειτουργήσει θετικά αποκλείοντας μεγάλο μέρος των πεπτιδίων που θα πρέπει να ελεγχθούν για μοναδικότητα. Αν το παράθυρο που ελέγχουμε περιέχει ένα ή περισσότερα CrUP τότε ο αλγόριθμος μπορεί να εκτελέσει ένα «άλμα» ξεκινώντας την επόμενή του αναζήτηση από το δεύτερο αμινοξύ του τελευταίου CrUP που περιέχεται εντός του παραθύρου.

Ένα παράδειγμα αυτού φαίνεται στην εικόνα 30 όπου το προς έλεγχο παράθυρο είναι μεγέθους 14 αμινοξέων. Στο παράθυρο αυτό υπάρχουν ήδη τα πεπτίδια **ELLK** και **FAVIM** τα οποία είναι CrUP, οπότε οποιοδήποτε πεπτίδιο τα περιέχει δεν μπορεί να είναι επίσης CrUP. Ο αλγόριθμος μπορεί να παραβλέψει το κομμάτι της αλληλουχίας με το πράσινο πλαίσιο **MLELLKSLVF** και να ξεκινήσει την αναζήτηση αμέσως μετά από αυτό με πρώτο πεπτίδιο το **AVIMVPVVMAIILG**.

Εικόνα 30 Παράδειγμα αποκλεισμού πεπτιδίων από την αναζήτηση

3.5 Παράλληλη εκτέλεση λειτουργιών

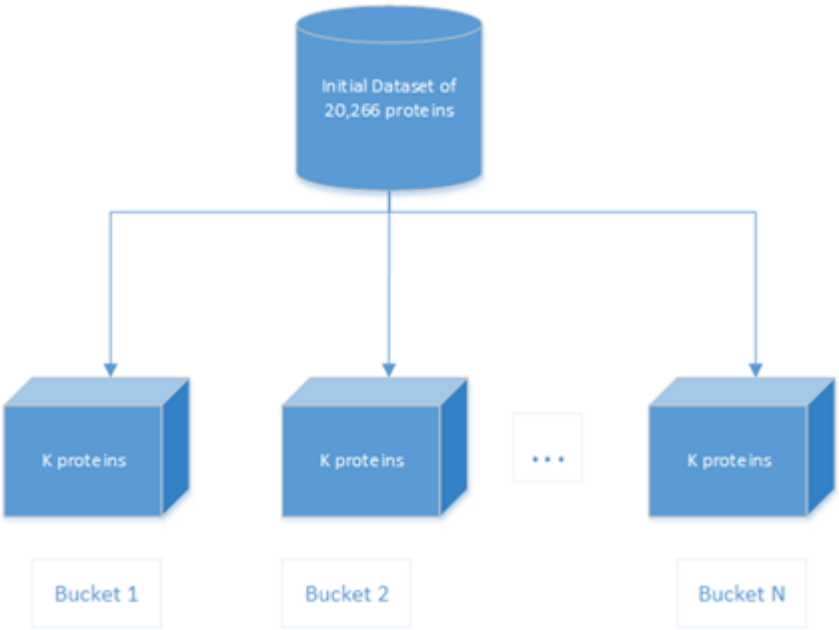
Μετά την ανάπτυξη των βελτιστοποιήσεων του αλγορίθμου εντός του ίδιου νήματος εκτέλεσης που αφορούν την αναζήτηση εντός μιας πρωτεΐνης καθώς και τον χωρισμό των πεπτιδίων που τελικά θα πρέπει να αναζητηθούν για ύπαρξη ή όχι στις άλλες πρωτεΐνες του

συνόλου. Περαιτέρω θα αναλυθούν οι βελτιστοποιήσεις που έγιναν στον αλγόριθμο ώστε να μπορεί να εκτελεστεί παράλληλα και να εκμεταλλευτεί τα σύγχρονα πολυπύρηνα συστήματα.

3.5.1 Χωρισμός του πρωτεώματος (βήμα 1)

Με το δεδομένο ότι η αναζήτηση ενός πεπτιδίου σε κάθε πρωτεΐνη είναι μια εντελώς ανεξάρτητη διαδικασία, αν θέλουμε να αναζητήσουμε ένα πεπτίδιο σε Χ πρωτεΐνες, η αναζήτηση του πεπτιδίου σε μία πρωτεΐνη δεν επηρεάζεται από την αναζήτηση του ίδιου πεπτιδίου και σε άλλες πρωτεΐνες. Αυτό που έχει σημασία είναι πως αν το πεπτίδιο εντοπιστεί σε μία πρωτεΐνη θα πρέπει να σταματήσουμε την περαιτέρω αναζήτηση γι' αυτό.

Παρατηρήθηκε ότι το πρωτέωμα μπορεί να χωρισθεί σε ομάδες και να ανατεθεί σε κάθε έναν επεξεργαστή η αναζήτηση ενός πεπτιδίου εντός μιας ομάδας πρωτεϊνών. Ονομάζουμε αυτά τα υποσύνολα δοχεία (bucket). Αν θεωρήσουμε πως τα διαφορετικά σύνολα έχουν περίπου το ίδιο μέγεθος αμινοξέων τότε έχουμε πετύχει μια καλή κατάτμηση του προβλήματος ευελπιστώντας σε βελτίωση ταχύτητας που θεωρητικά μπορεί να προσεγγίσει το N, όπου N είναι το πλήθος των δημιουργημένων υποομάδων (Εικόνα 31).



Εικόνα 31 Χωρισμός του πρωτεώματος σε N υποσύνολα

Η επιλογή του N μπορεί να επηρεάσει αρκετά την απόδοση του αλγορίθμου και θα πρέπει να επιλέγεται σε συνάρτηση με τους διαθέσιμους λογικούς πυρήνες του επεξεργαστή στον οποίο θα γίνει η εκτέλεση. Αν το N είναι αρκετά μεγάλο τότε το εκάστοτε υποσύνολο θα περιέχει ένα μικρό σχετικά πλήθος πρωτεϊνών που αυξάνει τις πιθανότητες να πετύχουμε νωρίς ένα πεπτίδιο. Για να γίνει καλύτερα κατανοητό τι εννοούμε με το τελευταίο, αν το προς αναζήτηση πεπτίδιο βρίσκεται στην θέση 15.000 του πρωτεώματος που αναλύουμε, τότε ο

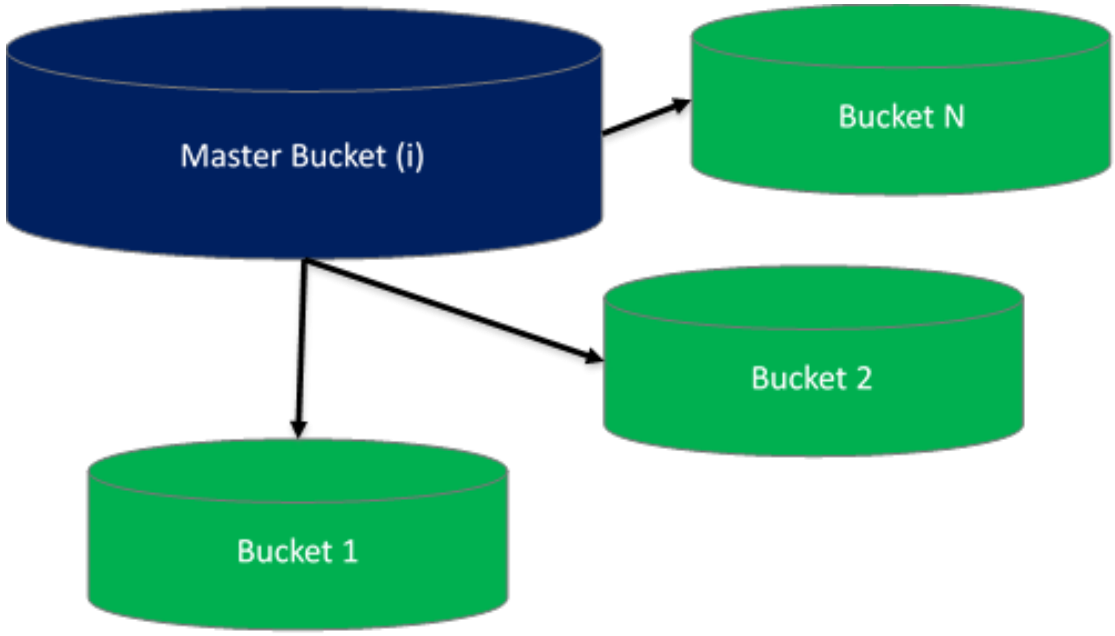
σειριακός αλγόριθμος θα πρέπει να ελέγξει 14.999 πρωτεΐνες μέχρι να βρει ότι υπάρχει πάλι το πεπτίδιο. Αν θέσουμε το N = 100 τότε κάθε κουβάς θα έχει να αναλύσει περίπου 200 πρωτεΐνες ανεβάζοντας κατά πολύ την πιθανότητα να βρεθεί πολύ σύντομα το πεπτίδιο, αφού θα είναι το πρώτο πεπτίδιο του κουβά 75. Όπως αναφέραμε όμως και η εναλλαγή του λειτουργικού μεταξύ των νημάτων έχει κόστος. Όσο μεγαλύτερο είναι το N από το πλήθος των λογικών πυρήνων του επεξεργαστή τόσο θα ανεβάζει το κόστος εναλλαγής των νημάτων για αναζήτηση στα υπόλοιπα δοχεία δεδομένων (buckets).

3.5.2 Εκτέλεση (βήμα 2)

Ένα από τα βασικότερα προβλήματα των παράλληλων εφαρμογών είναι η ανάγκη για προσπέλαση κοινών πόρων ανά επεξεργαστική μονάδα. Όποτε χρειάζεται να γίνει αυτό, η μονάδα που θέλει να προσπελάσει τον πόρο, πρέπει να κλειδώσει την πρόσβαση σε αυτόν ώστε να αποτρέψει την μη ορθή χρήση και την ενδεχόμενη φθορά των δεδομένων του. Το κλείδωμα αυτό όμως είναι ιδιαίτερα ακριβό σε πόρους συστήματος, ενώ, για την ώρα που το κλείδωμα είναι σε λειτουργία, ακυρώνεται η όποια παραλληλία της υλοποίησης αφού τα υπόλοιπα νήματα περιμένουν την ολοκλήρωση του κλειδώματος για να συνεχίσουν την λειτουργία τους.

Στην περίπτωση της εφαρμογής που αναπτύχθηκε στην παρούσα διατριβή, ο κοινός πόρος είναι η βάση των Core Unique Peptides. Υπήρχε η ανάγκη να βρεθεί ένας τρόπος ώστε να μηδενιστεί η πιθανότητα για κλείδωμα των όποιων πόρων, συνεπώς να μειωθεί και η ανάγκη για ύπαρξη κοινών πόρων. Για την επίτευξη των παραπάνω χρησιμοποιήθηκε μέρος της λογικής του αλγορίθμου MAP- REDUCE.

Όπως αναφέρθηκε παραπάνω, το σύνολο των πρωτεϊνών χωρίστηκε σε μικρότερα υποσύνολα που ονομάσαμε «δοχεία δεδομένων» (Buckets). Κάθε ένα από αυτά εκτός από το να αποτελεί αποθετήριο (repository) αλληλουχιών πρωτεϊνών, αποτελεί και μια ξεχωριστή υπολογιστική μονάδα. Η προσέγγιση που αναπτύχθηκε βασίστηκε στην λογική να μην εκτελεί κάθε δοχείο τον ίδιο ρόλο στην διάρκεια του υπολογισμού του Uniquome. Η ομάδα των διαφορετικών buckets προσπελαύνεται σειριακά και σε κάθε επανάληψη το τρέχον δοχείο θεωρείται ως Master και μόνο αυτό έχει πρόσβαση στην βάση δεδομένων και σε όσους πόρους απαιτούν να γίνει κάποιου είδους κλείδωμα. Τα υπόλοιπα δοχεία (Search Buckets) εκτελούν την ίδια λειτουργία μεταξύ τους δηλαδή την εύρεση ή όχι ενός πεπτιδίου στις πρωτεΐνες του αποθετηρίου τους, άρα δεν έχουν καμία αλληλεπίδραση μεταξύ τους. Η συγκεκριμένη αρχιτεκτονική μπορεί εύκολα να αντιπαραβληθεί με την Client-Server όπως αυτή περιγράφηκε στην ενότητα των κατανεμημένων συστημάτων, όπου ο Master Bucket έχει τον ρόλο του Client και οι Search Buckets τον ρόλο του Server.



Εικόνα 32 Αποστολή Περιεχομένου Ερώτησης από τον Master στους Search Buckets

3.5.3 Λειτουργία Master Bucket

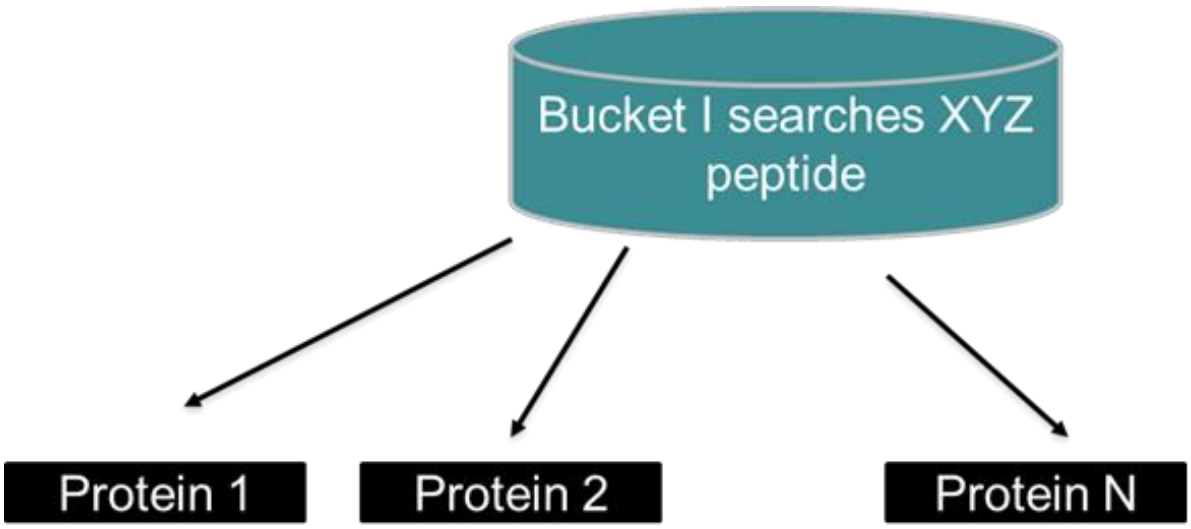
Ο ρόλος του Master Bucket είναι να επιτελέσει πολύ περισσότερα από τον ρόλο του Search Bucket. Εντός της υπολογιστικής μονάδας Master Bucket ακολουθείται ο σειριακός αλγόριθμος που περιγράφηκε νωρίτερα.
Συγκεκριμένα για κάθε πρωτεΐνη εντός του Master Bucket εκτελούνται οι παρακάτω διεργασίες:

- Επιλογή πεπτιδίου
- Αναζήτηση αν το πεπτίδιο περιέχει κάποιο Core Unique που έχει ήδη εντοπιστεί για την πρωτεΐνη (σε αυτή την περίπτωση είναι unique αλλά όχι Core Unique)
- Αποστολή πεπτιδίου προς αναζήτηση στους Search Buckets και αναμονή αποτελεσμάτων.
- Συγχρονισμός αναζήτησης των Search Buckets και παύση της αναζήτησής εντός αυτών όταν πρέπει.

Το Master Bucket παρακολουθεί την αναζήτηση των Search. Αν οποιοσδήποτε από αυτούς εντοπίσει το πεπτίδιο που αναζητά τότε στέλνει μήνυμα στον Master Bucket ότι βρέθηκε, ο οποίος με την σειρά του ενημερώνει τους υπόλοιπους Search Buckets να σταματήσουν την αναζήτηση του τρέχοντος πεπτιδίου και να περιμένουν για το επόμενο προς αναζήτηση. Η προσέγγιση αυτή επιτρέπει τον σύντομο τερματισμό της αναζήτησης

για ένα πεπτίδιο, από την στιγμή που βρεθεί πως δεν είναι Core Unique. Φυσικά ο μόνος τρόπος για να γίνει αποδεκτό ένα πεπτίδιο ως Core Unique Peptide είναι η εξαντλητική αναζήτηση στο σύνολο των πρωτεϊνών. Μπορεί εύκολα κάποιος να υποστηρίξει πως ο χρόνος εύρεσης ενός CrUP μειώνεται μόνο από τις βελτιστοποιήσεις που κάναμε εντός του single thread.

Ο σχεδιασμός του αλγορίθμου επιτρέπει την δημιουργία διαφορετικών ερωτήσεων στις οποίες ο Search Bucket μπορεί να απαντάει με NAI / OXI ή ακόμη και με πιο πολύπλοκες απαντήσεις όπως η αποστολή κατηγορηματικών δεδομένων ή ακόμη και αριθμητικών. Για να απαντηθεί το ερώτημα αν ένα πεπτίδιο είναι Core Unique εντός μιας οικογένειας πρωτεϊνών, τότε αρκεί σταματήσουμε την αναζήτηση για το πεπτίδιο εντός του Master Bucket. Άρα αν δεν βρεθεί σε άλλο δοχείο είναι μοναδικό σε αυτόν (τον Master Bucket) άρα και στην οικογένεια. Φυσικά απαραίτητη προϋπόθεση γι' αυτό είναι ο χωρισμός του πρωτεώματος σε δοχεία ανάλογα με την οικογένεια κάθε πρωτεΐνης.

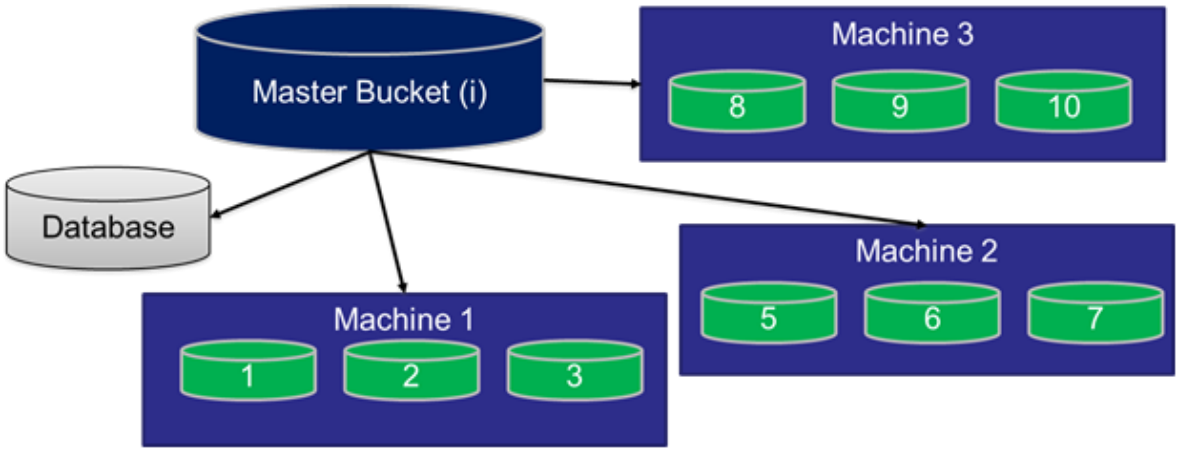


Εικόνα 33 Αναζήτηση πεπτιδίου εντός του Search Bucket

3.6 Καταμερισμένη εργασία σε πολλαπλά συστήματα

Η παράλληλη υλοποίηση του αλγορίθμου επιτρέπει την περαιτέρω βελτίωση της απόδοσης μέσω καταμερισμού της αναζήτησης σε περισσότερα τους ενός αυτόνομων συστημάτων. Όπως φαίνεται στην Εικόνα 34 μπορούμε να χωρίσουμε τα δοχεία του πρωτεώματος από το πρώτο βήμα του αλγορίθμου σε περισσότερα συστήματα. Η λογική παραμένει η ίδια όπως και στην παράλληλη έκδοση του αλγορίθμου. Μόνο ένα δοχείο μπορεί να έχει τον ρόλο του Master Bucket στην εκτέλεση. Το πρώτο στάδιο της αναζήτησης

αν το πεπτίδιο περιέχει ήδη κάποιο Core Unique peptide παραμένει επίσης το ίδιο. Όμως τώρα το Master Bucket στέλνει το προς αναζήτηση πεπτίδιο στους υπόλοιπους υπολογιστές που έχουν οριστεί. Εντός αυτών εκτελείται πάλι η ίδια παράλληλη διαδικασία με την αποστολή της απάντησης τον Master Bucket. Αυτή η επικοινωνία μπορεί να επιτευχθεί μέσω TCP/IP.



Εικόνα 34 Καταμερισμός αναζήτησης σε περισσότερα συστήματα

Ο υπολογιστής που εκκινεί την κατασκευή είναι υπεύθυνος για τον συντονισμό της διαδικασίας. Κατά την προεργασία και αφού χωριστούν οι πρωτεΐνες σε δοχεία, αποστέλλει σε κάθε δευτερεύων μηχάνημα ένα υποσύνολο δοχείων. Το σύστημα γνωρίζει την δυναμική κάθε επιμέρους υποσυστήματος οπότε ο χωρισμός του συνόλου των κουβάδων μπορεί να γίνει με βάση αυτή την γνώση, οπότε να στείλει στα λιγότερο δυνατά μηχανήματα (ή σε μηχανήματα που δεν πρέπει να κάνει χρήση του 100 % των πόρων) ένα μικρότερο υποσύνολο. Παράλληλα το ίδιο μηχάνημα έχει ένα πλήρες αντίγραφο όλων των κουβάδων του συστήματος ώστε να μπορεί να φέρνει τον καθένα στον ρόλο master. Το ίδιο δεν είναι απαραίτητο να κρατήσει κάποιο υποσύνολο από search buckets. Όμως πρέπει να διατηρεί μια σωστή εικόνα της λειτουργίας της διαδικασίας. Αν κάποιος κόμβος βγει εκτός συστήματος τότε μπορεί να υποκαταστήσει την λειτουργία του, χρησιμοποιώντας τον εαυτό του και ως ρόλο Search για τα δοχεία που περιείχε ο κόμβος που βγήκε εκτός λειτουργίας.

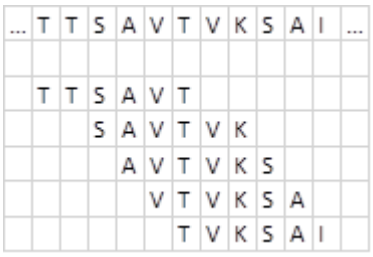
Από τα παραπάνω διαπιστώνεται ότι η επικοινωνία συστημάτων εισάγει μια σημαντική καθυστέρηση στην όλη διαδικασία. Επομένως για να είναι εποικοδομητική αυτή η διαχείριση, θα πρέπει τα υπόλοιπα συστήματα να έχουν αρκετούς πόρους (threads) διαθέσιμους για την λειτουργία. Εάν δεν υπάρχουν διαθέσιμοι αρκετοί πόροι η κατάτμηση της αναζήτησης σε περισσότερα μηχανήματα μπορεί να προκαλέσει ακόμη μεγαλύτερη καθυστέρηση.

3.7 Μετά-ανάλυση

Το παραγόμενο uniquome ενός οργανισμού είναι το πιο βαρύ υπολογιστικά κομμάτι της τρέχουσας διατριβής. Από την στιγμή της δημιουργίας του μπορεί να χρησιμοποιηθεί σε συνδυασμό με το αρχείο πρωτεώματος από το οποίο παρήχθη ή με αρχεία που προέρχονται από άλλες βιολογικές βάσεις δεδομένων για να μας δώσει αρκετές χρήσιμες πληροφορίες. Όλες αυτές τις θεωρούμε μέρος της μετά-ανάλυσης του uniquome και ακολουθεί η μεθοδολογία για τις μετά-αναλύσεις που εκτελέσθηκαν για τους σκοπούς της παρούσας διατριβής.

Σύνθετα Μοναδικά Πεπτίδια (CmUP)

Για την κατασκευή των CmUP η εφαρμογή ταξινομεί το uniquome κάθε πρωτεΐνης και στη συνέχεια ενώνει τα υπερ-καλυπτόμενα CrUP ώστε να συνθέσει το αντίστοιχο Σύνθετο Μοναδικό Πεπτίδιο (Composite unique peptide, CmUP). Ο αλγόριθμος παραγωγής μπορεί να δεχτεί ως είσοδο τη μέγιστη επιθυμητή απόσταση που μπορούν να έχουν δυο CrUP ώστε να θεωρηθούν μέρος του ίδιου CmUP. Στην παρούσα διατριβή η προεπιλογή της μέγιστης επιθυμητής απόστασης είναι το 0 άρα θα πρέπει τα CrUPs να αλληλεπικαλύπτονται κατά την σύνθεση του CmUP. Στην συνέχεια, μετά την κατασκευή του CmUP, ο αλγόριθμος αποθηκεύει την θέση του εντός της πρωτεΐνης καθώς και τα CrUP από τα οποία αποτελείται και τις θέσεις αυτών.



Εικόνα 35 Κατασκευή Composite Unique Peptide από τα αντίστοιχα CrUP

Στατιστικά μοναδικών πεπτιδίων

Πέρα από την καταγραφή/καταλογογράφηση των μοναδικών πεπτιδίων (τόσο των CrUPs όσο και των CmUPs) ενός οργανισμού, μέσω της εφαρμογής που αναπτύχθηκε παραπάνω, έχουν διερευνηθεί εκτεταμένα τα χαρακτηριστικών τους σε ένα υψηλά συστημικό και συνθετικό επίπεδο. Η διερεύνηση περιλαμβάνει:

- Την κατανομή του μήκους των Core Unique και Composite Unique Peptides,
- Την στατιστική κατανομή της θέσης έναρξης του κάθε πεπτιδίου μέσα στην ομοειδή πρωτεΐνη στο σύνολο των πρωτεϊνών του ανθρώπινου πρωτεώματος,

- Την κατανομή της πυκνότητας από μοναδικά πεπτίδια στο σύνολο των πρωτεϊνών του ανθρώπινου πρωτεώματος, και
- Την κατανομή της κάλυψης από μοναδικά πεπτίδια στο σύνολο των πρωτεϊνών του ανθρώπινου πρωτεώματος.

Αναζήτηση όμοιων μοναδικών πεπτιδίων μεταξύ Uniquome διαφορετικών οργανισμών

Μία από τις δυνατότητες της εφαρμογής για την μετά-ανάλυση των μοναδικών πεπτιδίων είναι η αναζήτηση όμοιων πεπτιδίων μεταξύ uniquome διαφορετικών οργανισμών. Για την εύρεση των όμοιων μοναδικών πεπτιδίων στα επιθυμητά πρωτεώματα οργανισμών, η τεχνική που ακολουθείται δέχεται ως είσοδο τα uniquomes των οργανισμών. Ο αλγόριθμος εντοπίζει το μικρότερο uniquome και αναζητά κάθε μοναδικό πεπτίδιο (CrUP ή CmUP ανάλογα την επιλογή) αυτού μέσα στα uniquomes των υπόλοιπων οργανισμών. Αν βρεθεί σε όλα τα επιθυμητά προς μελέτη uniquomes θεωρείται κοινό.

Μοναδικά πεπτίδια σε συνδυασμό με άλλες βιολογικές βάσεις δεδομένων.

Πληροφορίες από την βάση δεδομένων των Uniquomes αντλήθηκαν με σκοπό την συνδυαστική ανάλυσή τους με άλλες βιολογικές βάσεις δεδομένων για την εξαγωγή περαιτέρω σύνθετων πληροφοριών για τα μοναδικά πεπτίδια. Στις συγκεκριμένες αναλύσεις χρησιμοποιήθηκε η εφαρμογή που αναπτύχθηκε σε συνδυασμό με δεδομένα που προέρχονται από άλλες βάσεις δημιουργώντας κάθε φορά νέους αλγορίθμους ανάλογα με τα δεδομένα που αναζητήθηκαν. Με αυτή την διαδικασία έγινε:

- Αναζήτηση μοναδικών πεπτιδίων στα χρωμοσώματα του ανθρώπου και των λοιπών οργανισμών που μελετήθηκαν
- Αναζήτηση μοναδικών πεπτιδίων σε ομάδες πρωτεϊνών του ανθρώπου και των λοιπών οργανισμών που μελετήθηκαν
- Αναζήτηση μοναδικών πεπτιδίων σε άλλες βάσεις δεδομένων με ανθρώπινα πεπτίδια όπως:
    - Ανοσοπεπτίδια – immune epitopes peptides
    - Αντιγονικά καρκινικά πεπτίδια – Cancer Antigenic peptides

4. Αποτελέσματα και Εφαρμογές

4.1 Αποτελέσματα για το Uniquome του ανθρώπου – *Homo sapiens*

Πέρα από τα αποτελέσματα για το Uniquome του ανθρώπινου πρωτεώματος που περιλαμβάνουν την απλή καταγραφή των Core Unique και Composite Unique Peptides στην παρούσα διατριβή έχει επιτευχθεί και η διερεύνηση των χαρακτηριστικών τούς σε ένα υψηλά συστημικό και συνθετικό επίπεδο. Κάποια από τα σημαντικότερα χαρακτηριστικά τα οποία αναλύθηκαν είναι:

Μήκος Μοναδικών Πεπτιδίων

Μία από τις βασικότερες αναλύσεις των χαρακτηριστικών του Uniquome είναι η ομαδοποίηση των μοναδικών πεπτιδίων του, ανάλογα με τον αριθμό από αμινοξέα που αποτελούνται (μήκος πεπτιδίου). Το μήκος των μοναδικών πεπτιδίων ελαχίστου μήκους για τα Uniquome των οργανισμών κυμαίνεται από 4 αμινοξέα (είναι ο μικρότερος αριθμός αμινοξέων που δημιουργεί μοναδικό πεπτίδιο) έως 100 αμινοξέα (είναι ο μέγιστος αριθμός που αποτελείται ένα πεπτίδιο εξ ορισμού).

Σχετική θέση εμφάνισης Μοναδικών Πεπτιδίων

Ένα από τα σημαντικότερα χαρακτηριστικά του Uniquome είναι η θέση εμφάνισης των μοναδικών πεπτιδίων μέσα στην πρωτεΐνη. Για την ανάλυση του συγκεκριμένου χαρακτηριστικού ορίσθηκε η σχετική θέση εμφάνισης ενός μοναδικού πεπτιδίου η οποία αντιστοιχεί στην θέση του πρώτου αμινοξέος του μοναδικού πεπτιδίου ως προς το σύνολο των αμινοξέων της πρωτεΐνης (%).

Δημιουργία σύνθετων μοναδικών πεπτιδίων από μοναδικά πεπτίδια ελαχίστου μήκους

Ένα άλλο σημαντικό χαρακτηριστικό του Uniquome είναι το πλήθος των μοναδικών πεπτιδίων ελαχίστου μήκους που συμμετέχουν για την δημιουργία ενός σύνθετου μοναδικού πεπτιδίου. Βάση αυτής της πληροφορίας περιγράφεται η διατήρηση της μοναδικότητας των σύνθετων μοναδικών πεπτιδίων καθώς από όσα πιο πολλά μοναδικά πεπτίδια ελαχίστου μήκους αποτελείται ένα σύνθετο μοναδικό πεπτίδιο τόσο μεγαλύτερη μοναδικότητα θα έχει και συνεπώς θα είναι πιο ανθεκτικό στις μεταλλάξεις.

Πυκνότητα Μοναδικών Πεπτιδίων

Ένα χαρακτηριστικό του Uniquome που χρησιμοποιείται για τον προσδιορισμό της ποσότητας από μοναδικά πεπτίδια των πρωτεϊνών είναι η πυκνότητα. Ο ορισμός της πυκνότητας μιας πρωτεΐνης είναι ο λόγος του συνολικού αριθμού Μοναδικών Πεπτιδίων (CrUP ή CmUP) μιας πρωτεΐνης προς το σύνολο των αμινοξέων της πρωτεΐνης.

Μοναδική Κάλυψη

Τέλος, ένα άλλο χαρακτηριστικό με το οποίο προσδιορίζεται το ποσοστό των αμινοξέων που συμβάλουν στον σχηματισμό μοναδικών πεπτιδίων σε επίπεδο πρωτεϊνών είναι η κάλυψη. Ο ορισμός της κάλυψης μιας πρωτεΐνης είναι ο λόγος του συνολικού αριθμού των αμινοξέων της πρωτεΐνης που εμπεριέχονται έστω μία φορά στον σχηματισμό μοναδικών πεπτιδίων προς το σύνολο των αμινοξέων της πρωτεΐνης.

Τόσο ο όρος πυκνότητα όσο και ο όρος κάλυψη μπορούν να χρησιμοποιηθούν όχι μόνο για την περιγραφή των πρωτεϊνών αλλά και για την περιγραφή ολόκληρου του πρωτεώματος του ανθρώπου ως συνολική πυκνότητα του οργανισμού και μοναδική κάλυψη του οργανισμού αντίστοιχα.

Στον οργανισμό του ανθρώπου αναλύθηκαν 20.430 θεωρημένες/επιβεβαιωμένες ανθρώπινες πρωτεΐνες όπως έχουν καταγραφεί στην βάση δεδομένων Uniprot (version 10/2019). Σε 20.282 πρωτεΐνες καταγράφηκαν 7.263.888 μοναδικά πεπτίδια ελαχίστου μήκους (CrUP). Η συνολική πυκνότητα μοναδικών πεπτιδίων ελαχίστου μήκους για τον οργανισμού του ανθρώπου είναι 64% (δηλαδή το συνολικό πρωτέωμα του ανθρώπου ανά 100 αμινοξέα έχει 64 μοναδικά πεπτίδια ελαχίστου μήκους). Τα 52.765 από το σύνολο των μοναδικών πεπτιδίων ελαχίστου μήκους εμφανίζονται παραπάνω από μία φορά στην ίδια πρωτεΐνη (τα πεπτίδια αυτά εξακολουθούν να ορίζονται ως μοναδικά πεπτίδια καθώς ναι μεν εμφανίζονται παραπάνω από μία φορά αλλά εμφανίζονται μόνο σε μία πρωτεΐνη). Τα 7.263.888 μοναδικά πεπτίδια ελαχίστου μήκους δημιουργούν 77.697 σύνθετα μοναδικά πεπτίδια (CmUP), αριθμός ο οποίος αντιστοιχεί σε συνολική πυκνότητα από σύνθετα μοναδικά πεπτίδια 0,68%. Ο οργανισμός του ανθρώπου έχει 93% συνολική κάλυψη από μοναδικά πεπτίδια (δηλαδή στο συνολικό πρωτέωμα του ανθρώπου ανά 100 αμινοξέα τα 93 συμμετέχουν στο σχηματισμό μοναδικών πεπτιδίων). Τέλος, από τις 20.430 επιβεβαιωμένες πρωτεΐνες υπάρχουν 148 πρωτεΐνες (0,72%) οι οποίες δεν περιλαμβάνουν κανένα μοναδικό πεπτίδιο (μήκους 4 έως 100 αμινοξέα). Έτσι διαπιστώνεται ότι το ανθρώπινο Uniquome (Human Uniquome) αποτελείται από 7.263.888 CrUPs kai 77.697 CmUPs (Πίνακας 4).

.

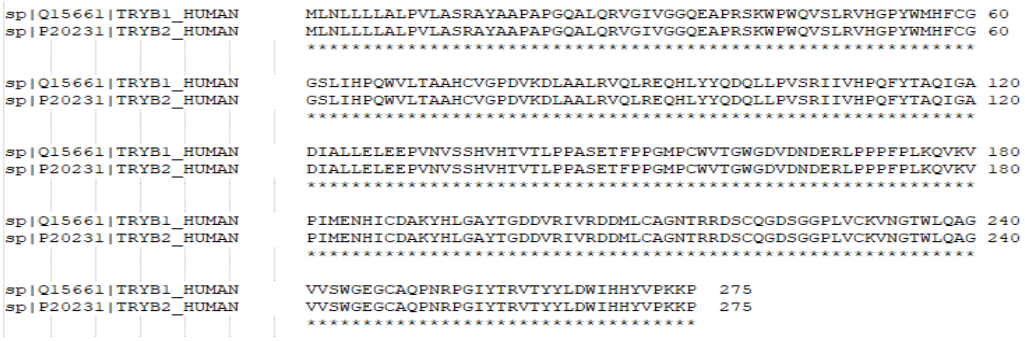
Human Uniquome	
Πρωτεΐνες (reviewed)	20.430
Πρωτεΐνες με μοναδικά πεπτίδια	20.282
Πρωτεΐνες χωρίς μοναδικά πεπτίδια	148
Μοναδικά πεπτίδια ελαχίστου μήκους	7.263.888
Μοναδικά πεπτίδια ελαχίστου μήκους >1 φορά	52.765
Σύνθετα μοναδικά πεπτίδια	77.697
Συνολική πυκνότητα μοναδικών πεπτιδίων ελαχίστου μήκους	64%
Συνολική πυκνότητα σύνθετων μοναδικών πεπτιδίων	0,68%
Συνολική Κάλυψη	93%

Πίνακας 4 Συνοπτικός πίνακας των αποτελεσμάτων του Uniquome του οργανισμού του ανθρώπου

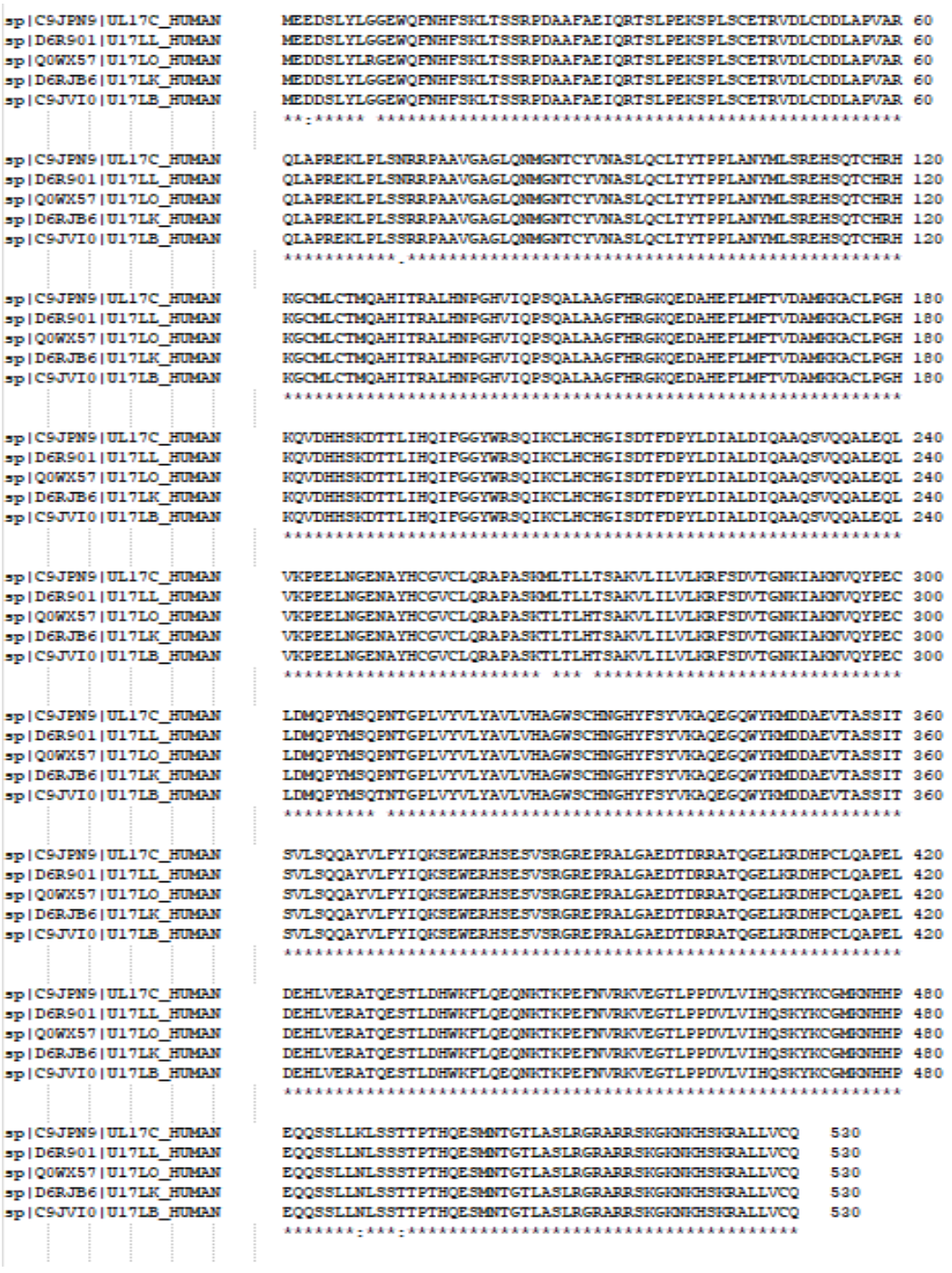
Οι 148 πρωτεΐνες που εντοπίστηκαν χωρίς μοναδικά πεπτίδια αναλύθηκαν περαιτέρω με σκοπό την αναζήτηση του λόγου της μη εμφάνισης μοναδικότητας στην αλληλουχία τους. Οι πρωτεΐνες αυτές όπως παρατηρήθηκε, ανήκουν ανά ζεύγη, ή και ανά ομάδες με περισσότερα από 2 μέλη, σε 51 οικογένειες πρωτεϊνών. Για τους σκοπούς της ανάλυσης, μελετήθηκαν οι οικογένειες G-protein coupled receptor 1 (GPCR), Peptidase C19 και Peptidase S1 οι οποίες περιλαμβάνουν 5 , 5 και 2 πρωτεΐνες αντίστοιχα χωρίς μοναδικά πεπτίδια (12% του συνόλου των 148 πρωτεϊνών χωρίς μοναδικά πεπτίδια). Λόγω του ότι η οικογένεια πρωτεϊνών G-protein coupled receptor 1 (GPCR) είναι η οικογένεια που αποτελείται από τον μεγαλύτερο αριθμό πρωτεϊνών (724) οι πρωτεΐνες της που δεν περιλαμβάνουν μοναδικά πεπτίδια χωρίστηκαν σε δύο μικρότερες ομάδες. Στις συγκεκριμένες πρωτεΐνες που δεν περιλαμβάνουν μοναδικά πεπτίδια πραγματοποιήθηκε αμινοξική ευθυγράμμιση και υπολογίστηκαν τα ποσοστά ομοιότητας τους. Τα αποτελέσματα έδειξαν πώς οι ομάδες των πρωτεϊνών της οικογένειας GPCR καθώς και οι δύο πρωτεΐνες της οικογένειας Peptidase S1 είχαν ποσοστό ομοιότητας της αμινοξικής τους αλληλουχίας 100% (ταυτόσημη αλληλουχία) ενώ οι πέντε πρωτεΐνες της οικογένειας Peptidase C19 είχαν ποσοστό ομοιότητας 99,4% (Εικόνα 36,37 και 38).

```
sp|P0DN77|OPSG2_HUMAN      MAQQWSLQRLAGRHPQDSYEDSTQSSIFTYTNSNSTRGPFEGPNYHIAPRWVYHLTSVWM 60
sp|P0DN78|OPSG3_HUMAN      MAQQWSLQRLAGRHPQDSYEDSTQSSIFTYTNSNSTRGPFEGPNYHIAPRWVYHLTSVWM 60
sp|P04001|OPSG_HUMAN       MAQQWSLQRLAGRHPQDSYEDSTQSSIFTYTNSNSTRGPFEGPNYHIAPRWVYHLTSVWM 60
                           ************************************************************

sp|P0DN77|OPSG2_HUMAN      IFVVIASVFTNGLVLAATMKFKKLRHPLNWILVNLAVADLAETVIASTISVVNQVYGYFV 120
sp|P0DN78|OPSG3_HUMAN      IFVVIASVFTNGLVLAATMKFKKLRHPLNWILVNLAVADLAETVIASTISVVNQVYGYFV 120
sp|P04001|OPSG_HUMAN       IFVVIASVFTNGLVLAATMKFKKLRHPLNWILVNLAVADLAETVIASTISVVNQVYGYFV 120
                           ************************************************************

sp|P0DN77|OPSG2_HUMAN      LGHPMCVLEGYTVSLCGITGLWSLAIISWERWMVVCKPFGNVRFDAKLAIVGIAFSWIWA 180
sp|P0DN78|OPSG3_HUMAN      LGHPMCVLEGYTVSLCGITGLWSLAIISWERWMVVCKPFGNVRFDAKLAIVGIAFSWIWA 180
sp|P04001|OPSG_HUMAN       LGHPMCVLEGYTVSLCGITGLWSLAIISWERWMVVCKPFGNVRFDAKLAIVGIAFSWIWA 180
                           ************************************************************

sp|P0DN77|OPSG2_HUMAN      AVWTAPPIFGWSRYWPHGLKTSCGPDVFSGSSYPGVQSYMIVLMVTCCITPLSIIVLCYL 240
sp|P0DN78|OPSG3_HUMAN      AVWTAPPIFGWSRYWPHGLKTSCGPDVFSGSSYPGVQSYMIVLMVTCCITPLSIIVLCYL 240
sp|P04001|OPSG_HUMAN       AVWTAPPIFGWSRYWPHGLKTSCGPDVFSGSSYPGVQSYMIVLMVTCCITPLSIIVLCYL 240
                           ************************************************************

sp|P0DN77|OPSG2_HUMAN      QVWLAIRAVAKQQKESESTQKAEKEVTRMVVVMVLAFCFCWGPYAFFACFAAANPGYPFH 300
sp|P0DN78|OPSG3_HUMAN      QVWLAIRAVAKQQKESESTQKAEKEVTRMVVVMVLAFCFCWGPYAFFACFAAANPGYPFH 300
sp|P04001|OPSG_HUMAN       QVWLAIRAVAKQQKESESTQKAEKEVTRMVVVMVLAFCFCWGPYAFFACFAAANPGYPFH 300
                           ************************************************************

sp|P0DN77|OPSG2_HUMAN      PLMAALPAFFAKSATIYNPVIYVFMNRQFRNCILQLFGKKVDDGSELSSASKTEVSSVSS 360
sp|P0DN78|OPSG3_HUMAN      PLMAALPAFFAKSATIYNPVIYVFMNRQFRNCILQLFGKKVDDGSELSSASKTEVSSVSS 360
sp|P04001|OPSG_HUMAN       PLMAALPAFFAKSATIYNPVIYVFMNRQFRNCILQLFGKKVDDGSELSSASKTEVSSVSS 360
                           ************************************************************

sp|P0DN77|OPSG2_HUMAN      VSPA 364
sp|P0DN78|OPSG3_HUMAN      VSPA 364
sp|P04001|OPSG_HUMAN       VSPA 364
                           ****


sp|P0DQD5|NPY42_HUMAN      MNTSHLLALLLPKSPQGENRSKPLGTPYNFSEHCQDSVDVMVFIVTSYSIETVVGVLGNL 60
sp|P50391|NPY4R_HUMAN      MNTSHLLALLLPKSPQGENRSKPLGTPYNFSEHCQDSVDVMVFIVTSYSIETVVGVLGNL 60
                           ************************************************************

sp|P0DQD5|NPY42_HUMAN      CLMCVTVRQKEKANVTNLLIANLAFSDFLMCLLCQPLTAVYTIMDYWIFGETLCKMSAFI 120
sp|P50391|NPY4R_HUMAN      CLMCVTVRQKEKANVTNLLIANLAFSDFLMCLLCQPLTAVYTIMDYWIFGETLCKMSAFI 120
                           ************************************************************

sp|P0DQD5|NPY42_HUMAN      QCMSVTVSILSLVLVALERHQLIINPTGWKPSISQAYLGIVLIWVIACVLSLPFLANSIL 180
sp|P50391|NPY4R_HUMAN      QCMSVTVSILSLVLVALERHQLIINPTGWKPSISQAYLGIVLIWVIACVLSLPFLANSIL 180
                           ************************************************************

sp|P0DQD5|NPY42_HUMAN      ENVFHKNHSKALEFLADKVVCTESWPLAHHRTIYTTFLLLFQYCLPLGFILVCYARIYRR 240
sp|P50391|NPY4R_HUMAN      ENVFHKNHSKALEFLADKVVCTESWPLAHHRTIYTTFLLLFQYCLPLGFILVCYARIYRR 240
                           ************************************************************

sp|P0DQD5|NPY42_HUMAN      LQRQGRVFHKGTYSLRAGHMKQVNVVLVVMVVAFAVLWLPLHVFNSLEDWHHEAIPICHG 300
sp|P50391|NPY4R_HUMAN      LQRQGRVFHKGTYSLRAGHMKQVNVVLVVMVVAFAVLWLPLHVFNSLEDWHHEAIPICHG 300
                           ************************************************************

sp|P0DQD5|NPY42_HUMAN      NLIFLVCHLLAMASTCVNPFIYGFLNTNFKKEIKALVLTCQQSAPLEESEHLPLSTVHTE 360
sp|P50391|NPY4R_HUMAN      NLIFLVCHLLAMASTCVNPFIYGFLNTNFKKEIKALVLTCQQSAPLEESEHLPLSTVHTE 360
                           ************************************************************

sp|P0DQD5|NPY42_HUMAN      VSKGSLRLSGRSNPI  375
sp|P50391|NPY4R_HUMAN      VSKGSLRLSGRSNPI  375
                           ***************
```

Εικόνα 36 Αμινοξική ευθυγράμμιση των υποομάδων πρωτεϊνών NPY και OPSG της οικογενείας G-protein coupled receptor 1.



Εικόνα 37 Αμινοξική ευθυγράμμιση των πρωτεϊνών της οικογενείας Peptidase S1.

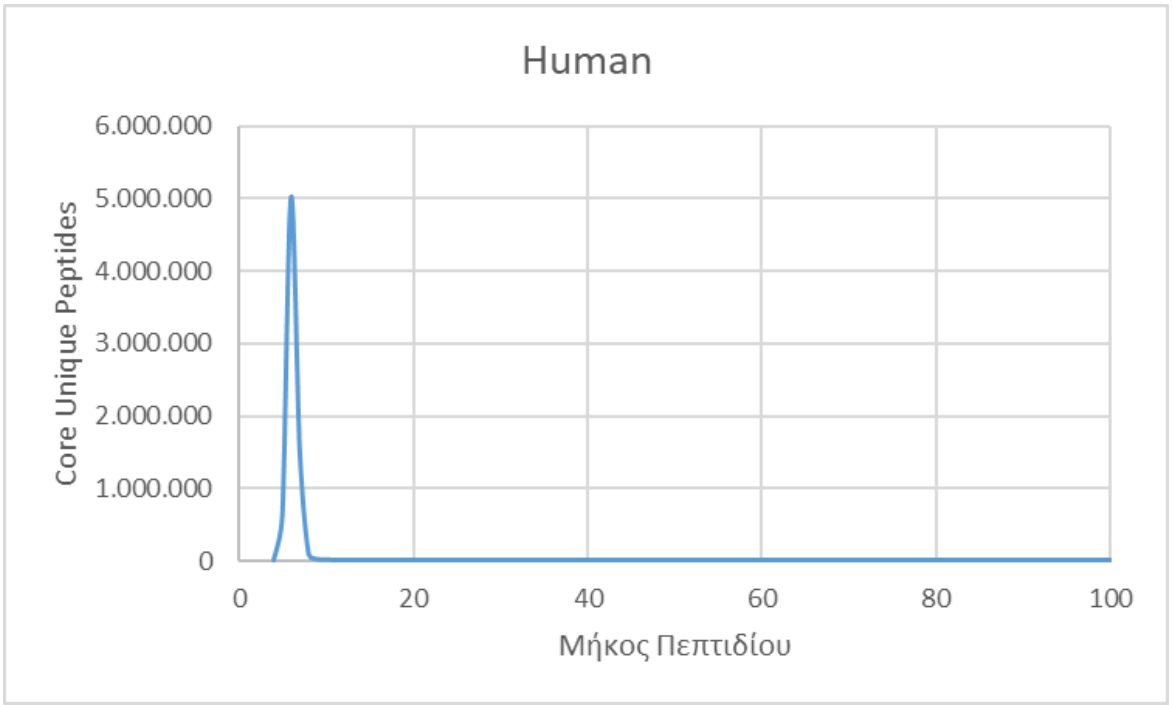


Εικόνα 38 Αμινοξική ευθυγράμμιση των πρωτεϊνών της οικογενείας Peptidase C19.

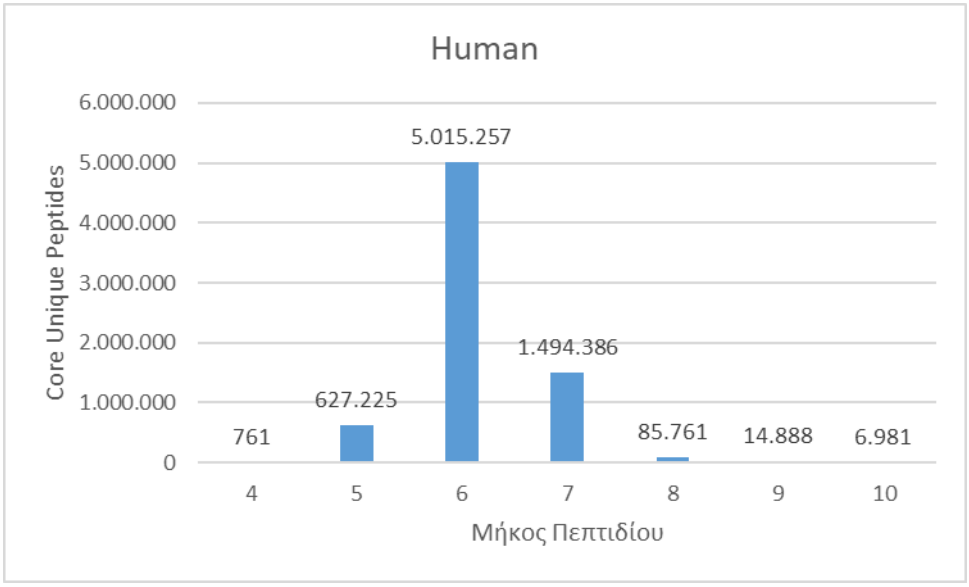
4.1.1 Αποτελέσματα μοναδικών πεπτιδίων ελαχίστου μήκους

Στη συνέχεια της ανάλυσης του Uniquome του ανθρώπου, διερευνήθηκαν τα χαρακτηριστικά των μοναδικών πεπτιδίων ανάλογα με την κατηγορία μοναδικότητας που ανήκουν (CrUP και CmUP).

Τα μοναδικά πεπτίδια ελαχίστου μήκους αποτελούνται από 4 -100 αμινοξέα. Η συντριπτική πλειοψηφία των μοναδικών πεπτιδίων ελαχίστου μήκους είναι τα πεπτίδια με μήκος από 5 έως 7 αμινοξέα (7.136.868 τα οποία αντιστοιχούν στο 97% του συνόλου των CrUP στον άνθρωπο). Ειδικότερα, τα μοναδικά πεπτίδια ελαχίστου μήκους με μέγεθος 6 αμινοξέων είναι η ομάδα με το μεγαλύτερο αριθμό μοναδικών πεπτιδίων (5.015.527) ακολουθούμενη από τα 7-πεπτίδια (1.494.386) και τα 5-πεπτίδια (627.225). Η ομάδα των μοναδικών πεπτιδίων ελαχίστου μήκους με το μικρότερο μήκος περιλαμβάνει 761 μοναδικά πεπτίδια τα οποία είναι και τα πεπτίδια με το μικρότερο δυνατό μήκος που είναι μοναδικά. Τέλος, παρατηρείται ότι τα μοναδικά πεπτίδια ελαχίστου μήκους τα οποία ανήκουν στις ομάδες με μήκους μεγαλύτερο από 8 αμινοξέα εμφανίζονται με πολύ μικρό αριθμό από πεπτίδια (Εικόνα 39 και Εικόνα 40).



Εικόνα 39 Πλήθος CrUP ανάλογα το μήκος του πεπτιδίου (4-100) στον άνθρωπο

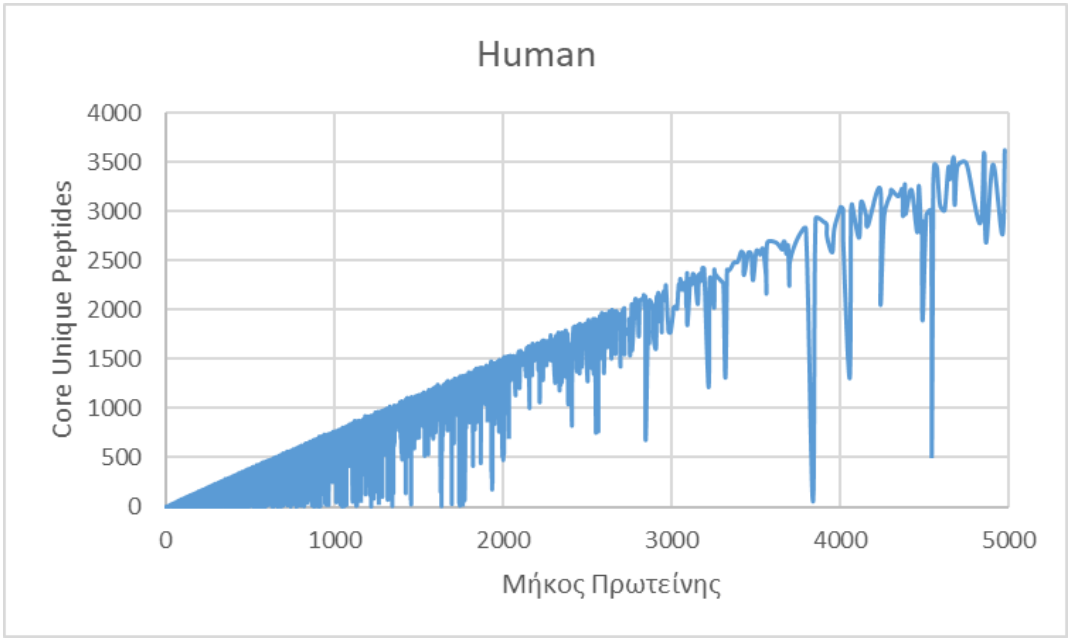


Εικόνα 40 Πλήθος CrUP ανάλογα το μήκος του πεπτιδίου (4-10) στον άνθρωπο

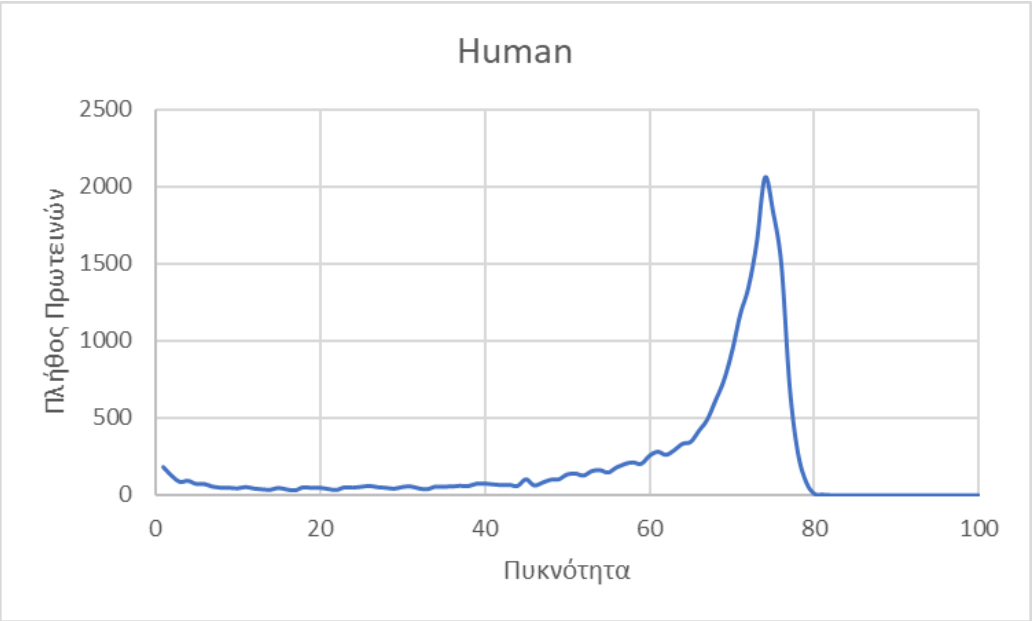
Στη συνέχεια της καταγραφής των χαρακτηριστικών των αποτελεσμάτων για τα μοναδικά πεπτίδια ελαχίστου μήκους στον οργανισμό του ανθρώπου, αναλύθηκε η σχετική θέση στην οποία τα CrUP εμφανίζονται μέσα στις πρωτεΐνες. Η συγκεκριμένη ανάλυση έδειξε πως τα μοναδικά πεπτίδια ελαχίστου μήκους εμφανίζονται με την ίδια συχνότητα στις διάφορες θέσεις των πρωτεϊνών του ανθρώπου καθώς όλες οι σχετικές θέσεις στις οποίες εμφανίζονται τα μοναδικά πεπτίδια ελαχίστου μήκους συγκεντρώνονται με περίπου τον ίδιο αριθμό. Μικρή εξαίρεση αποτελούν οι πολύ τελευταίες σχετικές θέσεις των πρωτεϊνών καθώς όπως είναι λογικό δεν μπορούν να σχηματιστούν πεπτίδια με μήκους μεγαλύτερο από 4 αμινοξέα (Εικόνα 41) .

Εικόνα 41 Πλήθος CrUP ανά σχετική θέση εμφάνισης στον άνθρωπο

Ένα από τα σημαντικότερα χαρακτηριστικά που αναλύθηκε για τα μοναδικά πεπτίδια ελαχίστου μήκους στον οργανισμό του ανθρώπου είναι η πυκνότητα των πρωτεϊνών από CrUPs. Στο πρώτο επίπεδο τής ανάλυσης οι πρωτεΐνες ταξινομήθηκαν με βάση το μέγεθός τους. Η ανάλυση αυτή έδειξε ότι ο αριθμός των μοναδικών πεπτιδίων ελαχίστου μήκους έχει άμεση σχέση με το μέγεθος της πρωτεΐνης καθώς όσο μεγαλύτερες είναι οι πρωτεΐνες (σε αριθμό αμινοξέων) τόσο περισσότερα μοναδικά πεπτίδια εμπεριέχονται σε αυτές (Εικόνα 42). Στο δεύτερο επίπεδο της ανάλυσης αυτής, υπολογίστηκε η πυκνότητα τους με βάση τον αριθμό από μοναδικά πεπτίδια ελαχίστου μήκους που εμπεριέχονται σε αυτές ως προς τον αριθμό των αμινοξέων τους. Η ανάλυση αυτή έδειξε ότι στο σύνολο τους οι πρωτεΐνες του ανθρώπου έχουν πυκνότητα 64% με το μεγαλύτερο πλήθος των πρωτεϊνών να έχει πυκνότητα από 60% έως 75%, με αρκετές όμως εξαιρέσεις καθώς υπάρχουν πρωτεΐνες με πυκνότητα ακόμα και μικρότερη από 5% αλλά και πρωτεΐνες με πυκνότητα μεγαλύτερη του 80%. Περεταίρω ανάλυση σε αυτές τις ομάδες πρωτεϊνών ανέδειξε πως οι πρωτεΐνες με μικρή πυκνότητα ανήκουν σε οικογένειες πρωτεϊνών με πανομοιότυπη αμινοξική αλληλουχία συνεπώς και δυσκολία στον σχηματισμό μοναδικών πεπτιδίων. Αντίθετα, οι πρωτεΐνες που εμφανίζουν μεγαλύτερη πυκνότητα από μοναδικά πεπτίδια ελαχίστου μήκους ανήκουν σε ομάδες στις οποίες η αμινοξική αλληλουχία τους ήταν μοναδική ως προς το συνολικό πρωτέωμα του ανθρώπου (Εικόνα 43).



Εικόνα 42 Αριθμός CrUP στις πρωτεΐνες, ταξινομημένες ως προς το μέγεθος τους από αμινοξέα (για πρωτεΐνες με μέγεθος μέχρι 5.000 αμινοξέα) στον άνθρωπο



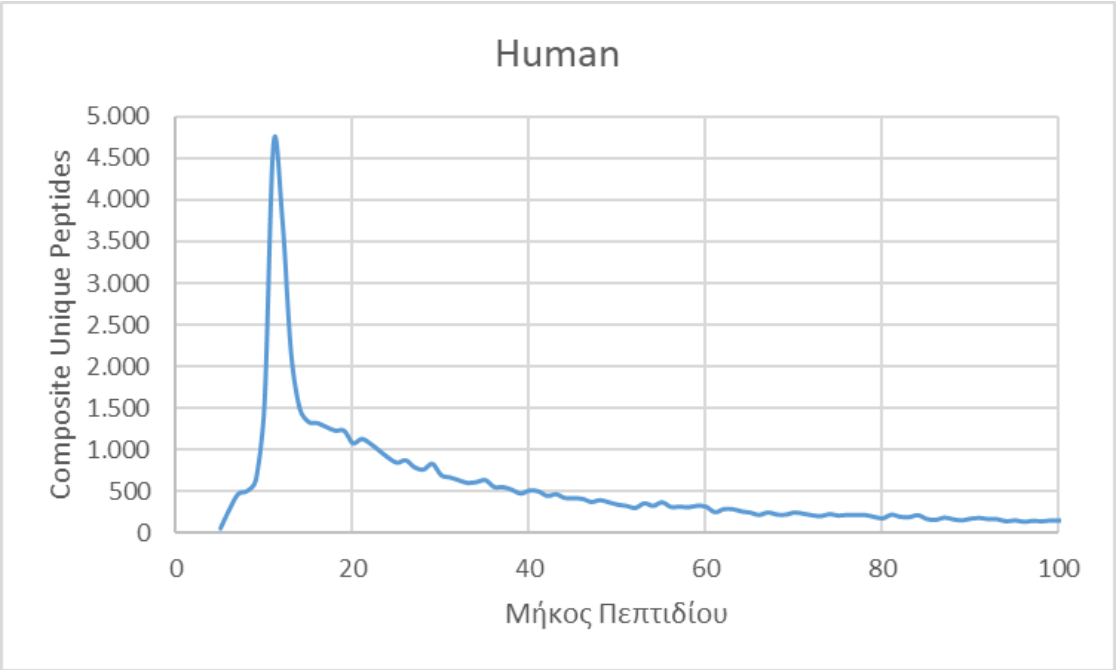
Εικόνα 43 Πλήθος πρωτεϊνών ως προς την πυκνότητά τους απο CrUP στον άνθρωπο

4.1.2 Αποτελέσματα σύνθετων μοναδικών πεπτιδίων

Σε συνέχεια της διερεύνησης των χαρακτηριστικών των μοναδικών πεπτιδίων αναλύθηκαν τα αποτελέσματα, για την δεύτερη κατηγορία των μοναδικών πεπτιδίων, τα σύνθετα μοναδικά πεπτίδια.

Ως προς το μήκος των σύνθετων μοναδικών πεπτιδίων η ανάλυση έδειξε ότι τα σύνθετα μοναδικά πεπτίδια μεγέθους 11 αμινοξέων αποτελούν την πολυπληθέστερη ομάδα (4.676 πεπτίδια) ακολουθούμενη από τα 12- και 13-πεπτίδια (3.769 και 2.155 πεπτίδια αντίστοιχα), αντίθετα τα σύνθετα μοναδικά πεπτίδια τα οποία αποτελούνται από 5 έως 9 αμινοξέα σχηματίζουν ομάδες με μικρό αριθμό πεπτιδίων. Παρατηρείται επίσης ότι ο αριθμός των σύνθετων μοναδικών πεπτιδίων αρχίζει να μειώνεται αισθητά στις ομάδες με μήκος μεγαλύτερο των 40 αμινοξέων. Τα μικρότερου μεγέθους σύνθετα μοναδικά πεπτίδια που καταγράφηκαν έχουν μήκος 5 αμινοξέων (61 πεπτίδια) ενώ το μεγαλύτερο CmUP που καταγράφηκε αποτελείται είχε μήκος 33.679 αμινοξέων (Εικόνα 44, Πίνακας 5).

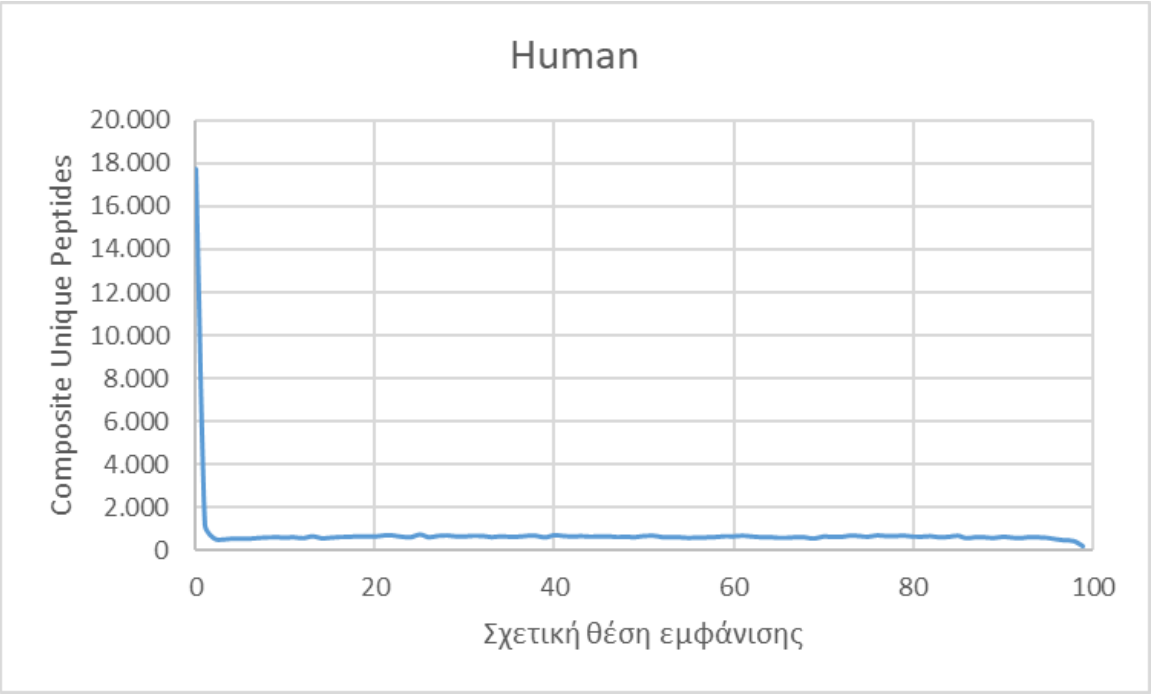
Στην συνέχεια της ανάλυσης των χαρακτηριστικών των σύνθετων μοναδικών πεπτιδίων, καταγράφηκε η σχετική θέση εμφάνισης των σύνθετων μοναδικών πεπτιδίων στις ανθρώπινες πρωτεΐνες. Σε αντίθεση με την ομοιόμορφη κατανομή που καταγράφηκε από την σχετική θέση εμφάνισης των μοναδικών πεπτιδίων ελαχίστου μήκους η πλειοψηφία των σύνθετων μοναδικών πεπτιδίων έχει σαν θέση έναρξης τα αμινοξέα που βρίσκονται στις αρχικές θέσεις των πρωτεϊνών του οργανισμού του ανθρώπου (Εικόνα 45).



Εικόνα 44 Πλήθος CmUP ανάλογα το μήκος του πεπτιδίου (5-100 αμινοξέα) στον άνθρωπο

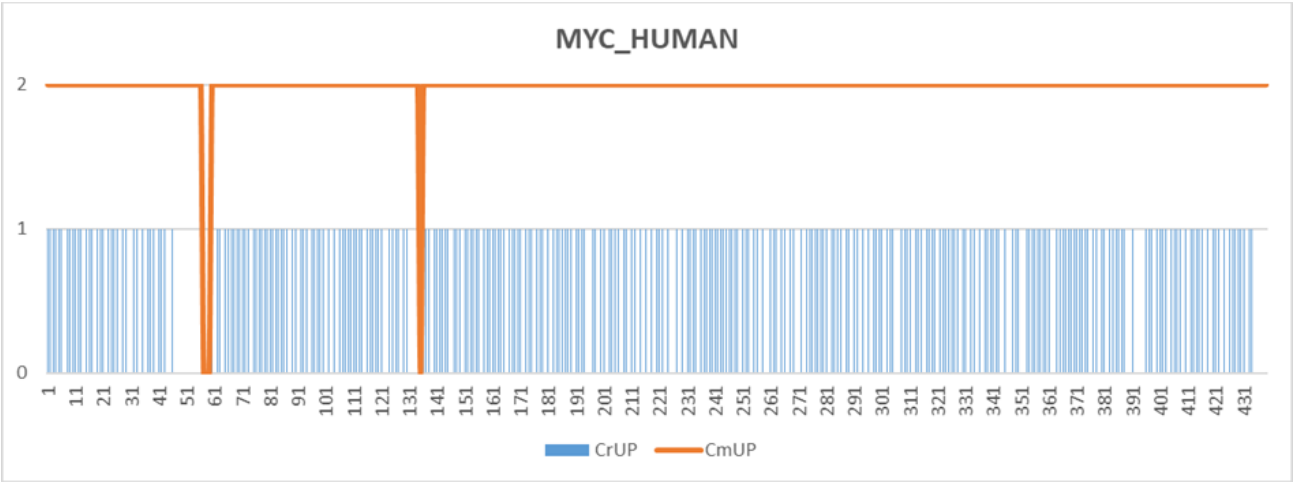
Μήκος Πεπτιδίου	Πλήθος	%
5	61	0,08%
6	288	0,37%
7	473	0,61%
8	508	0,65%
9	638	0,82%
10	1.582	2,04%
11	4.676	6,02%
12	3.769	4,85%
13	2.155	2,77%
14	1.489	1,92%
15	1.336	1,72%
16	1.323	1,70%
17	1.276	1,64%
18	1.233	1,59%
19	1.232	1,59%
20	1.082	1,39%

Πίνακας 5 Πλήθος CmUP ανάλογα το μήκος του πεπτιδίου (5-20 αμινοξέα) στον άνθρωπο



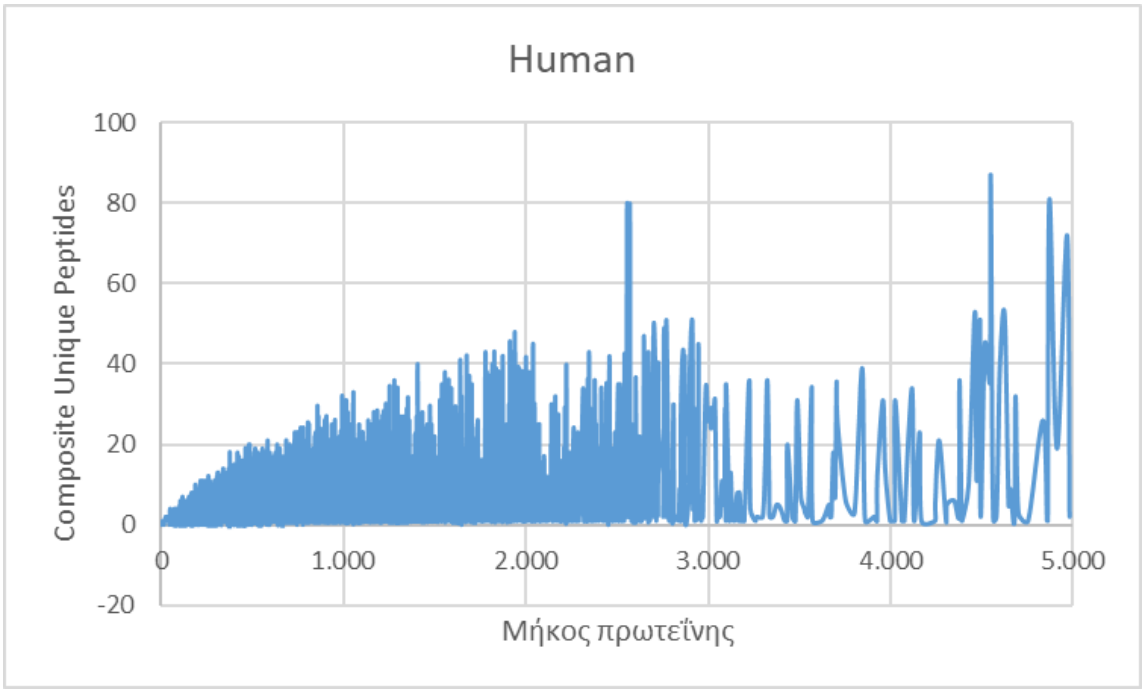
Εικόνα 45 Πλήθος CmUP ανά σχετική θέση εμφάνισης στον άνθρωπο

Για την καλύτερη κατανόηση της σχετικής θέσης εμφάνισης των μοναδικών πεπτιδίων μέσα στις πρωτεΐνες του ανθρώπου, και της εξήγησης του λόγου για τον οποίο τα σύνθετα μοναδικά πεπτίδια εντοπίζονται στις αρχικές θέσεις των πρωτεϊνών αναλύθηκε σαν παράδειγμα η πρωτεΐνη MYC. Η πρωτεΐνη MYC αποτελείται από 439 αμινοξέα στα οποία εντοπίζονται 314 μοναδικά πεπτίδια ελαχίστου μήκους τα οποία συνθέτουν 3 σύνθετα μοναδικά πεπτίδια. Στην εικόνα 46 παρουσιάζεται η κατανομή τόσο των μοναδικών πεπτιδίων ελαχίστου μήκους όσο και των σύνθετων μοναδικών πεπτιδίων στην θέση μέσα στην πρωτεΐνη που εμφανίζονται. Η θέση αυτή, ορίζεται σαν το σημείο έναρξης του αντίστοιχου μοναδικού πεπτιδίου δηλαδή στην θέση που εμφανίζεται το πρώτο του αμινοξύ.



Εικόνα 46 Κατανομή των μοναδικών πεπτιδίων της πρωτεΐνης MYC

Όπως και για τα μοναδικά πεπτίδια ελαχίστου μήκους, έτσι και για τα σύνθετα μοναδικά πεπτίδια, αναλύθηκε ο αριθμός των CmUP στις πρωτεΐνες ως προς το μέγεθος της πρωτεΐνης. Η ανάλυση αυτή ανέδειξε πως ο αριθμός των μοναδικών πεπτιδίων δεν εξαρτάται άμεσα από το μέγεθος της πρωτεΐνης από αμινοξέα (Εικόνα 47).

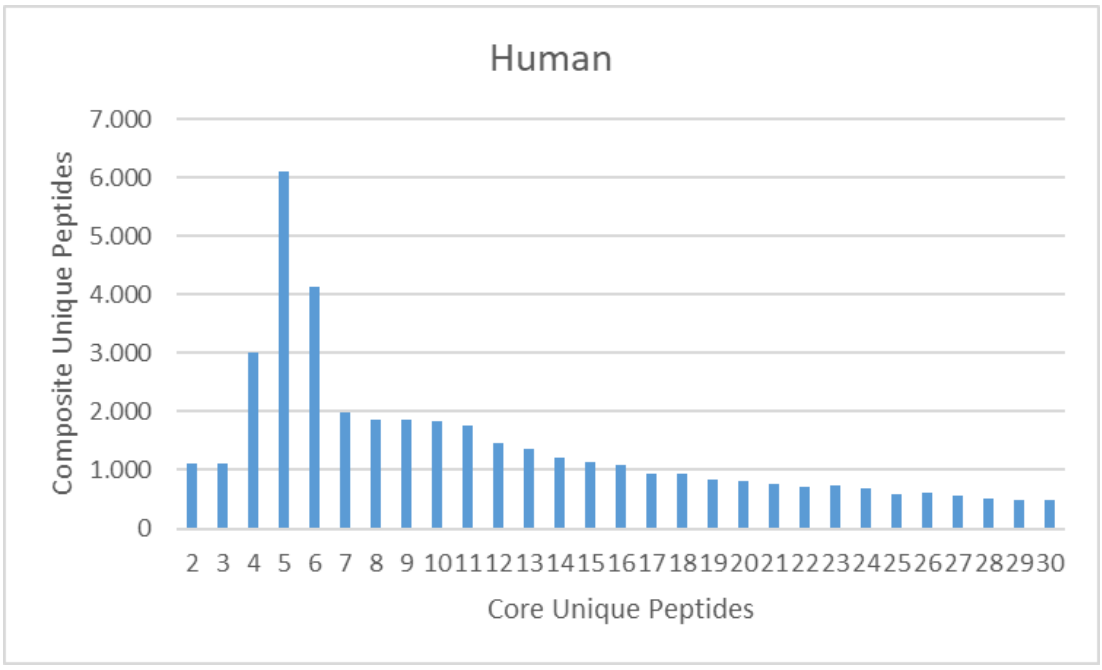


Εικόνα 47 Αριθμός CrUP στις πρωτεΐνες, ταξινομημένες ως προς το μέγεθος τους από αμινοξέα (για πρωτεΐνες με μέγεθος μέχρι 5.000 αμινοξέα) στον άνθρωπο

Τέλος, για την καλύτερη ερμηνεία της σχέσης μεταξύ των δύο κατηγοριών μοναδικών πεπτιδίων για το Uniquome του ανθρώπου αναλύθηκε η πληροφορία με την οποία αποτυπώνεται ο τρόπος με τον οποίο τα CrUPs συνθέτουν τα CmUPs. Τα αποτελέσματα αυτής της ανάλυσης έδειξαν πως για τον οργανισμό του ανθρώπου υπάρχουν σύνθετα μοναδικά πεπτίδια τα οποία συνθέτονται από 2 μέχρι και 25.055 μοναδικά πεπτίδια ελαχίστου μήκους. Τα μεγαλύτερα σε αριθμό σύνθετα μοναδικά πεπτίδια είναι αυτά που αποτελούνται από 5 και 6 μοναδικά πεπτίδια ελαχίστου μήκους (6.103 και 4,123 πεπτίδια αντίστοιχα). Τέλος τα σύνθετα μοναδικά πεπτίδια που δημιουργούνται μέχρι και με 20 μοναδικά πεπτίδια ελαχίστου μήκους εμφανίζονται με υπολογίσιμη σε αριθμό πεπτιδίων ποσότητα (>1% ως προς το πλήθος των σύνθετων μοναδικών πεπτιδίων) ενώ τα σύνθετα μοναδικά πεπτίδια που δημιουργούνται με περισσότερα από 20 μοναδικά πεπτίδια ελαχίστου μήκους παρουσιάζουν μεγάλη μείωση (Εικόνα 48, πίνακας 6).

Αριθμός CrUP που συνθέτουν ένα CmUP	Πλήθος CmUP	%
2	1.123	1,45%
3	1.120	1,44%
4	3.003	3,87%
5	6.103	7,85%
6	4.123	5,31%
7	1.979	2,55%
8	1.873	2,41%
9	1.867	2,40%
10	1.844	2,37%
11	1.751	2,25%
12	1.467	1,89%
13	1.358	1,75%
14	1.217	1,57%
15	1.148	1,48%
16	1.078	1,39%
17	934	1,20%
18	936	1,20%
19	829	1,07%
20	807	1,04%
21	762	0,98%
22	715	0,92%
23	727	0,94%
24	696	0,90%
25	596	0,77%
26	602	0,77%
27	554	0,71%
28	525	0,68%
29	493	0,63%
30	496	0,64%

Πίνακας 6 Πλήθος CmUP ανάλογα τον αριθμό από CrUP που αποτελούνται (2-30 πεπτίδια) στον άνθρωπο



Εικόνα 48 Πλήθος CmUP ανάλογα τον αριθμό από CrUP που αποτελούνται (2-30 πεπτίδια) στον άνθρωπο

4.1.3 Μοναδικά πεπτίδια και Χρωμοσώματα

Για την καλύτερη κατανόηση των μοναδικών πεπτιδίων της βάσης δεδομένων των Uniquomes και τις ιδιότητες τους, κατηγοριοποιήθηκαν τα μοναδικά πεπτίδια βάσει των χρωμοσωμάτων από τα οποία προέρχεται η αμινοξική τους αλληλουχία. Χρησιμοποιώντας τα εργαλεία της Βάσης δεδομένων Uniprot ομαδοποιήθηκαν πρωτεΐνες του ανθρώπου βάση το χρωμόσωμα στο οποίο ανήκουν. Στην συνέχεια χρησιμοποιώντας τη Βάση δεδομένων του ανθρώπινου Uniquome (Human Uniquome) εντοπίστηκαν τα μοναδικά πεπτίδια που αντιστοιχούν σε κάθε χρωμόσωμα. Πέρα από την απλή καταγραφή του αριθμού των μοναδικών πεπτιδίων για κάθε χρωμόσωμα συλλέχθηκαν και τα εξής χαρακτηριστικά :

- Αριθμός πρωτεϊνών ανά χρωμόσωμα χωρίς μοναδικά πεπτίδια
- Πυκνότητα μοναδικών πεπτιδίων ελαχίστου μήκους ανά χρωμόσωμα
- Πυκνότητα σύνθετων μοναδικών πεπτιδίων ανά χρωμόσωμα
- Συνολική κάλυψη μοναδικών πεπτιδίων ανά χρωμόσωμα

Από τα πρώτα αποτελέσματα αυτής της ανάλυσης παρατηρείται πώς τα χαρακτηριστικά των μοναδικών πεπτιδίων, ανάλογα με τα χρωμοσώματα στα ανήκουν οποία οι πρωτεΐνες που τα εμπεριέχουν εμφανίζονται με σημαντικές διαφοροποιήσεις (Πίνακας 12).

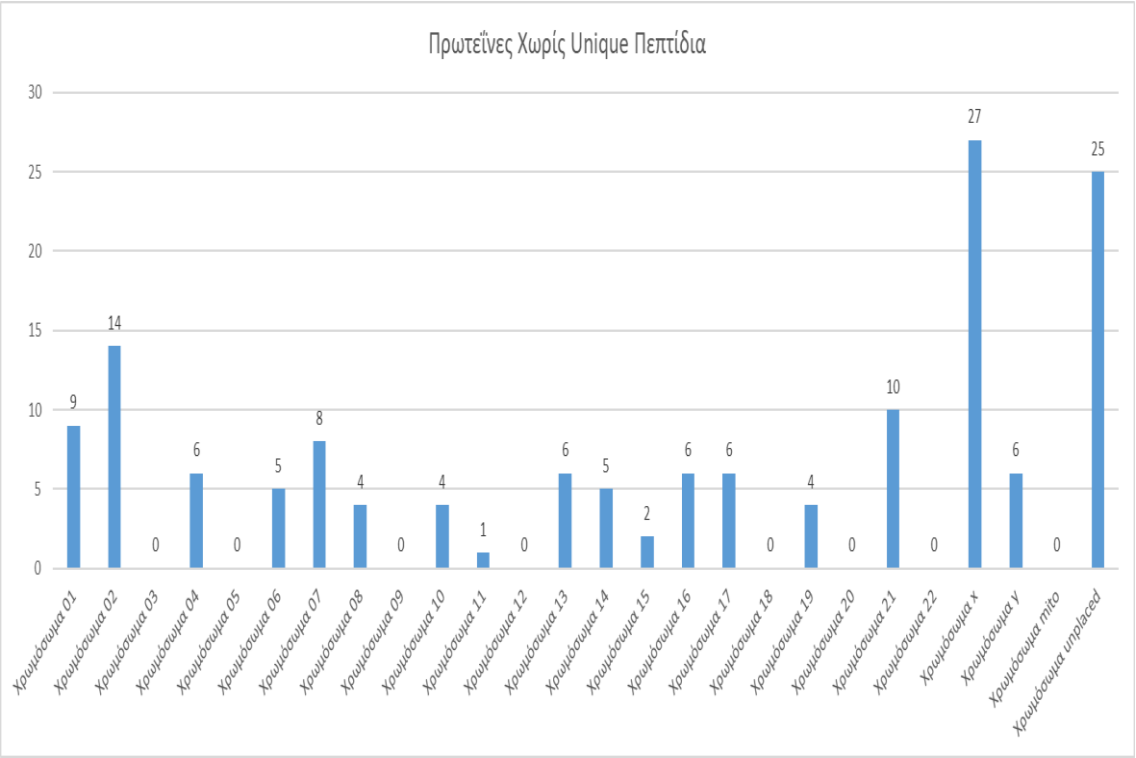
Χρωμόσωμα	Πρωτεΐνες	Πρωτεΐνες Χωρίς Core	CrUP	CmUP	πυκνότητα CrUP	πυκνότητα CmUP	Μοναδική κάλυψη
Χρωμόσωμα 01	1.992	9 (0,45%)	719.132	7.334	64%	0,65%	93%
Χρωμόσωμα 02	1.263	14 (1,11%)	532.141	4.749	65%	0,58%	93%
Χρωμόσωμα 03	1.034	0 (0,00%)	428.272	3.840	68%	0,61%	96%
Χρωμόσωμα 04	731	6 (0,82%)	289.377	2.750	66%	0,63%	95%
Χρωμόσωμα 05	848	0 (0,00%)	329.564	3.906	64%	0,76%	93%
Χρωμόσωμα 06	1.057	5 (0,47%)	370.191	3.619	63%	0,62%	92%
Χρωμόσωμα 07	919	8 (0,87%)	331.159	3.593	63%	0,69%	93%
Χρωμόσωμα 08	646	4 (0,62%)	245.462	2.305	66%	0,62%	95%
Χρωμόσωμα 09	746	0 (0,00%)	286.453	2.773	64%	0,62%	92%
Χρωμόσωμα 10	712	4 (0,56%)	281.380	2.646	66%	0,62%	94%
Χρωμόσωμα 11	1.260	1 (0,08%)	425.749	4.479	65%	0,69%	95%
Χρωμόσωμα 12	992	0 (0,00%)	371.684	4.163	65%	0,73%	95%
Χρωμόσωμα 13	313	6 (1,92%)	132.700	1.303	67%	0,66%	95%
Χρωμόσωμα 14	694	5 (0,72%)	236.113	2.370	65%	0,66%	94%
Χρωμόσωμα 15	567	2 (0,35%)	247.458	2.415	64%	0,63%	92%
Χρωμόσωμα 16	794	6 (0,76%)	299.265	2.453	66%	0,54%	93%
Χρωμόσωμα 17	1.126	6 (0,53%)	401.317	4.533	63%	0,72%	92%
Χρωμόσωμα 18	262	0 (0,00%)	116.714	1.057	67%	0,61%	95%
Χρωμόσωμα 19	1.388	4 (0,29%)	431.491	6.166	59%	0,85%	94%
Χρωμόσωμα 20	524	0 (0,00%)	176.515	1.795	67%	0,68%	96%
Χρωμόσωμα 21	221	10 (4,52%)	69.028	661	63%	0,61%	91%
Χρωμόσωμα 22	461	0 (0,00%	147.738	1.599	64%	0,69%	91%
Χρωμόσωμα x	795	27 (3,40%)	255.779	3.610	61%	0,87%	91%
Χρωμόσωμα y	37	6 (16,22%)	3.677	306	19%	1,56%	35%
Χρωμόσωμα Μιτοχονδριακό	14	0 (0,00%)	2.815	14	74%	0,37%	100%
Χρωμόσωμα αταξινόμητο	1.034	25 (2,42%)	132.714	3.258	43%	1,05%	71%

Πίνακας 7 Κατηγοριοποίηση μοναδικών πεπτιδίων στα χρωμοσώματα του ανθρώπου

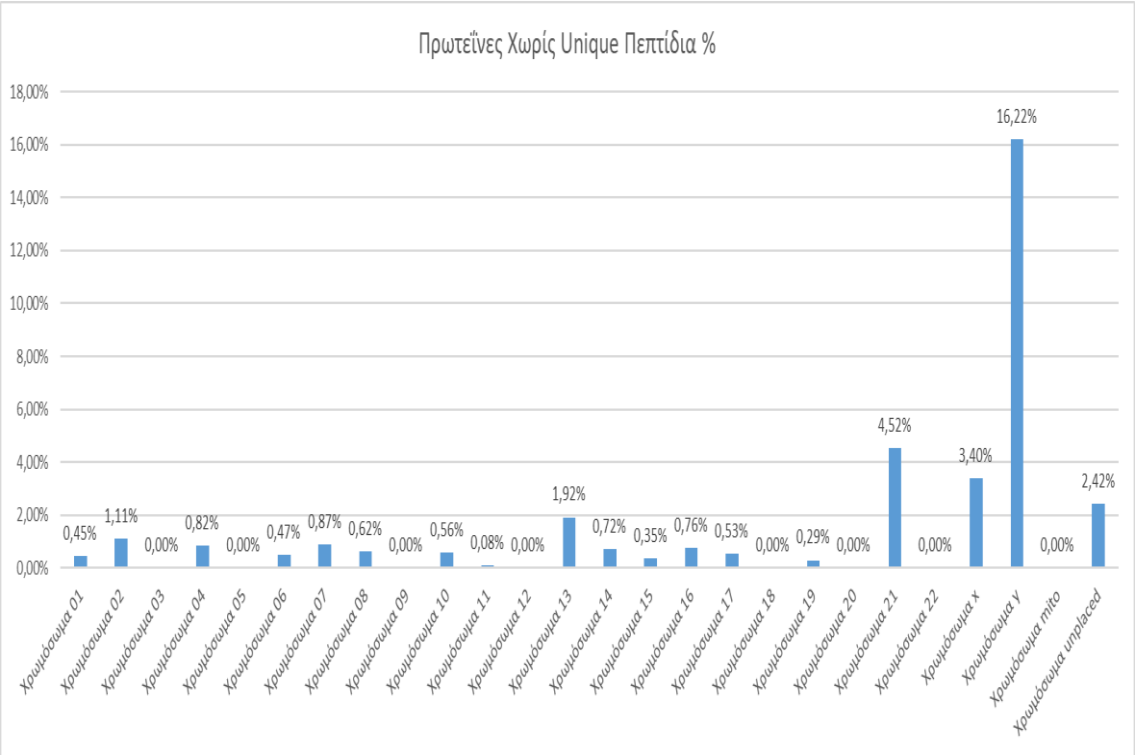
Έντονη διαφοροποίηση παρατηρείται στον αριθμό από πρωτεΐνες οι οποίες δεν εμφανίζουν μοναδικά πεπτίδια ανάλογα στο χρωμόσωμα το οποίο ανήκουν. Ειδικότερα στα χρωμοσώματα 3,5,9,12,18,20,22 καθώς και στο Μιτοχονδριακό χρωμόσωμα όλες οι πρωτεΐνες περιλαμβάνουν μοναδικά πεπτίδια ενώ αντίθετα στα υπόλοιπα χρωμοσώματα υπάρχουν από 1 έως 27 πρωτεΐνες χωρίς μοναδικά πεπτίδια (Εικόνα 49). Για την καλύτερη κατανόηση της διακύμανσης του αριθμού των πρωτεϊνών που δεν περιλαμβάνουν μοναδικά πεπτίδια μελετήθηκε ο αριθμός αυτός των πρωτεϊνών (ανά χρωμόσωμα) ως προς το σύνολο του αριθμού των πρωτεϊνών που περιλαμβάνονται στο χρωμόσωμα. Η ανάλυση αυτή ανέδειξε πως το χρωμόσωμα Y εμφανίζεται με το μεγαλύτερο ποσοστό καθώς το 16% των πρωτεϊνών του δεν περιέχει μοναδικά πεπτίδια (6 πρωτεΐνες). Στη συνέχεια ακολουθούν το χρωμόσωμα 21, X και 13 τα οποία και αυτά εμφανίζουν μεγάλο ποσοστό των πρωτεϊνών τους δεν περιέχει μοναδικά πεπτίδια (Εικόνα 50).

Διακύμανση στα αποτελέσματα παρατηρείται και στον αριθμό από μοναδικά πεπτίδια ελαχίστου μήκους στα χρωμοσώματα του ανθρώπου. Το χρωμόσωμα 1 εμφανίζεται με το μεγαλύτερο αριθμό (719.132) ενώ το Μιτοχονδριακό χρωμόσωμα με τον μικρότερο αριθμό από μοναδικά πεπτίδια ελαχίστου μήκους. Για την καλύτερη κατανόηση του αριθμού από μοναδικά πεπτίδια που περιλαμβάνουν τα χρωμοσώματα, έγινε ανάλυση ως προς την πυκνότητα των χρωμοσωμάτων από CrUP (υπολογίστηκε ο λόγος του συνολικού αριθμού των μοναδικών πεπτιδίων του κάθε χρωμοσώματος ως προς το συνολικό αριθμό από αμινοξέα που αντιστοιχεί στις πρωτεΐνες του χρωμοσώματος). Τα αποτελέσματα αυτά της ανάλυσης έδειξαν πώς το χρωμόσωμα Y έχει την μικρότερη πυκνότητα (19%) από μοναδικά πεπτίδια ελαχίστου μήκους (3.677 πεπτίδια) ενώ αντίθετα το Μιτοχονδριακό χρωμόσωμα εμφανίζεται με τη μεγαλύτερη πυκνότητα (73%) από μοναδικά πεπτίδια ελαχίστου μήκους (2.815 πεπτίδια). Τέλος όλα τα υπόλοιπα χρωμοσώματα εμφανίζονται με πυκνότητα μοναδικών πεπτιδίων ελαχίστου μήκους από 59% έως 68% (Εικόνα 51).

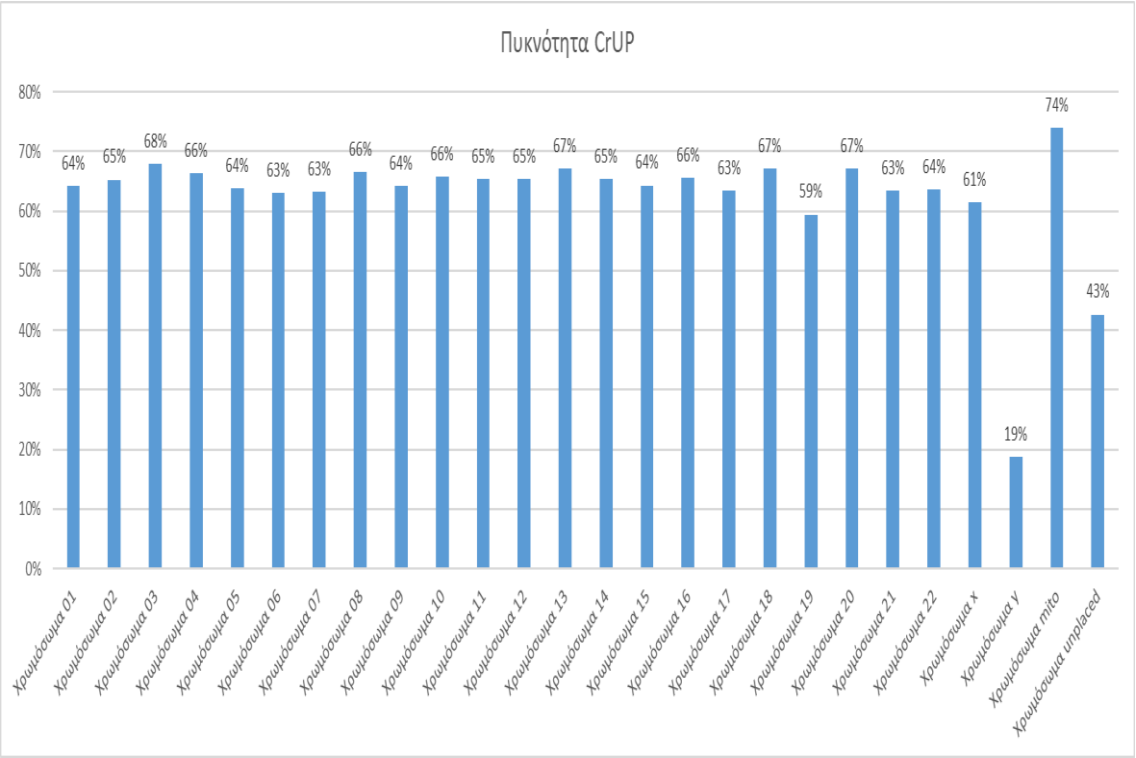
Τέλος, σημαντική διαφοροποίηση εμφανίζεται και στα αποτελέσματα της μοναδικής κάλυψης από μοναδικά πεπτίδια στα χρωμοσώματα του ανθρώπου. Για την ανάλυσης της μοναδικής κάλυψης υπολογίστηκε ο λόγος του συνολικού αριθμού των αμινοξέων που συμμετέχουν έστω μία φορά στον σχηματισμό μοναδικών πεπτιδίων του κάθε χρωμοσώματος ως προς το συνολικό αριθμό από αμινοξέα που αντιστοιχεί στις πρωτεΐνες του χρωμοσώματος. Η ανάλυση αυτή έδειξε πως το Μιτοχονδριακό χρωμόσωμα εμφανίζει 100% μοναδική κάλυψη (όλα τα αμινοξέα των πρωτεϊνών του συμμετέχουν στον σχηματισμό μοναδικών πεπτιδίων) ενώ αντίθετα το χρωμόσωμα Y έχει την μικρότερη μοναδική κάλυψη (35%) από μοναδικά πεπτίδια. Τέλος, τα υπόλοιπα χρωμοσώματα εμφανίζονται με μοναδική κάλυψη από 91% έως 96% (Εικόνα 52).



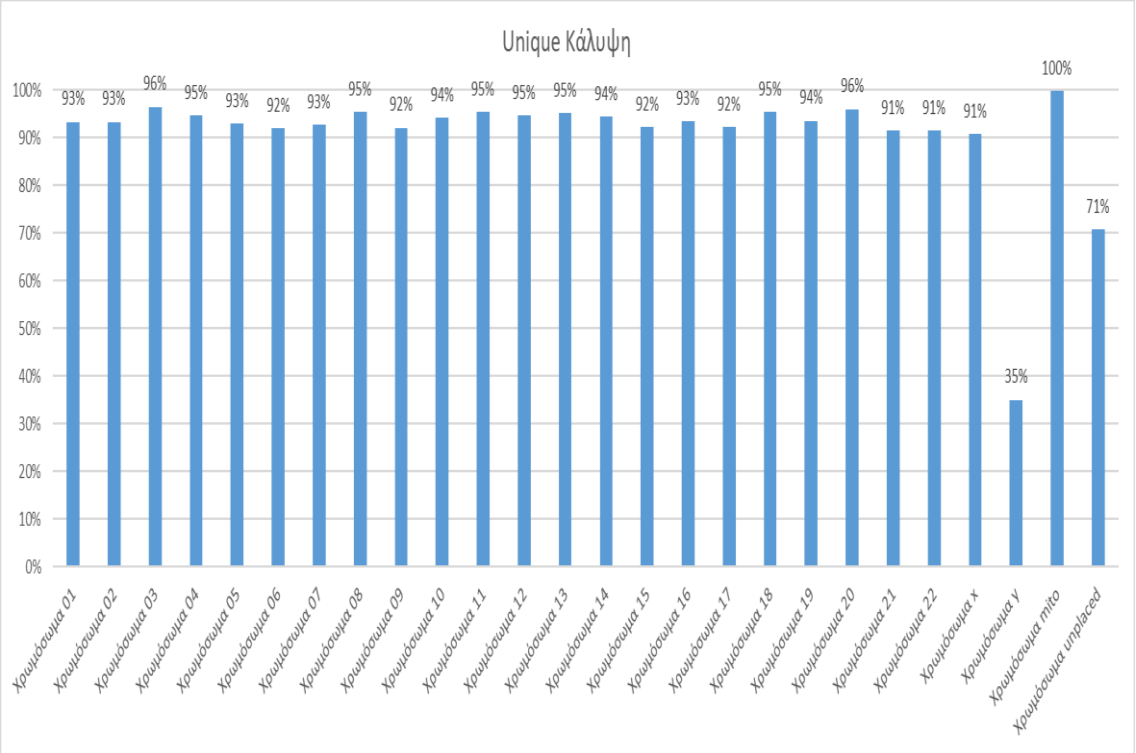
Εικόνα 49 Αριθμός πρωτεϊνών χωρίς μοναδικά πεπτίδια στα χρωμοσώματα το ανθρώπου



Εικόνα 50 Αριθμός πρωτεϊνών (%) χωρίς μοναδικά πεπτίδια στα χρωμοσώματα το ανθρώπου



Εικόνα 51 Πυκνότητα μοναδικών πεπτιδίων ελαχίστου μήκους στα χρωμοσώματα του ανθρώπου



Εικόνα 52 Μοναδική κάλυψη από μοναδικά πεπτίδια στα χρωμοσώματα του ανθρώπου

Από τις παραπάνω αναλύσεις έντονο ενδιαφέρον παρουσιάζουν τα χαρακτηριστικά μοναδικότητας του Μιτοχονδριακού χρωμοσώματος καθώς:

- Δεν έχει πρωτεΐνες που να μην περιλαμβάνουν μοναδικά πεπτίδια

- Έχει την μεγαλύτερη πυκνότητα από μοναδικά πεπτίδια ελαχίστου μήκους σε σχέση με τα υπόλοιπα χρωμοσώματα (μεγάλο αριθμό από CrUP)
- Έχει το μεγαλύτερο ποσοστό από μοναδική κάλυψη (100%).

Αντιθέτως για τα Χρωμοσώματα 13, 21, X και Y παρατηρείται πώς:

- Έχουν μεγάλο ποσοστό από πρωτεΐνες οι οποίες δεν περιλαμβάνουν μοναδικά πεπτίδια

Και πιο συγκεκριμένα για το χρωμόσωμα Y παρατηρείται πως:

- Έχει την μικρότερη πυκνότητα από μοναδικά πεπτίδια ελαχίστου μήκους σε σχέση με τα υπόλοιπα χρωμοσώματα
- Έχει το μικρότερο ποσοστό από μοναδική κάλυψη.

Από τα παραπάνω συμπεραίνεται πως οι πρωτεΐνες που βρίσκονται στο Μιτοχονδριακό χρωμόσωμα παρουσιάζουν μεγάλη μοναδικότητα ως προς την αμινοξική τους αλληλουχία σε σχέση με το υπόλοιπο πρωτέωμα του ανθρώπου, ενώ αντίθετα οι πρωτεΐνες του χρωμοσώματος Y παρουσιάζουν μικρή μοναδικότητα ως προς την αμινοξική τους αλληλουχία σε σχέση με το υπόλοιπο πρωτέωμα του ανθρώπου.

4.1.4 Μοναδικά πεπτίδια και οικογένειες πρωτεϊνών

Σε συνέχεια της ανάλυσης των αποτελεσμάτων των μοναδικών πεπτιδίων του ανθρώπου, μελετήθηκαν τα μοναδικά πεπτίδια βάση στην οικογένεια πρωτεϊνών του ανθρώπου που ανήκει η πρωτεΐνη από την οποία προέρχεται η αμινοξική τους αλληλουχία.

Χρησιμοποιώντας τα εργαλεία της Βάσης δεδομένων Uniprot ομαδοποιήθηκαν οι πρωτεΐνες του ανθρώπου βάση την ομάδα οικογένεια πρωτεϊνών στην οποία ανήκουν. Στην συνέχεια χρησιμοποιώντας τη Βάση δεδομένων του ανθρώπινου Uniquome (Human Uniquome) εντοπίστηκαν τα μοναδικά πεπτίδια που αντιστοιχούν σε κάθε οικογένεια πρωτεϊνών. Πέρα από την απλή καταγραφή του αριθμού των μοναδικών πεπτιδίων για κάθε οικογένεια πρωτεϊνών συλλέχθηκαν και τα χαρακτηριστικά (Πίνακας 8) :

- Αριθμός πρωτεϊνών ανά οικογένεια πρωτεϊνών χωρίς μοναδικά πεπτίδια
- Πυκνότητα μοναδικών πεπτιδίων ελαχίστου μήκους ανά οικογένεια πρωτεϊνών
- Πυκνότητα σύνθετων μοναδικών πεπτιδίων ανά οικογένεια πρωτεϊνών
- Συνολική κάλυψη μοναδικών πεπτιδίων ανά οικογένεια πρωτεϊνών

Οικογένεια Πρωτεϊνών	Πρωτεΐνες	Πρωτεΐνες χωρίς Unique	CrUP	CmUP	Πυκνότητα CrUP	Πυκνότητα CmUP	Unique κάλυψη
G-protein coupled receptor 1 family	724	5	137.815	3.081	55%	1,22%	88%
Krueppel C2H2-type zinc-finger protein family	544	0	170.255	2.952	50%	0,87%	93%
Protein kinase superfamily	490	0	265.716	3.553	63%	0,84%	93%
Small GTPase superfamily	162	0	19.844	560	52%	1,46%	82%
Immunoglobulin superfamily	130	0	44.976	718	59%	0,94%	90%
Peptidase S1 family	119	2	33.102	429	60%	0,77%	93%
Protein-tyrosine phosphatase family	93	0	47.984	581	66%	0,80%	96%
Tyr protein kinase family	90	0	55.711	759	65%	0,88%	96%
MHC class I family	85	0	3.012	291	10%	0,96%	37%
TRAFAC class myosin-kinesin ATPase superfamily	85	0	74.899	1.165	59%	0,92%	90%
Major facilitator superfamily	76	0	26.138	272	67%	0,70%	96%
Peptidase C19 family	76	5	37.149	564	55%	0,83%	80%
Intermediate filament family	75	0	19.576	539	48%	1,33%	83%
CAMK Ser/Thr protein kinase family	75	0	59.153	554	63%	0,59%	94%
Rab family	74	0	9.264	273	51%	1,49%	81%
Ser/Thr protein kinase family	67	0	34.545	305	67%	0,59%	95%
TRIM/RBCC family	64	0	20.682	261	60%	0,76%	87%
CMGC Ser/Thr protein kinase family	62	0	19.895	433	56%	1,21%	86%
Class I-like SAM-binding methyltransferase superfamily	62	0	19.810	135	69%	0,47%	95%
DEAD box helicase family	60	0	31.102	289	63%	0,59%	91%

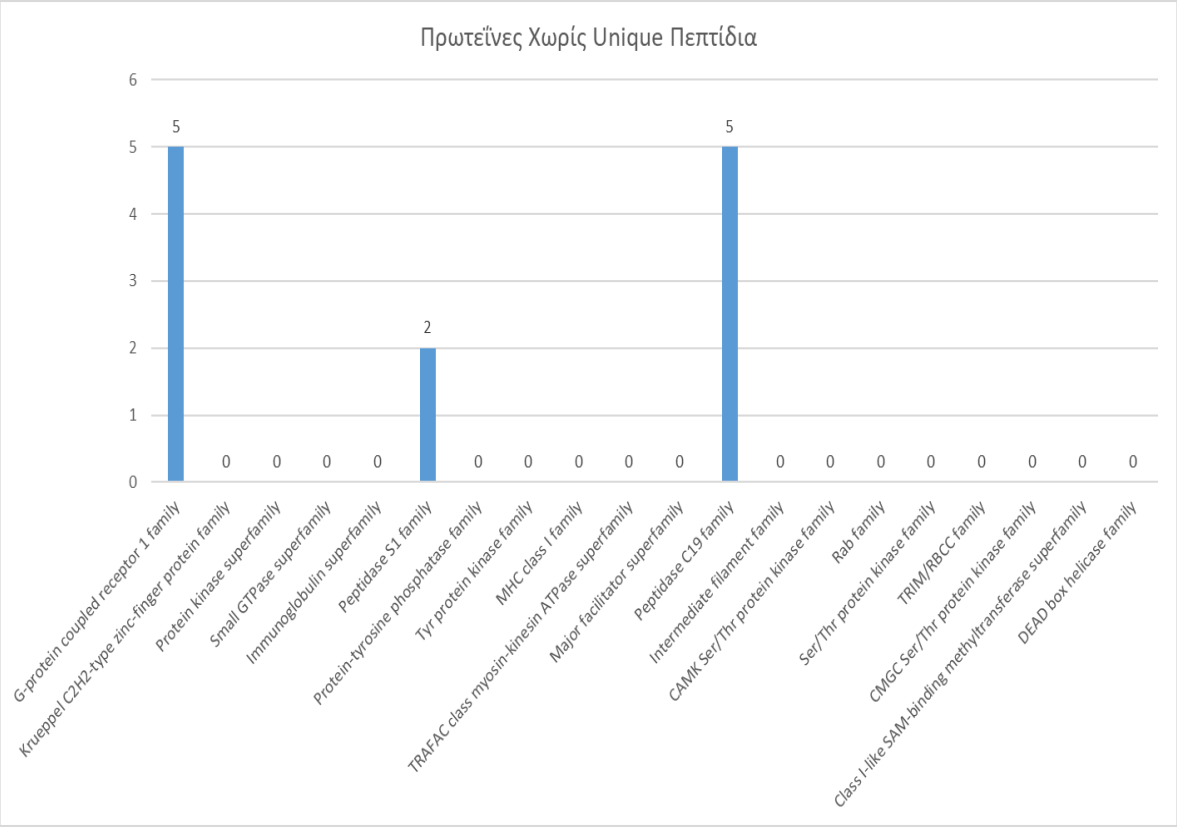
Πίνακας 8 Κατηγοριοποίηση μοναδικών πεπτιδίων στις οικογένειες πρωτεϊνών του ανθρώπου (για τις 20 μεγαλύτερες σε αριθμό πρωτεϊνών)

Όπως παρατηρείται από την παρουσίαση των αποτελεσμάτων των μοναδικών πεπτιδίων βάση την οικογένεια πρωτεϊνών που ανήκουν, στην οποία ανήκουν, τα χαρακτηριστικά που μελετήθηκαν δεν εμφανίζουν κάποια συγκεκριμένη ομοιομορφία. Περαιτέρω ανάλυση των χαρακτηριστικών των μοναδικών πεπτιδίων στις οικογένειες πρωτεϊνών ανέδειξε πώς τα χαρακτηριστικά αυτά συνδέονται άμεσα με την ομοιότητα σε αλληλουχία αμινοξέων που εμφανίζουν οι πρωτεΐνες της εκάστοτε οικογένειας.

Η ανάλυση που πραγματοποιήθηκε στις 20 μεγαλύτερες οικογένειες (σε αριθμό από πρωτεΐνες που αποτελούνται), οικογένειες έδειξε πώς υπάρχουν μόλις 3 οικογένειες πρωτεϊνών οι οποίες έχουν πρωτεΐνες χωρίς μοναδικά πεπτίδια (Εικόνα 53). Οι οικογένειες αυτές είναι οι:

- G-protein coupled receptor 1 family με 5 πρωτεΐνες.
- Peptidase C19 family με 5 πρωτεΐνες
- Peptidase S1 family με 2 πρωτεΐνες

Σύμφωνα με την ανάλυση που πραγματοποιήθηκε στις παραπάνω οικογένειες, οι πρωτεΐνες τους, εμφανίζονται σε μεγάλο ποσοστό με όμοια αμινοξική αλληλουχία.



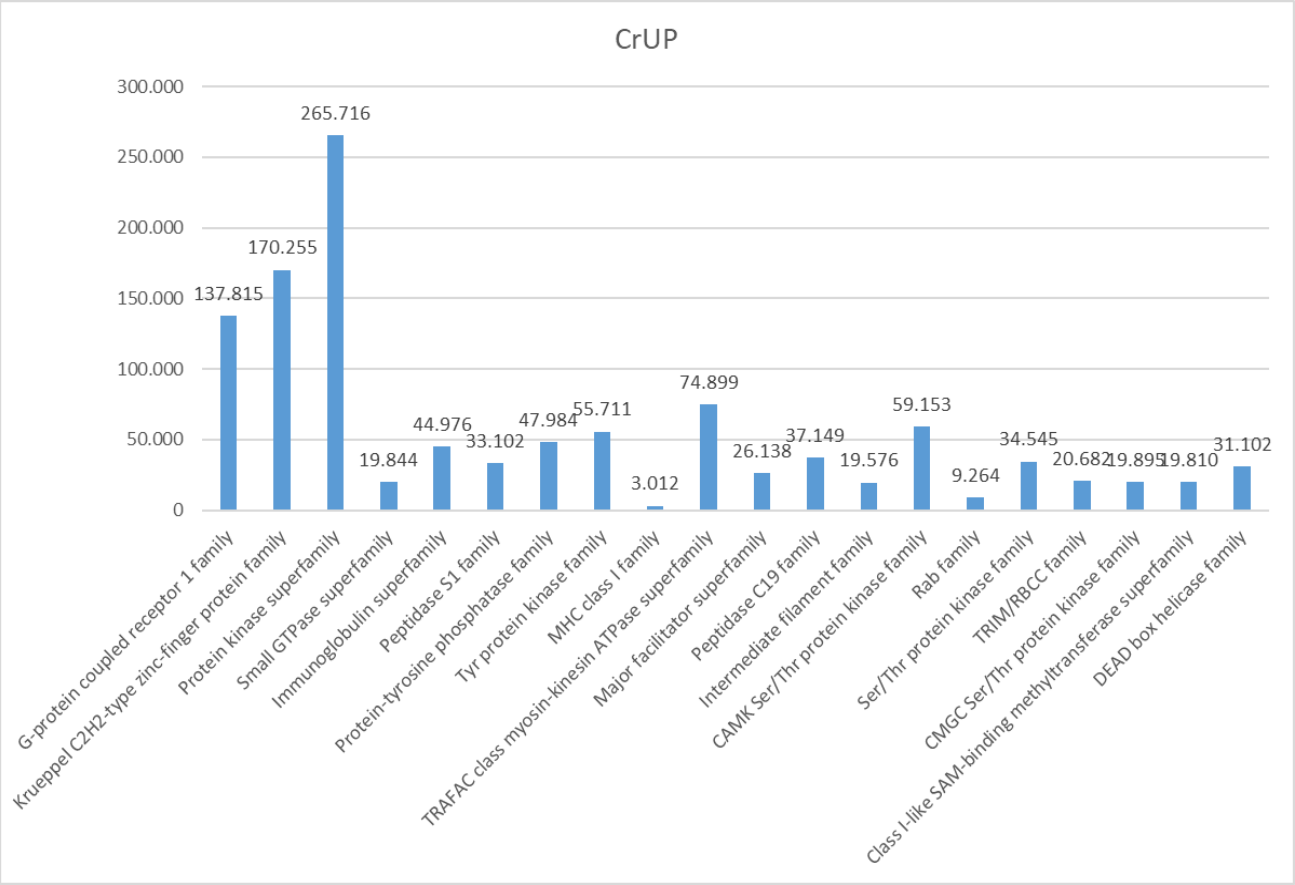
Εικόνα 53 Αριθμός πρωτεϊνών χωρίς μοναδικά πεπτίδια στις οικογένειες πρωτεϊνών του ανθρώπου

Μεγάλη διακύμανση εμφανίζεται και στις οικογένειες πρωτεϊνών ως προς την πυκνότητα των μοναδικών πεπτιδίων ελαχίστου μήκους. Πιο συγκεκριμένα στην πυκνότητα των μοναδικών πεπτιδίων ελαχίστου μήκους υπάρχουν ομάδες όπως η MHC class I family που έχει πυκνότητα από CrUP 10% με 3.012 πεπτίδια αλλά και ομάδες όπως η Class I-like SAM-binding methyltransferase superfamily που έχει πυκνότητα 69% με 19.810 πεπτίδια (Εικόνα 54, Εικόνα 55). Οι οικογένειες πρωτεϊνών στις 20 μεγαλύτερες, σε αριθμό από πρωτεΐνες, με χαμηλή πυκνότητα από μοναδικά πεπτιδίου ελαχίστου μήκους είναι οι :

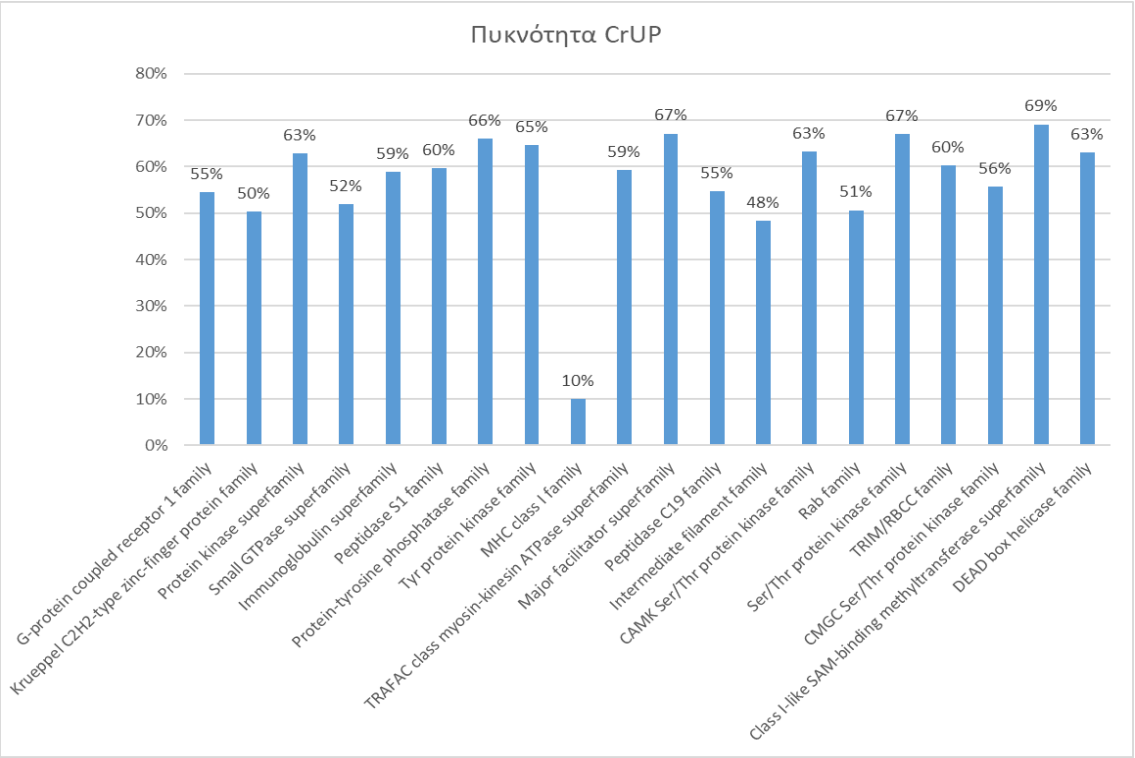
- MHC class I family (10%)
- Intermediate filament family (48%)
- Krueppel C2H2-type zinc-finger protein family (50%)

Αντίθετα οι οικογένειες πρωτεϊνών με την μεγαλύτερη πυκνότητα είναι οι :

- Class I-like SAM-binding methyltransferase superfamily (69%)
- Ser/Thr protein kinase family (67%)
- Major facilitator superfamily (67%)



Εικόνα 54 Αριθμός μοναδικών πεπτιδίων ελαχίστου μήκος στις οικογένειες πρωτεϊνών του ανθρώπου



Εικόνα 55 Πυκνότητα μοναδικών πεπτιδίων ελαχίστου μήκους στις οικογένειες πρωτεϊνών του ανθρώπου

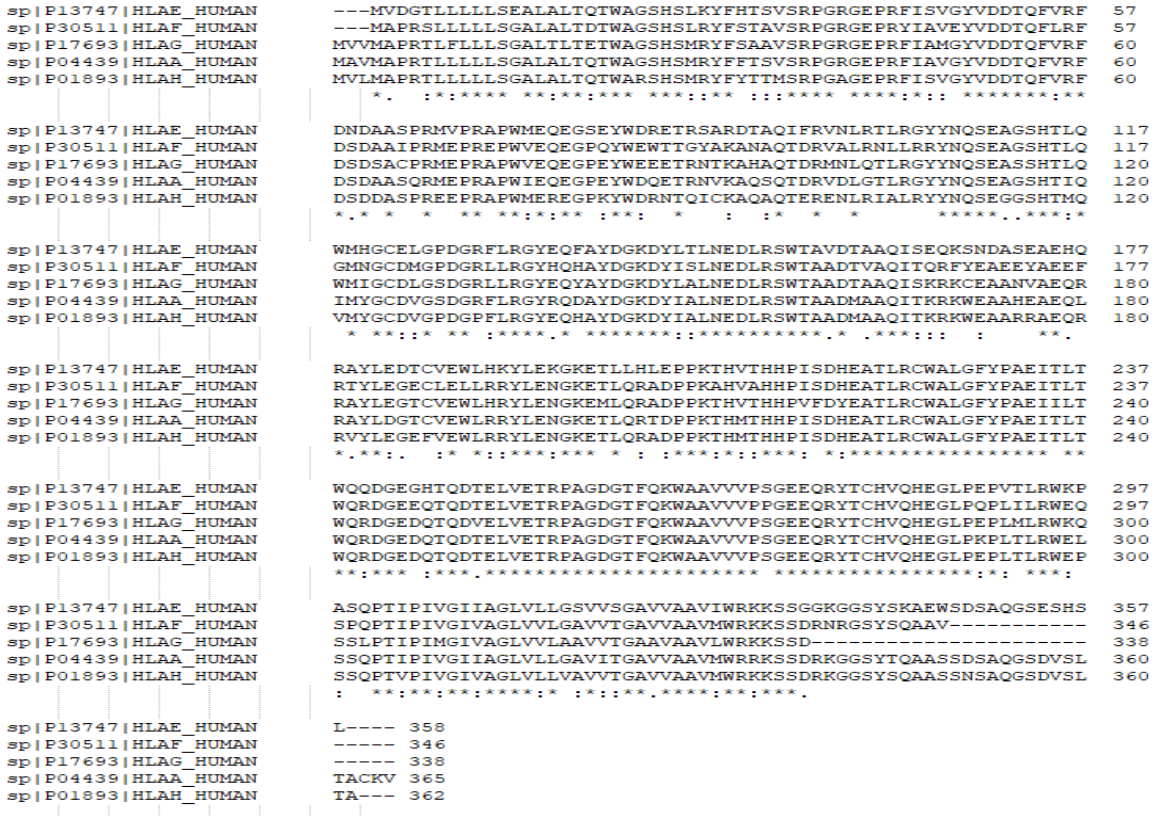
Με σκοπό την κατανόηση των χαμηλών ποσοστών πυκνότητας από μοναδικά πεπτίδια ελαχίστου μήκους της οικογένειας πρωτεϊνών MHC Class I αναλύθηκαν περαιτέρω τρεις ομάδες της. Συγκεκριμένα πραγματοποιήθηκε ευθυγράμμιση των αμινοξικών ακολουθιών των ομάδων πρωτεϊνών MIC, (που περιλαμβάνει τις πρωτεΐνες MICA και MICB) ULBP (που περιλαμβάνει τις πρωτεΐνες ULBP1, ULBP2, ULBP3, ULBP5 και ULBP6) και της ομάδας πρωτεϊνών HLA (που περιλαμβάνει τις πρωτεΐνες HLAA, HLAE, HLAF, HLAG και HLAH) και υπολογισμός του ποσοστού ομοιότητας μεταξύ τους. Τα αποτελέσματα έδειξαν πως οι πρωτεΐνες της ομάδας MIC είχαν μεταξύ τους ομοιότητα 82,51% ενώ η ομάδα πρωτεϊνών ULBP εμφανίζεται με ποσοστά ομοιότητας από 54,32 έως 94,31% ανά ζεύγη. Τέλος οι πρωτεΐνες της ομάδας πρωτεϊνών HLA εντοπίστηκαν με ποσοστά ομοιότητα από 72,54% έως 85,64% (Εικόνα 56,57 και 58) .

```
sp|Q29980|MICB_HUMAN      MGLGRVLLFLAVAFPFAPPAAAAEPHSLRYNLMVLSQDGSVQSGFLAEGHLDGQPFLRYD  60
sp|Q29983|MICA_HUMAN      MGLGPVFLLLAGIFPFAPPGAAAEPHSLRYNLTVLSWDGSVQSGFLTEVHLDGQPFLRCD  60
                          **** *:*:**  ******.************ *** *********:* ********* *

sp|Q29980|MICB_HUMAN      RQKRRAKPQGQWAENVLGAKTWDTETEDLTENGQDLRRTLTHIKDQKGGLHSLQEIRVCE  120
sp|Q29983|MICA_HUMAN      RQKCRAKPQGQWAEDVLGNKTWDRETRDLTGNGKDLRMTLAHIKDQKEGLHSLQEIRVCE  120
                          *** **********:*** **** **.*** **:*** **:****** ************

sp|Q29980|MICB_HUMAN      IHEDSSTRGSRHFYYDGELFLSQNLETQESTVPQSSRAQTLAMNVTNFWKEDAMKTKTHY  180
sp|Q29983|MICA_HUMAN      IHEDNSTRSSQHFYYDGELFLSQNLETKEWTMPQSSRAQTLAMNVRNFLKEDAMKTKTHY  180
                          ****.***.*:*****************:* *:************ ** **********

sp|Q29980|MICB_HUMAN      RAMQADCLQKLQRYLKSGVAIRRTVPPMVNVTCSEVSEGNITVTCRASSFYPRNITLTWR  240
sp|Q29983|MICA_HUMAN      HAMHADCLQELRRYLKSGVVLRRTVPPMVNVTRSEASEGNITVTCRASGFYPWNITLSWR  240
                          :**:*****:*:*******.:*********** **.************.*** ****:**

sp|Q29980|MICB_HUMAN      QDGVSLSHNTQQWGDVLPDGNGTYQTWVATRIRQGEEQRFTCYMEHSGNHGTHPVPSGKA  300
sp|Q29983|MICA_HUMAN      QDGVSLSHDTQQWGDVLPDGNGTYQTWVATRICQGEEQRFTCYMEHSGNHSTHPVPSGKV  300
                          ********:************************ ******************.********.

sp|Q29980|MICB_HUMAN      LVLQSQRTDFPYVSAAMPCFVIIIILCVPCCKKKTSAAEGPELVSLQVLDQHPVGTGDHR  360
sp|Q29983|MICA_HUMAN      LVLQSHWQTFHVSAVAAAAIFVIIIFYVRCCKKKTSAAEGPELVSLQVLDQHPVGTSDHR  360
                          *****:   *   :.*  .:.:***: * ****************************.***

sp|Q29980|MICB_HUMAN      DAAQLGFQPLMSATGSTGSTEGT   383
sp|Q29983|MICA_HUMAN      DATQLGFQPLMSDLGSTGSTEGA   383
                          **:*********  ********:
```

Εικόνα 56 Αμινοξική ευθυγράμμιση των πρωτεϊνών της ομάδας MIC της οικογένειας MHC I.

```
sp|Q9BZM4|ULBP3_HUMAN     MAAAASPAILPRLAILPYLLFDWSGTGRADAHSLWYNFTIIHLPRHGQQWCEVQSQVDQK 60
sp|Q9BZM6|ULBP1_HUMAN     MAAAASPAFLLCLPLL-HLLSGWSRAGWVDTHCLCYDFIITPKSRPEPQWCEVQGLVDER 59
sp|Q6H3X3|ULBP5_HUMAN     MAAAASPAFLLRLPLL-LLLSSWCRTGLADPHSLCYDITVIPKFRPGPRWCAVQGQVDEK 59
sp|Q9BZM5|ULBP2_HUMAN     MAAAAATKILLCLPLL-LLLSGWSRAGRADPHSLCYDITVIPKFRPGPRWCAVQGQVDEK 59
sp|Q5VY80|ULBP6_HUMAN     MAAAAIPALLLCLPLL-FLLFGWSRARRDDPHSLCYDITVIPKFRPGPRWCAVQGQVDEK 59
                          *****   :*  * :*  ** .*. :   * *.* *:: :    *   :** **. **::

sp|Q9BZM4|ULBP3_HUMAN     NFLSYDCGSDKVLSMGHLEEQLYATDAWGKQLEMLREVGQRLRLELADTELEDFTPSGPL 120
sp|Q9BZM6|ULBP1_HUMAN     PFLHYDCVNHKAKAFASLGKKVNVTKTWEEQTETLRDVVDFLKGQLLDIQVENLIPIEPL 119
sp|Q6H3X3|ULBP5_HUMAN     TFLHYDCGSKTVTPVSPLGKKLNVTTAWKAQNPVLREVVDILTEQLLDIQLENYIPKEPL 119
sp|Q9BZM5|ULBP2_HUMAN     TFLHYDCGNKTVTPVSPLGKKLNVTTAWKAQNPVLREVVDILTEQLRDIQLENYTPKEPL 119
sp|Q5VY80|ULBP6_HUMAN     TFLHYDCGNKTVTPVSPLGKKLNVTMAWKAQNPVLREVVDILTEQLLDIQLENYTPKEPL 119
                           ** *** .... .. * ::: .* :*  *   **:* : *  :* * ::*:  *  **

sp|Q9BZM4|ULBP3_HUMAN     TLQVRMSCECEADGYIRGSWQFSFDGRKFLLFDSNNRKWTVVHAGARRMKEKWEKDSGLT 180
sp|Q9BZM6|ULBP1_HUMAN     TLQARMSCEHEAHGHGRGSWQFLFNGQKFLLFDSNNRKWTALHPGAKKMTEKWEKNRDVT 179
sp|Q6H3X3|ULBP5_HUMAN     TLQARMSCEQKAEGHGSGSWQLSFDGQIFLLFDSENRMWTTVHPGARKMKEKWENDKDMT 179
sp|Q9BZM5|ULBP2_HUMAN     TLQARMSCEQKAEGHSSGSWQFSFDGQIFLLFDSEKRMWTTVHPGARKMKEKWENDKVVA 179
sp|Q5VY80|ULBP6_HUMAN     TLQARMSCEQKAEGHSSGSWQFSIDGQTFLLFDSEKRMWTTVHPGARKMKEKWENDKDVA 179
                          ***.***** :*.*:  ****: ::*: ******::* **.:* **::*.****::  ::

sp|Q9BZM4|ULBP3_HUMAN     TFFKMVSMRDCKSWLRDFLMHRKKRLEPT--APPTMAPGLAQPKAIATTLSPWSFLII-L 237
sp|Q9BZM6|ULBP1_HUMAN     MFFQKISLGDCKMWLEEFLMYWEQMLDPT--KPPSLAPGTTQPKAMATTLSPWSLLIIFL 237
sp|Q6H3X3|ULBP5_HUMAN     MSFHYISMGDCTGWLEDFLMGMDSTLEPSAGAPPTMSSGTAQPRATATTLILCCLLIMCL 239
sp|Q9BZM5|ULBP2_HUMAN     MSFHYFSMGDCIGWLEDFLMGMDSTLEPSAGAPLAMSSGTTQLRATATTLILCCLLIILP 239
sp|Q5VY80|ULBP6_HUMAN     MSFHYISMGDCIGWLEDFLMGMDSTLEPSAGAPLAMSSGTTQLRATATTLILCCLLIILP 239
                            *: .*: **  **.:***  .. *:*:   * ::: * :* :* ****   .:**:

sp|Q9BZM4|ULBP3_HUMAN     CFILPGI---------------------------------------------------- 244
sp|Q9BZM6|ULBP1_HUMAN     CFILAGR---------------------------------------------------- 244
sp|Q6H3X3|ULBP5_HUMAN     LICSRHSLTQSHGHHPQSLQPPPHPPLLHPTWLLRRVLWSDSYQIAKRPLSGGHVTRVTL 299
sp|Q9BZM5|ULBP2_HUMAN     CFILPGI---------------------------------------------------- 246
sp|Q5VY80|ULBP6_HUMAN     CFILPGI---------------------------------------------------- 246
                           :

sp|Q9BZM4|ULBP3_HUMAN     ----------------------------------  244
sp|Q9BZM6|ULBP1_HUMAN     ----------------------------------  244
sp|Q6H3X3|ULBP5_HUMAN     PIIGDDSHSLPCPLALYTINNGAARYSEPLQVSIS  334
sp|Q9BZM5|ULBP2_HUMAN     ----------------------------------  246
sp|Q5VY80|ULBP6_HUMAN     ----------------------------------  246
```

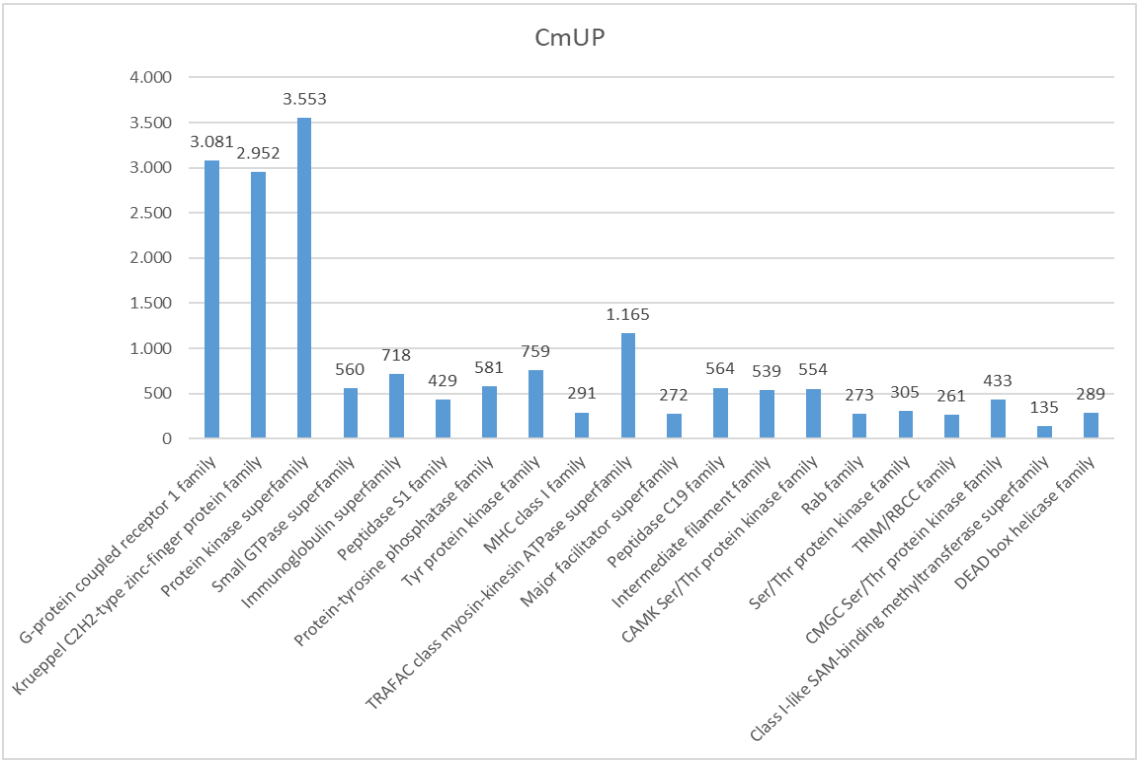
Εικόνα 57 Αμινοξική ευθυγράμμιση των πρωτεϊνών της ομάδας ULBP της οικογένειας MHC I.



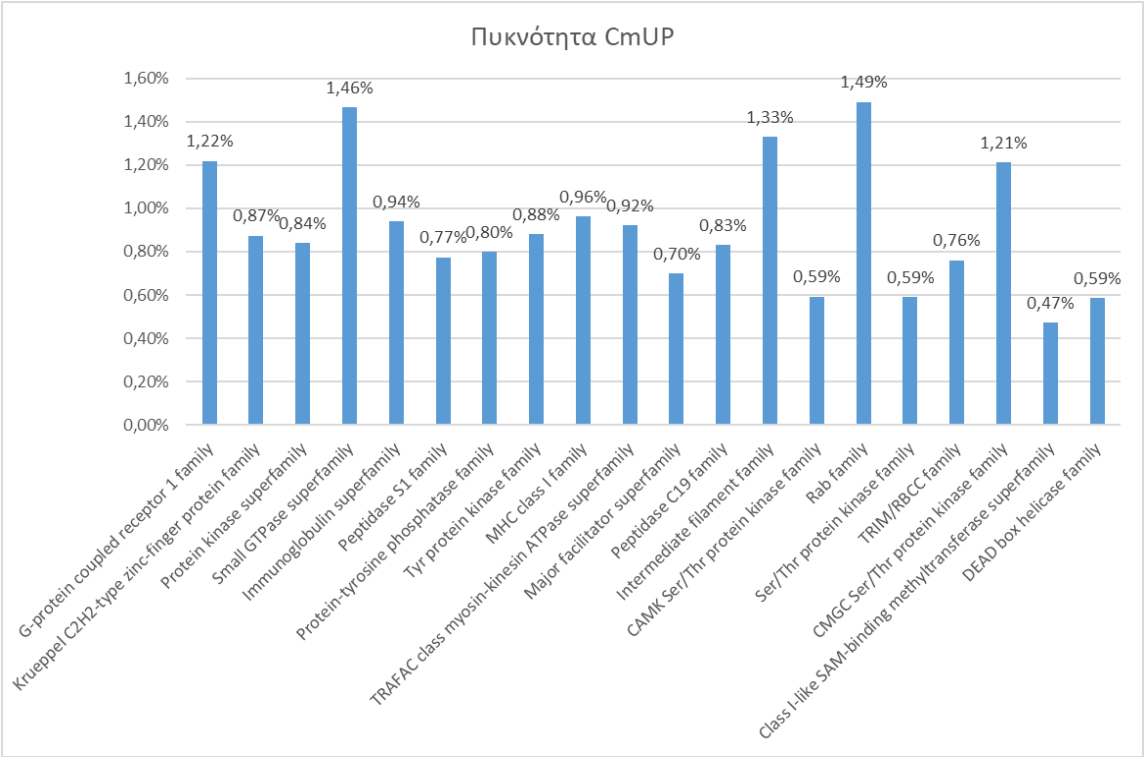
Εικόνα 58 Αμινοξική ευθυγράμμιση των πρωτεϊνών της ομάδας HLA της οικογένειας MHC I.

Τα αποτελέσματα της ανάλυσης των σύνθετων μοναδικών πεπτιδίων ως προς την οικογένεια των πρωτεϊνών που ανήκουν τα πεπτίδια αυτά, οδηγούν στα ίδια συμπεράσματα με την αντίστοιχη ανάλυση των μοναδικών πεπτιδίων ελαχίστου μήκους. Αυτό που παρατηρείται και στην συγκεκριμένη ανάλυση είναι ότι ο αριθμός των σύνθετων μοναδικών πεπτιδίων εξαρτάται από την ομοιότητα σε αμινοξική αλληλουχία που εμφανίζουν οι πρωτεΐνες της κάθε οικογενείας στην οποία ανήκουν. Η οικογένεια πρωτεϊνών με την μεγαλύτερη πυκνότητα από σύνθετα μοναδικά πεπτίδια είναι η Rab Family που εμπεριέχει 273 πεπτίδια ενώ η οικογένεια πρωτεϊνών με την μικρότερη πυκνότητα από σύνθετα μοναδικά πεπτίδια είναι η Class I-like SAM-binding methyltransferase superfamily που εμπεριέχει 135 πεπτίδια (Εικόνα 59, εικόνα 60).

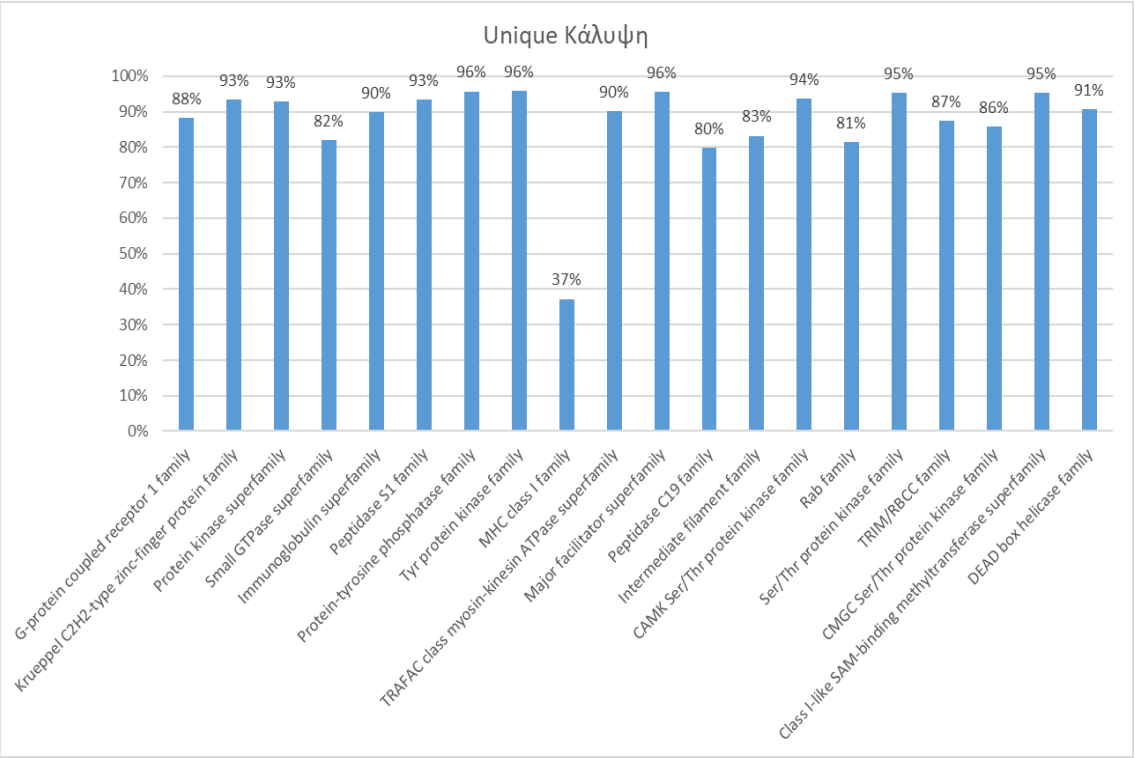
Τέλος, ένα άλλο χαρακτηριστικό που μελετήθηκε για ως προς την μοναδικότητα στις οικογένειες πρωτεϊνών για τον άνθρωπο είναι η μοναδική κάλυψη. Όπως παρατηρείται και σε αυτό το χαρακτηριστικό τα αποτελέσματα εμφανίζονται με μεγάλη διακύμανση. Η οικογένεια πρωτεϊνών MHC class I ξεχωρίζει καθώς έχει την μικρότερη κάλυψη από μοναδικά αμινοξέα (37%) ενώ αντίθετα οι οικογένειες protein – tyrosine phosphatase family, Tyr protein kinase family και Major facilitator superfamily εμφανίζονται με τα μεγαλύτερα ποσοστά (96%) από μοναδική κάλυψη (Εικόνα 61).



Εικόνα 59 Αριθμός σύνθετων μοναδικών πεπτιδίων στις οικογένειες πρωτεϊνών του ανθρώπου



Εικόνα 60 Πυκνότητα σύνθετων μοναδικών πεπτιδίων στις οικογένειες πρωτεϊνών του ανθρώπου



Εικόνα 61 Μοναδική κάλυψη από μοναδικά πεπτίδια στις οικογένειες πρωτεϊνών του ανθρώπου

Για την καλύτερη κατανόηση των μοναδικών πεπτιδίων και τον τρόπο εμφάνισης τους στις οικογένειες πρωτεϊνών του ανθρώπου, μελετήθηκαν πιο αναλυτικά η υπό-οικογένεια πρωτεϊνών Ras(H-Ras, K-Ras και N-Ras) η οποία ανήκει στην οικογένεια Small GTPase superfamily.

Η οικογένεια πρωτεϊνών Ras

Οι RAS πρωτεΐνες ανήκουν στην υπεροικογένεια των μικρών GTP-άσων και εναλλάσσονται μεταξύ μιας ανενεργής (που προσδένει το GDP) και μιας ενεργής (που προσδένει το GTP) μορφής. Οι μεταλλάξεις του γονιδίου οδηγούν συνήθως σε πρωτεΐνη που βρίσκεται συνεχώς στην ενεργή μορφή και ενεργοποιεί σηματοδοτικά μονοπάτια που ελέγχουν τον πολλαπλασιασμό, την απόπτωση και άλλες κυτταρικές λειτουργίες.

Η οικογένεια πρωτεϊνών Ras (H-Ras, K-Ras και N-Ras) μελετήθηκε εκτεταμένα ως προς τα χαρακτηριστικά των μοναδικών πεπτιδίων της. Αρχικά, με τη βοήθεια εργαλείων από την βάση δεδομένων της Uniprot έγινε αμινοξική ευθυγράμμιση (alignment) στις αλληλουχίες των 3 αυτών πρωτεϊνών (Εικόνα 62). Όπως διαπιστώθηκε από την στοίχιση, στο μεγαλύτερο μέρος των ακολουθιών τους και στις τρείς πρωτεΐνες εντοπίζεται πανομοιότυπη σύσταση από αμινοξέα. Η ακριβής ομοιότητα στην αμινοξική αλληλουχία των πρωτεϊνών εντοπίζεται έντονα στο πρώτο μισό των πρωτεϊνών.

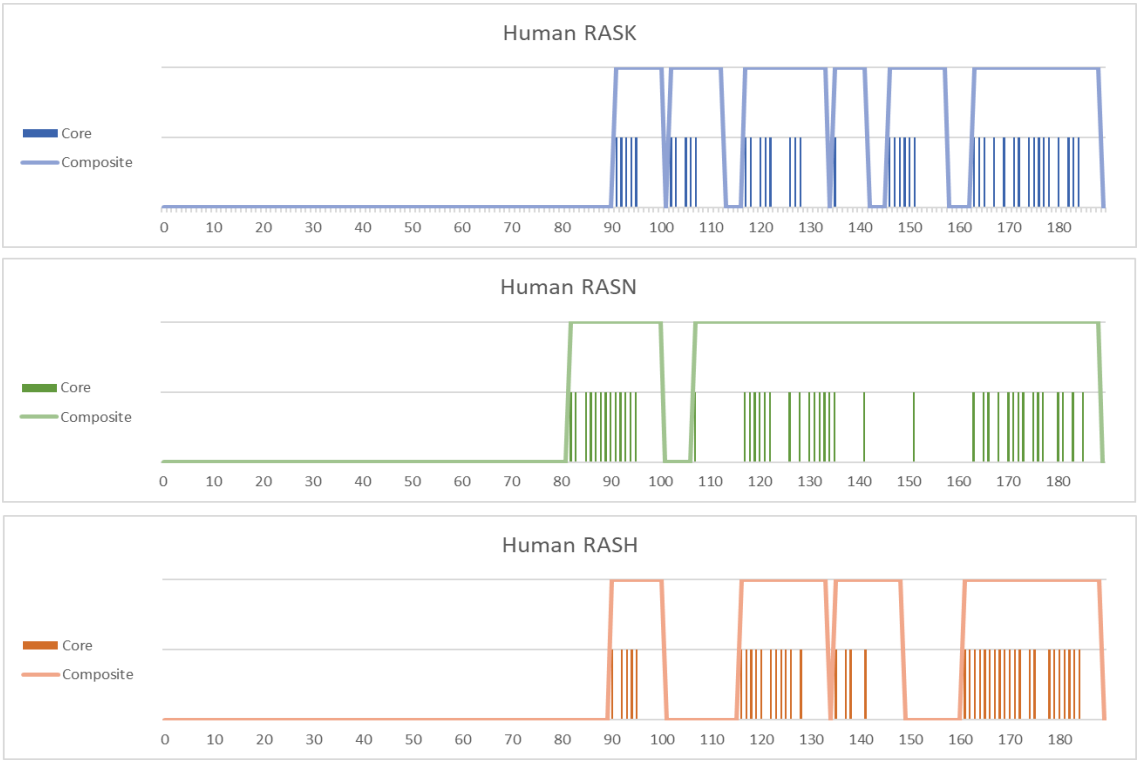
Εικόνα 62 Αμινοξική ευθυγράμμιση (alignment) αμινοξικών ακολουθιών στην ομάδα των πρωτεϊνών Ras (K-Ras, H-Ras και N-Ras) στον οργανισμό του ανθρώπου

Στην συνέχεια της ανάλυσης, καταγράφηκαν και μελετήθηκαν τα χαρακτηριστικά της μοναδικότητας (CrUP, CmUP, πυκνότητα καθώς και μοναδική κάλυψη) των πρωτεϊνών της οικογένειας Ras (H-Ras, K-Ras και N-Ras). Τα αποτελέσματα έδειξαν πώς και οι 3 πρωτεΐνες εμφανίζουν μικρό αριθμό από μοναδικά πεπτίδια τα οποία έχουν και πανομοιότυπα χαρακτηρίστηκα. Πιο συγκεκριμένα τόσο τα μοναδικά πεπτίδια ελαχίστου μήκους όσο και τα σύνθετα μοναδικά πεπτίδια είναι σχεδόν τα ίδια σε αριθμό και στις 3 πρωτεΐνες (Πίνακας 9).

Homo Sapiens	RASK	RASN	RASH
Μήκος Πρωτεΐνης	189	189	189
CrUP	41	45	41
CmUP	6	2	4
Πυκνότητα CrUP	22%	24%	22%
Πυκνότητα CmUP	3,17%	1,05%	2,11%
Μοναδική Κάλυψη	44%	54%	38%

Πίνακας 9 Χαρακτηρίστηκα μοναδικότητας στην οικογένεια πρωτεϊνών Ras (k-ras, n-ras, h-ras)

Στο επόμενο στάδιο της ανάλυσης αποτυπώθηκαν οι θέσεις εμφάνισης των μοναδικών πεπτιδίων (CrUP και CmUP) και στις τρεις υπό μελέτη πρωτεΐνες με σκοπό την καλύτερη απεικόνιση των αποτελεσμάτων για τα χαρακτηριστικά της μοναδικότητας τους. Η απεικόνιση αυτή ανέδειξε πως η ομοιομορφία στη θέση εμφάνισης των μοναδικών πεπτιδίων στις πρωτεΐνες της οικογένειας Ras αντιστοιχεί με την ομοιότητα που παρατηρήθηκε από το alignment των πρωτεϊνών. Αναλυτικότερα, οι πρωτεΐνες στο πρώτο μισό της αμινοξικής τους αλληλουχίας δεν εμφανίζουν κανένα μοναδικό πεπτίδιο (κάτι τέτοιο ήταν αναμενόμενο καθώς έχουν την ίδια αμινοξική αλληλουχία) σε αντίθεση με το δεύτερο μισό των πρωτεϊνών στο οποίο εμφανίζονται τα μοναδικά τους πεπτίδια. Επιπλέον τα μοναδικά πεπτίδια στις περισσότερες περιπτώσεις εμφανίζονται ακριβώς στην ίδια θέση μέσα στην αμινοξική αλληλουχία των πρωτεϊνών (Εικόνα 63).



Εικόνα 63 Θέση εμφάνισης μοναδικών πεπτιδίων στην οικογένεια πρωτεϊνών Ras (k-ras, n-ras, h-ras) του ανθρώπου

Τέλος, με σκοπό την περαιτέρω σύγκριση των αποτελεσμάτων τόσο για τη θέση εμφάνισης των πεπτιδίων όσο και για την αμινοξική τους αλληλουχία στις τρεις υπό μελέτη πρωτεΐνες της οικογένειας RAS επιλέχθηκαν και αναλύθηκαν περεταίρω κάποια πεπτίδια (Πίνακας 10). Για την καλύτερη κατανόηση των αποτελεσμάτων στην ανάλυση αναλύθηκαν όχι μόνο τα μοναδικά πεπτίδια, αλλά και πεπτίδια που δεν ήταν μοναδικά αλλά παρουσίαζαν κάποια ιδιαίτερα χαρακτηριστικά (πανομοιότυπη αμινοξική αλληλουχία και ίδια θέση εμφάνισης).

Θέση	K-Ras	N-Ras	H-Ras
64	EEYSAM	EEYSAM	EEYSAM
83	AINNTK	AINNSK*	AINNTK
95	HYREQI*	LYREQI*	QYREQI*
116	KCDLPT*	KCDLPS*	KCDLAA*
135	RSYGIPF*	KSYGIPF*	RSYGIPY*
165	QYRLKK*	QYRMK*	QHKLRK*
184	KKCIIM*	PCVVM*	CKCVLS*

*Πίνακας 10 Σύγκριση αμινοξικής αλληλουχίας σε πεπτίδια (με * παρουσιάζονται τα μοναδικά πεπτίδια ελαχίστου μήκους) στην οικογένεια πρωτεϊνών Ras (k-ras, n-ras, h-ras) του ανθρώπου*

Πιο συγκεκριμένα για τον παραπάνω πίνακα παρατηρείται:

- Στην θέση 64, στις τρείς πρωτεΐνες εμφανίζεται ένα 6-πεπτίδιο με ακριβώς την ίδια αμινοξική αλληλουχία (EEYSAM). Όπως είναι λογικό (από τον ορισμό των μοναδικών πεπτιδίων) το πεπτίδιο αυτό δεν είναι μοναδικό πεπτίδιο. Όμως μια περεταίρω αναζήτηση στο πρωτέωμα του ανθρώπου έδειξε πως το συγκεκριμένο πεπτίδιο εμφανίζεται μόνο στις συγκεκριμένες πρωτεΐνες. Συνεπώς το πεπτίδιο EEYSAM είναι μοναδικό ως προς την οικογένεια πρωτεϊνών Ras (k-ras, n-ras, h-ras) του ανθρώπου. Τέτοια πεπτίδια ορίστηκαν ως Family unique Peptides.
- Στην θέση 83, στις πρωτεΐνες k-ras και h-ras υπάρχει ακριβώς το ίδιο 6-πεπτίδιο (AINNTK) συνεπώς δεν είναι και μοναδικό πεπτίδιο. Αντίθετα στην πρωτεΐνη n-ras

υπάρχει το 6-πεπτίδιο (AINNSK) το οποίο είναι και μοναδικό πεπτίδιο ελαχίστου μήκους (CrUP).

- Στην θέση 95, στις τρεις πρωτεΐνες υπάρχει το πεπτίδιο (XYREQI) όπου X, H για την K-ras, L για την N-ras και Q για την H-ras. Τα τρία αυτά πεπτίδια είναι μοναδικά πεπτίδια ελαχίστου μήκους (CrUP). Στην συγκεκριμένη θέση εντοπίζεται το μοτίβο (XYREQI) το οποίο ανάλογα με το αμινοξύ της πρώτης θέσης (H,L ή Q) χαρακτηρίζει μοναδικά την αντίστοιχη πρωτεΐνη.
- Στην θέση 116 παρουσιάζεται μια παρόμοια περίπτωση με την θέση 95. Εντοπίζεται το μοτίβο (KCDLXX) το οποίο ανάλογα με την αλληλουχία των δύο τελευταίων αμινοξέων (PT, PS ή AA) προσδιορίζεται και η μοναδικότητα της πρωτεΐνης K-ras, N-ras ή H-ras αντίστοιχα αφού με αυτό τον τρόπο δημιουργούνται μοναδικά πεπτίδια ελαχίστου μήκους (CrUP).
- Στην θέση 135 παρουσιάζεται μια ακόμη περίπτωση παρόμοια με τις δύο προηγούμενες. Εντοπίζεται το μοτίβο (XSYGIPX) στο οποίο τη μοναδικότητα καθορίζει το πρώτο και το τελευταίο αμινοξύ. Πιο συγκεκριμένα αν ο συνδυασμός των δύο αυτών αμινοξέων είναι RF τότε χαρακτηρίζεται μοναδικά η πρωτεΐνη K-ras, αν είναι KF η N-ras και αν είναι PY η H-ras αφού με αυτό τον τρόπο δημιουργούνται μοναδικά πεπτίδια ελαχίστου μήκους (CrUP).
- Τέλος, στη θέση 184 υπάρχουν τρία διαφορετικά πεπτίδια ως προς την αμινοξική τους αλληλουχία. Το 6-πεπτίδιο KKCCIIM για την πρωτεΐνη K-ras, το 5-πεπτίδιο PCVVM για την N-ras και το 6-πεπτίδιο CKCVLS για την H-ras. Και τα τρία αυτά πεπτίδια είναι μοναδικά πεπτίδια ελαχίστου μήκους (CrUP) και κάθε ένα από αυτά χαρακτηρίζει μοναδικά την αντίστοιχη πρωτεΐνη στην οποία ανήκει.

4.1.5 Προσομοιωμένο πρωτέωμα (simulated) και το ανθρώπινο Uniquome

Με σκοπό την καλύτερη ερμηνεία των χαρακτηριστικών της μοναδικότητας στο Uniquome του ανθρώπου, κατασκευάστηκε ένα τεχνητό πρωτέωμα (simulated) που εξομοιώνει το ανθρώπινο πρωτέωμα στο οποίο πραγματοποιήθηκε η ανάλυση της μοναδικότητας των πεπτιδίων. Η προσομοίωση του τεχνητού πρωτεώματος βασίσθηκε στα χαρακτηριστικά του πρωτεώματος του ανθρώπου (Uniprot 10/2019). Η μεθοδολογία για την κατασκευή του προσομοιωμένου πρωτεώματος στηρίχτηκε στις μεθόδους Monte Carlo [81,82] ακολουθώντας τις εξής παραδοχές:

- Να έχει τον ίδιο αριθμό πρωτεϊνών με αυτό του ανθρώπινου πρωτεώματος.
- Το μήκος κάθε πρωτεΐνης να αντιστοιχεί στο μήκος μίας πρωτεΐνης του ανθρώπινου πρωτεώματος.

- Σε κάθε θέση μέσα στην πρωτεΐνη η πιθανότητα εμφάνισης του κάθε αμινοξέος να ισούται με την πιθανότητα εμφάνισης του κάθε αμινοξέος στο σύνολο των πρωτεϊνών του ανθρώπινου πρωτεώματος.

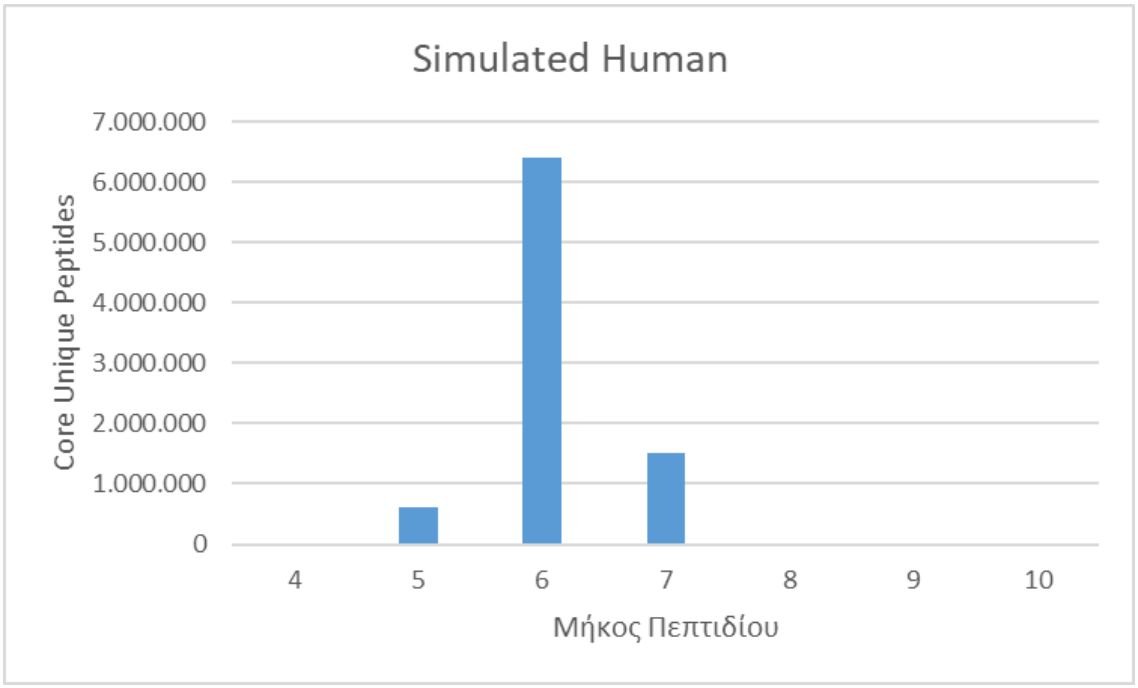
Στο προσομοιωμένο πρωτέωμα του ανθρώπου αναλύθηκαν 20.430 θεωρημένες/επιβεβαιωμένες υποθετικές πρωτεΐνες. Σε 20.426 πρωτεΐνες καταγράφηκαν 8.542.941 μοναδικά πεπτίδια ελαχίστου μήκους (CrUP). Η συνολική πυκνότητα μοναδικών πεπτιδίων ελαχίστου μήκους είναι 75%. Μόλις 182 από το σύνολο των μοναδικών πεπτιδίων ελαχίστου μήκους του προσομοιωμένου ανθρώπινου πρωτεώματος εμφανίζονται παραπάνω από μία φορά στην ίδια πρωτεΐνη (τα πεπτίδια αυτά εξακολουθούν να ορίζονται σαν μοναδικά πεπτίδια καθώς ναι μεν εμφανίζονται παραπάνω από μία φορά αλλά εμφανίζονται μόνο σε μία πρωτεΐνη). Το σύνολο των μοναδικών πεπτιδίων ελαχίστου μήκους δημιουργούν 20.461 σύνθετα μοναδικά πεπτίδια (CmUP), αριθμός ο οποίος αντιστοιχεί σε συνολική πυκνότητα από σύνθετα μοναδικά πεπτίδια 0,18%. Το προσομοιωμένο πρωτέωμα του ανθρώπινου πρωτεώματος έχει 93% συνολική κάλυψη από μοναδικά πεπτίδια. Τέλος, από τις 20.430 υποθετικές πρωτεΐνες υπάρχουν μόνο 4 πρωτεΐνες οι οποίες δεν περιλαμβάνουν κανένα μοναδικό πεπτίδιο (Πίνακας 11).

Simulated Human Uniquome	
Πρωτεΐνες (υποθετικές)	20.430
Πρωτεΐνες με μοναδικά πεπτίδια	20.426
Πρωτεΐνες χωρίς μοναδικά πεπτίδια	4
Μοναδικά πεπτίδια ελαχίστου μήκους	8.542.941
Μοναδικά πεπτίδια ελαχίστου μήκους >1 φορά	182
Σύνθετα μοναδικά πεπτίδια	20.461
Συνολική πυκνότητα μοναδικών πεπτιδίων ελαχίστου μήκους	75%
Συνολική πυκνότητα σύνθετων μοναδικών πεπτιδίων	0,18%
Συνολική Κάλυψη	100%

Πίνακας 11 Χαρακτηριστικά μοναδικότητας για το Uniquome του Προσομοιωμένου ανθρώπινου πρωτεώματος

Μοναδικά πεπτίδια Ελαχίστου μήκους

Στο Simulated πρωτέωμα του ανθρώπου τα μοναδικά πεπτίδια ελαχίστου μήκους έχουν μήκος από 4 – 9 αμινοξέα καθώς σύμφωνα με τα αποτελέσματα δεν υπάρχει κανένα μοναδικό πεπτίδιο το οποίο να έχει μήκος μεγαλύτερο των 9 αμινοξέων. Αναλυτικότερα, η συντριπτική πλειοψηφία των μοναδικών πεπτιδίων ελαχίστου μήκους αποτελείται από 5, 6 και 7 αμινοξέα. Τα μοναδικά 6-πεπτίδια ελαχίστου μήκους είναι τα πεπτίδια που εμφανίζονται με το μεγαλύτερο αριθμό (6.406.417) ενώ ακολουθούν τα πεπτίδια που αποτελούνται από 7 και 5 αμινοξέα με 1.507.402 και 612.531 πεπτίδια αντίστοιχα. Τέλος, για τον simulated πρωτέωμα του ανθρώπου, τα 4, 8 και 9 πεπτίδια είναι ελάχιστα (Εικόνα 64, πίνακας 12).

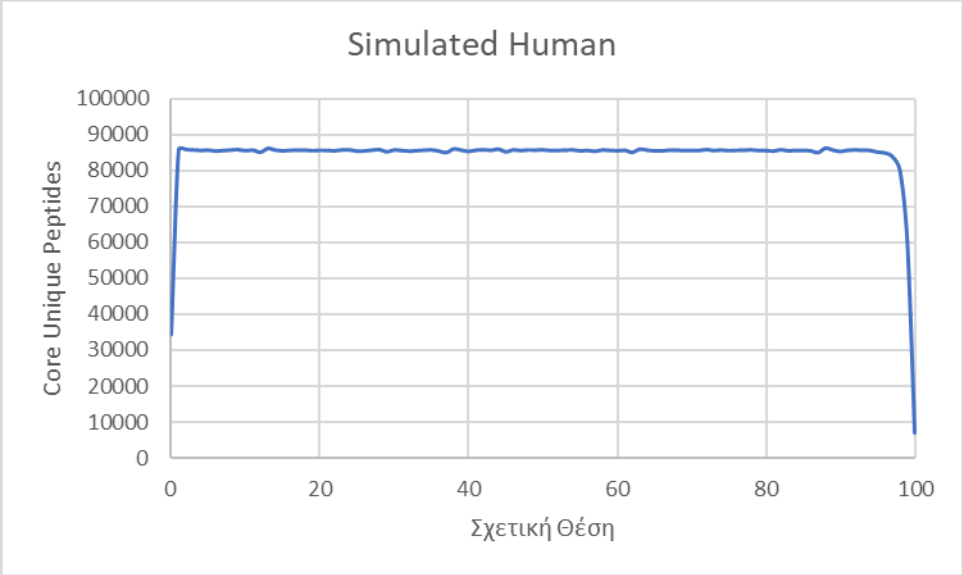


Εικόνα 64 Πλήθος CrUP ανάλογα το μήκος του πεπτιδίου στο Simulated πρωτέωμα του ανθρώπου

Μήκος Πεπτιδίου	Πλήθος	%
4	604	0%
5	612.531	7%
6	6.406.417	75%
7	1.507.402	18%
8	15.928	0%
9	59	0%
10	0	0%

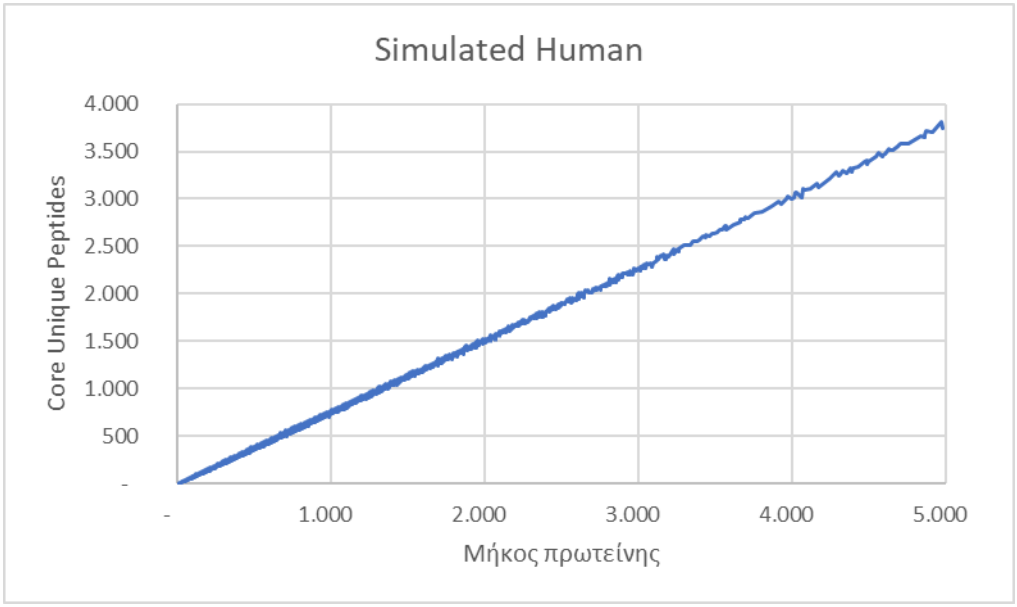
Πίνακας 12 Πλήθος CrUP ανάλογα το μήκος του πεπτιδίου (4-10) στο Simulated πρωτέωμα του ανθρώπου

Σε συνέχεια της διερεύνησης των χαρακτηριστικών για τα μοναδικά πεπτίδια ελαχίστου μήκους στο Simulated Uniquome του ανθρώπου αναλύθηκαν τα μοναδικά πεπτίδια ως προς την σχετική θέση εμφάνισης τους στις υποθετικές πρωτεΐνες. Τα αποτελέσματα έδειξαν πώς τα μοναδικά πεπτίδια ελαχίστου μήκους εμφανίζονται με την ίδια συχνότητα στις διάφορες θέσεις των πρωτεϊνών του προσομοιωμένου πρωτεώματος του ανθρώπου (Εικόνα 65).

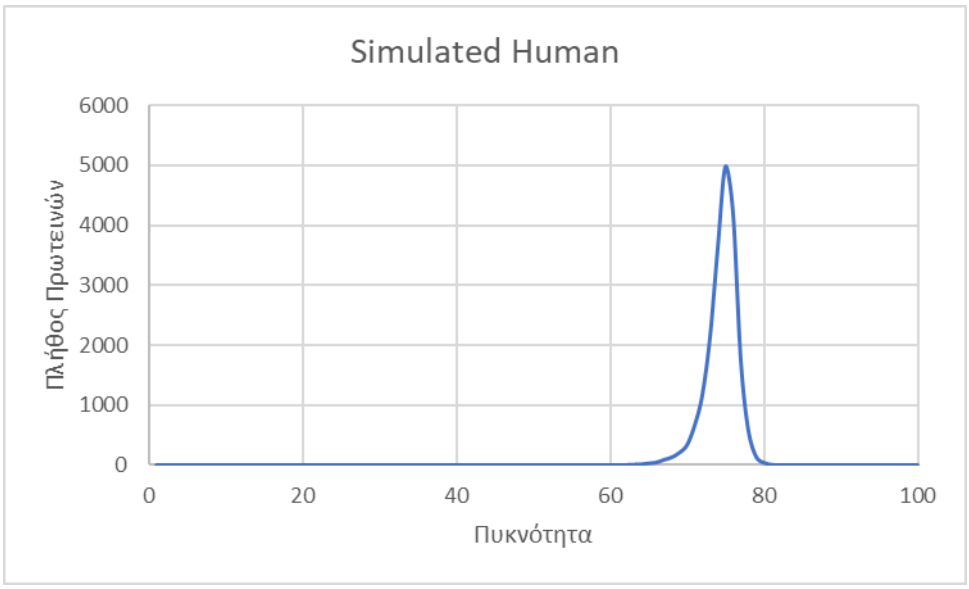


Εικόνα 65 Πλήθος CrUP ανά σχετική θέση εμφάνισης στο Simulated πρωτέωμα του ανθρώπου

Όπως και στο Uniquome του ανθρώπου έτσι και στο προσομοιωμένο Uniquome που κατασκευάστηκε, μελετήθηκε η πυκνότητα των υποθετικών πρωτεϊνών του από μοναδικά πεπτίδια ελαχίστου μήκους. Ακολουθώντας τα ίδια βήματα, αρχικά ταξινομήθηκαν οι πρωτεΐνες του ως προς το μέγεθός τους και τα μοναδικά πεπτίδια ελαχίστου μήκους που περιλαμβάνουν. Τα αποτελέσματα ανέδειξαν πως υπάρχει μια απόλυτη αναλογία στον αριθμό από μοναδικά πεπτίδια ελαχίστου μήκους που περιλαμβάνει η κάθε υποθετική πρωτεΐνη σε σχέση με το μήκος της, καθώς όσο αυξάνεται ο αριθμός από αμινοξέα που περιλαμβάνονται στις πρωτεΐνες τόσο αυξάνεται και ο αριθμός από μοναδικά πεπτίδια ελαχίστου μήκους που δημιουργούνται (Εικόνα 66). Στο επόμενο βήμα αυτής της ανάλυσης υπολογίστηκε η πυκνότητα της κάθε πρωτεΐνης από μοναδικά πεπτίδια. Η πυκνότητα για τις υποθετικές πρωτεΐνες του προσομοιωμένου πρωτεώματος του ανθρώπου κυμαινόταν αποκλειστικά στο διάστημα από 62 έως 82% με την συντριπτική πλειοψηφία να εμφανίζεται σε ποσοστά μεταξύ του 72 και του 77%. Τέλος, στο ποσοστό πυκνότητας της τάξης του 75% (ανά 100 αμινοξέα να αντιστοιχούν 75 μοναδικά πεπτίδια) αντιστοιχεί το μεγαλύτερο πλήθος υποθετικών πρωτεϊνών με 4.991 σε αριθμό (Εικόνα 67).



Εικόνα 66 Αριθμός CrUP στις υποθετικές πρωτεΐνες, ταξινομημένες ως προς το μέγεθος τους από αμινοξέα (για πρωτεΐνες με μέγεθος μέχρι 5.000 αμινοξέα) στο προσομοιωμένο πρωτέωμα του ανθρώπου.



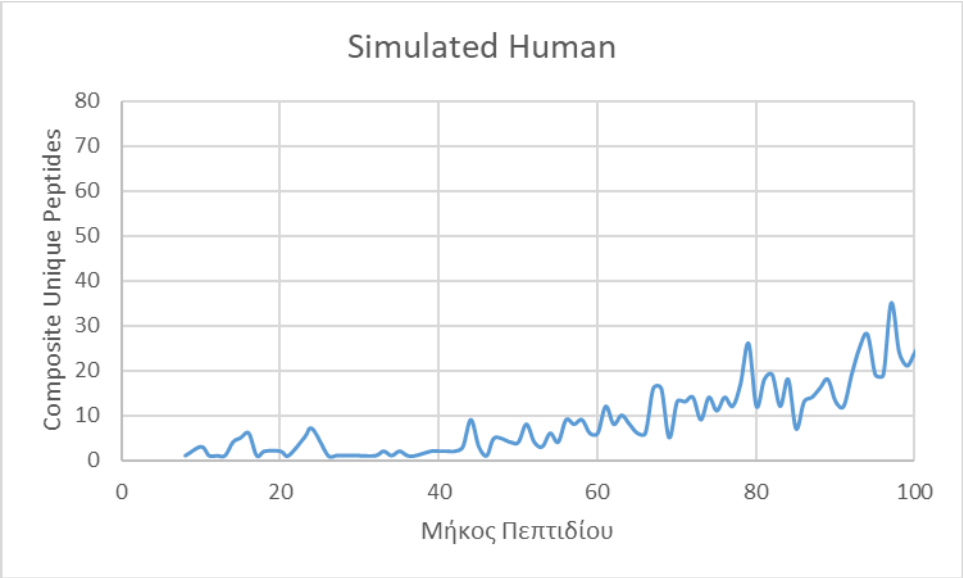
Εικόνα 67 Πλήθος πρωτεϊνών ως προς την πυκνότητά τους από CrUP στον άνθρωπο

Σύνθετα Μοναδικά πεπτίδια

Σε συνέχεια της καταγραφής των αποτελεσμάτων για τα χαρακτηριστικά των μοναδικών πεπτιδίων του προσομοιωμένου πρωτεώματος του ανθρώπου, αναλύθηκαν τα σύνθετα μοναδικά πεπτίδια.

Τα αποτελέσματα της ανάλυσης ως προς το μήκος των σύνθετων μοναδικών πεπτιδίων στο προσομοιωμένο πρωτέωμα του ανθρώπου έδειξαν πως το μήκος τους

κυμαίνεται από 8 έως 34.349 αμινοξέα. Αναλυτικότερα παρατηρείται πως τα σύνθετα μοναδικά πεπτίδια που αποτελούνται από αριθμό αμινοξέων <60, αντίθετα όσο ο αριθμός από αμινοξέα των σύνθετων μοναδικών πεπτιδίων αυξάνεται τόσο αυξάνεται και το πλήθος των σύνθετων αμινοξέων (Εικόνα 68). Ο μεγαλύτερος αριθμός από σύνθετα μοναδικά πεπτίδια εμφανίζεται σε πεπτίδια τα οποία έχουν μήκους μεγαλύτερο από 100 αμινοξέα. Αναλυτικότερα τη μεγαλύτερη ομάδα από σύνθετα μοναδικά πεπτίδια την συναντάμε σε πεπτίδια που έχουν μήκους 314 αμινοξέων (75 πεπτίδια) και ακολουθούν οι ομάδες των πεπτιδίων με μήκος 312 και 117 αμινοξέα αποτελούμενες από 73 και 62 πεπτίδια αντίστοιχα (Πίνακας 13).



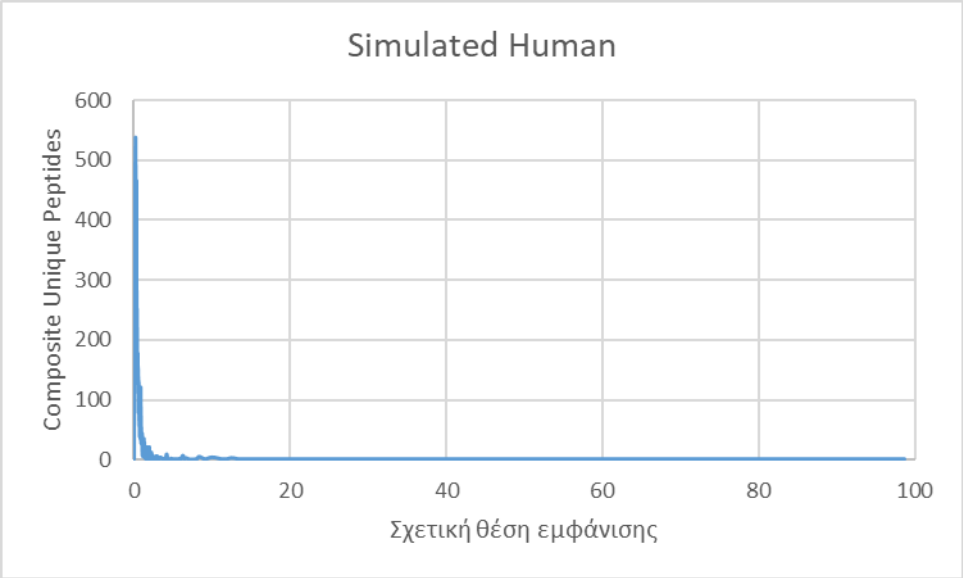
Εικόνα 68 Πλήθος CmUP ανάλογα το μήκος του πεπτιδίου (5-100) στο Simulated πρωτέωμα του ανθρώπου

Μήκος Πεπτιδίου	Πλήθος	%
314	75	0,37%
312	73	0,36%
117	62	0,30%
311	62	0,30%
114	61	0,30%
116	57	0,28%
115	56	0,27%
309	55	0,27%
313	54	0,26%
316	54	0,26%

Πίνακας 13 Πλήθος CmUP ανάλογα το μήκος του πεπτιδίου για τα 10 μήκη με το μεγαλύτερο πλήθος από CmUP στο Simulated πρωτέωμα του ανθρώπου

Η ανάλυση των χαρακτηριστικών για τα σύνθετα μοναδικά πεπτίδια ελαχίστου μήκους συνεχίστηκε με την διερεύνηση της σχετικής θέσης εμφάνισης τους μέσα στις υποθετικές πρωτεΐνες για το προσομοιωμένο πρωτέωμα του ανθρώπου. Τα αποτελέσματα

αυτής της ανάλυσης ανέδειξαν πως η πλειοψηφία των σύνθετων μοναδικών πεπτιδίων έχει σαν θέση έναρξης τα αμινοξέα που βρίσκονται στις αρχικές θέσεις των υποθετικών πρωτεϊνών του προσομοιωμένου πρωτεώματος (Εικόνα 69).

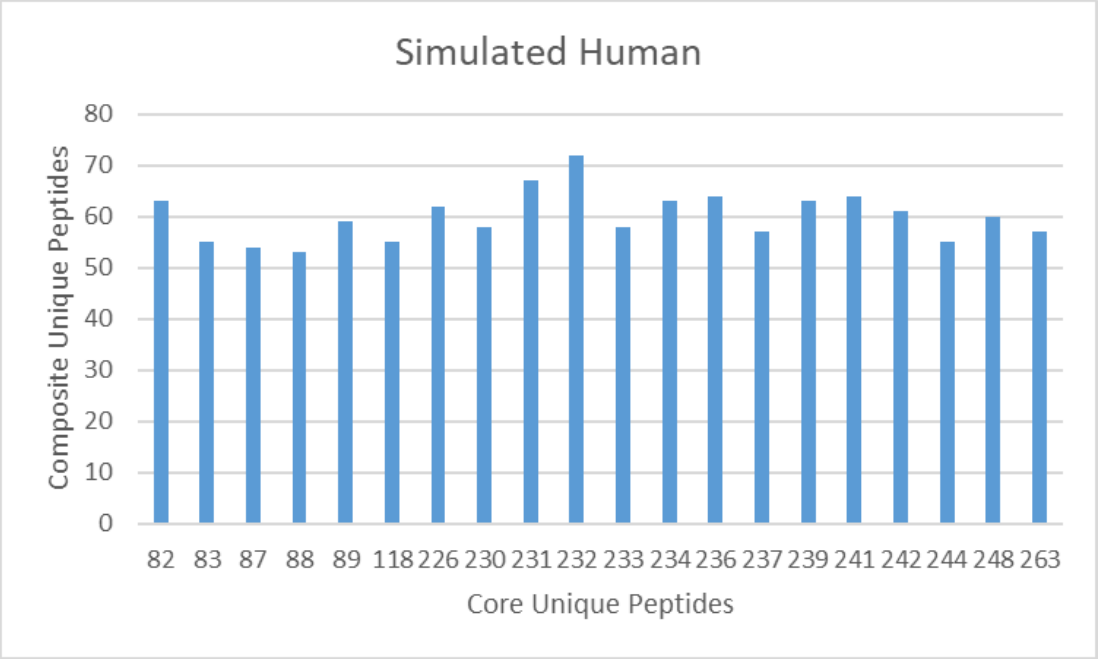


Εικόνα 69 Πλήθος CmUP ανά σχετική θέση εμφάνισης στο Simulated πρωτέωμα του ανθρώπου

Τέλος, μελετήθηκε το χαρακτηριστικό των σύνθετων μοναδικών πεπτιδίων στο προσομοιωμένο πρωτέωμα του ανθρώπου από το οποίο προκύπτει η πληροφορία για το πώς συνθέτονται από τα μοναδικά πεπτίδια ελαχίστου μήκους. Η συγκεκριμένη ανάλυση έδειξε πώς τα σύνθετα μοναδικά πεπτίδια που συνθέτονται από 2 μέχρι και 26.078 μοναδικά πεπτίδια ελαχίστου μήκους. Περεταίρω διερεύνηση της πληροφορίας για τον αριθμό από μοναδικά πεπτίδια ελαχίστου μήκους που συμμετέχουν στον σχηματισμό των σύνθετων μοναδικών πεπτιδίων ανέδειξε πως οι μεγαλύτερες ομάδες σύνθετων μοναδικών πεπτιδίων σχηματίζονται στα διαστήματα στα οποία τα πεπτίδια αυτά απαρτίζονται από 82 έως 89 και από 226 έως 248 μοναδικά πεπτίδια ελαχίστου μήκους. Ένα σημαντικό αποτέλεσμα που προέκυψε κατά την μελέτη της περιεκτικότητας από μοναδικά πεπτίδια ελαχίστου μήκους στα σύνθετα μοναδικά πεπτίδια είναι πώς στο σύνολο τους τα σύνθετα μοναδικά πεπτίδια όλων των ομάδων ανάλογα με τον αριθμό από μοναδικά πεπτίδια ελαχίστου μήκους που αποτελούνται δεν εμφανίζουν κάποια ομάδα από πεπτίδια που να ξεχωρίζει έντονα από τις υπόλοιπες. Αναλυτικότερα η μεγαλύτερη ομάδα από σύνθετα μοναδικά πεπτίδια είναι αυτή που τα πεπτίδια της δημιουργούνται από 232 μοναδικά πεπτίδια ελαχίστου μήκους και περιλαμβάνει μόλις 72 σύνθετα μοναδικά πεπτίδια. Στη συνέχεια ακολουθούν οι ομάδες των σύνθετων μοναδικών πεπτιδίων που τα πεπτίδια τους δημιουργούνται από 231 και 236 μοναδικά πεπτίδια ελαχίστου μήκους και αποτελούνται από 27 και 64 σύνθετα μοναδικά πεπτίδια αντίστοιχα (Εικόνα 70, πίνακα 14).

Αριθμός CrUP που συνθέτουν ένα CmUP	Πλήθος CmUP	%
232	72	0,35%
231	67	0,33%
236	64	0,31%
241	64	0,31%
82	63	0,31%
234	63	0,31%
239	63	0,31%
226	62	0,30%
242	61	0,30%
248	60	0,29%
89	59	0,29%
230	58	0,28%
233	58	0,28%
237	57	0,28%
263	57	0,28%
83	55	0,27%
118	55	0,27%
244	55	0,27%
87	54	0,26%
88	53	0,26%

Πίνακας 14 Πλήθος CmUP ανάλογα τον αριθμό από CrUP που αποτελούνται για το προσομοιωμένο πρωτέωμα του ανθρώπου για τις 20 πολυπληθέστερες ομάδες.



Εικόνα 70 Πλήθος CmUP ανάλογα τον αριθμό από CrUP που αποτελούνται για το προσομοιωμένο πρωτέωμα του ανθρώπου για τις 20 πολυπληθέστερες ομάδες.

Σύγκριση Μοναδικότητας σε ανθρώπινο πρωτέωμα και Προσομοιωμένο πρωτέωμα

Για την σύγκριση των αποτελεσμάτων που προέκυψαν στο Uniquome του προσομοιωμένου πρωτεώματος του ανθρώπου σε σχέση με το Uniquome του ανθρώπου, αναλύθηκαν τα μεταξύ τους χαρακτηριστικά. Τα αποτελέσματα έδειξαν πώς οι διαφορές στα δύο Uniquome είναι εμφανείς στα χαρακτηριστικά της μοναδικότητας των μοναδικών πεπτιδίων ελαχίστου μήκους καθώς και των σύνθετων μοναδικών πεπτιδίων (Πίνακας 15).

	Human	Simulated Human
Σύνολο πρωτεϊνών	20.430	20.430
Πρωτεΐνες με CrUP	20.282	20.426
Πρωτεΐνες χωρίς CrUP	148	4
Core Unique Peptides	7.263.888	8.542.941
Composite Unique Peptides	77.697	20.461
Πυκνότητα CrUP	64%	75%
Πυκνότητα CmUP	0,68%	0,18%
Συνολική κάλυψη	93%	100%

Πίνακας 15 Σύγκριση ανθρώπινου πρωτεώματος και προσομοιωμένου ως προς το Μοναδίωμα τους

Τα κυριότερα συμπεράσματα που προκύπτουν από την σύγκριση των δύο παραπάνω πρωτεωμάτων είναι ότι:

- Ο συνολικός αριθμός των CrUP του προσομοιωμένου πρωτεώματος σε σχέση με το ανθρώπινο πρωτέωμα είναι αυξημένος κατά ~17%.
- Ο συνολικός αριθμός των CmUP του προσομοιωμένου είναι περίπου το ¼ του συνολικού αριθμό των CmUP του ανθρώπινου πρωτεώματος.
- Στο προσομοιωμένο πρωτέωμα υπάρχουν μόλις 4 πρωτεΐνες που δεν εμφανίζουν κανένα CrUP ενώ στο ανθρώπινο 148.
- Η συνολική κάλυψη του προσομοιωμένου πρωτεώματος είναι 100% σε αντίθεση με του ανθρώπου που είναι 93%.
- Το προσομοιωμένο πρωτέωμα δεν εμφανίζει κανένα CrUP με μήκος μεγαλύτερο από 9 αμινοξέα ενώ του ανθρώπου εμφανίζει CrUP με μήκος έως 100 αμινοξέα (το μέγιστο εξ' ορισμού).
- Τα μοναδικά πεπτίδια ελαχίστου μήκους που συναντάμε με το μεγαλύτερο ποσοστό τόσο στο προσομοιωμένο όσο και στο ανθρώπινο πρωτέωμα είναι τα 6-πεπτίδια (75% και 69% αντίστοιχα).
- Τα σύνθετα μοναδικά πεπτίδια που αποτελούνται από 311 έως 314 αμινοξέα είναι αυτά που εμφανίζονται με το μεγαλύτερο ποσοστό για το προσομοιωμένο πρωτέωμα ενώ αντίθετα στο πρωτέωμα του ανθρώπου είναι τα πεπτίδια που αποτελούνται από 9 έως 13 αμινοξέα.

4.2 Διερεύνηση μοναδικότητας των νουκλεοτιδίων που κωδικοποιούν μοναδικά πεπτίδια

Η μελέτη του ανθρώπινου Uniquome επεκτάθηκε στην διερεύνηση της μοναδικότητας των νουκλεοτιδίων τα οποία κωδικοποιούν μοναδικά πεπτίδια και αντίστροφα. Συγκεκριμένα εξετάστηκε αν τα μοναδικά πεπτίδια κωδικοποιούνται από μοναδικά νουκλεοτίδια καθώς και αν από μοναδικά νουκλεοτίδια κωδικοποιούνται μοναδικά πεπτίδια. Η ανάλυση που πραγματοποιήθηκε για την απόδειξη της αμφίδρομης σχέσης μοναδικότητας νουκλεοτιδίων και πεπτιδίων περιορίστηκε σε συγκεκριμένες οικογένειες πρωτεϊνών λόγω της πολυπλοκότητας του γονιδιώματος. Μία από τις οικογένειες πρωτεϊνών που αναλύθηκε για την συγκεκριμένη έρευνα είναι η οικογένεια RAS που μελετήθηκε αναλυτικά ως προς την μοναδικότητα των πρωτεϊνών της στο υποκεφάλαιο 4.1.4.

Όπως παρουσιάστηκε παραπάνω η οικογένεια RAS (K-ras, N-ras, H-ras) αποτελείται από 127 μοναδικά πεπτίδια ελαχίστου μήκους (41, 45 και 41 αντίστοιχα). Αρχικά χρησιμοποιώντας δεδομένα της βάσης δεδομένων NCBI (National Center for Biotechnology Information) ανακτήθηκαν οι νουκλεοτιδικές αλληλουχίες από τις οποίες προέρχονται οι τρεις πρωτεΐνες (Εικόνα 71) και αντιστοιχήθηκαν τα μοναδικά πεπτίδια των τριών πρωτεϊνών με τα αντίστοιχα νουκλεοτίδια που τα κωδικοποιούν (Πίνακας 16,17,18).

CCDS Sequence Data

Blue highlighting indicates alternating exons.

Red highlighting indicates amino acids encoded across a splice junction.

Mouse over the nucleotide or protein sequence below and click on the highlighted codon or residue to select the pair.

Nucleotide Sequence (570 nt):

ATGACTGAATATAAACTTGTGGTAGTTGGAGCTGGTGGCGTAGGCAAGAGTGCCTTGACGATACAGCTAA
TTCAGAATCATTTTGTGGACGAATATGATCCAACAATAGAGGATTCCTACAGGAAGCAAGTAGTAATTGA
TGGAGAAACCTGTCTCTTGGATATTCTCGACACAGCAGGTCAAGAGGAGTACAGTGCAATGAGGGACCAG
TACATGAGGACTGGGGAGGGCTTTCTTTGTGTATTTGCCATAAATAATACTAAATCATTTGAAGATATTC
ACCATTATAGAGAACAAATTAAAAGAGTTAAGGACTCTGAAGATGTACCTATGGTCCTAGTAGGAAATAA
ATGTGATTTGCCTTCTAGAACAGTAGACACAAAACAGGCTCAGGACTTAGCAAGAAGTTATGGAATTCCT
TTTATTGAAACATCAGCAAAGACAAGACAGAGAGTGGAGGATGCTTTTTATACATTGGTGAGAGAGATCC
GACAATACAGATTGAAAAAAATCAGCAAAGAAGAAAAGACTCCTGGCTGTGTGAAAATTAAAAAATGCAT
TATAATGTAA

Translation (189 aa):

MTEYKLVVVGAGGVGKSALTIQLIQNHFVDEYDPTIEDSYRKQVVIDGETCLLDILDTAGQEEYSAMRDQ
YMRTGEGFLCVFAINNTKSFEDIHHYREQIKRVKDSEDVPMVLVGNKCDLPSRTVDTKQAQDLARSYGIP
FIETSAKTRQRVEDAFYTLVREIRQYRLKKISKEEKTPGCVKIKKCIIM

Εικόνα 71 Αντιστοίχιση μεταγραφόμενων νουκλεοτιδίων με την μεταφραζόμενη πρωτεΐνη για την K-ras χρησιμοποιώντας εργαλεία της NCBI

Τέλος, με τη χρήση του αλγόριθμου blastn και χρησιμοποιώντας εργαλεία του NCBI αναζητήθηκαν πού αλλού εντοπίζονται οι αλληλουχίες αυτές (γραμμική αλληλουχία) στο μεταγράφημα του ανθρώπου. Στην περίπτωση που οι αμινοξικές αλληλουχίες εντοπίζονταν και σε κάποια άλλη θέση στο ανθρώπινο μεταγράφημα θεωρήθηκαν ότι δεν είναι Unique στο μεταγράφημα του ανθρώπου (Πίνακας 16,17,18).

K-RAS			
Core Unique Peptides	**Nucleotides**	**Unique στο μεταγράφημα**	**Unique σε μεταφραζόμενη πρωτεΐνη**
EDIHH	GAAGATATTCACCAT	ΟΧΙ	ΝΑΙ
DIHHY	GATATTCACCATTAT	ΟΧΙ	ΝΑΙ
IHHYR	ATTCACCATTATAGA	ΝΑΙ	ΝΑΙ
HHYREQ	CACCATTATAGAGAACAA	ΝΑΙ	ΝΑΙ
HYREQI	CATTATAGAGAACAAATT	ΝΑΙ	ΝΑΙ
RVKDSE	AGAGTTAAGGACTCTGAA	ΝΑΙ	ΝΑΙ
VKDSEDV	GTTAAGGACTCTGAAGATGTA	ΝΑΙ	ΝΑΙ
DSEDVP	GACTCTGAAGATGTACCT	ΝΑΙ	ΝΑΙ
SEDVPM	TCTGAAGATGTACCTATG	ΝΑΙ	ΝΑΙ
EDVPMV	GAAGATGTACCTATGGTC	ΝΑΙ	ΝΑΙ
KCDLPS	AAATGTGATTTGCCTTCT	ΝΑΙ	ΝΑΙ
CDLPSR	TGTGATTTGCCTTCTAGA	ΝΑΙ	ΝΑΙ
LPSRTV	TTGCCTTCTAGAACAGTA	ΝΑΙ	ΝΑΙ
PSRTVD	CCTTCTAGAACAGTAGAC	ΝΑΙ	ΝΑΙ
SRTVDT	TCTAGAACAGTAGACACA	ΝΑΙ	ΝΑΙ
DTKQAQ	GACACAAAACAGGCTCAG	ΝΑΙ	ΝΑΙ
TKQAQD	ACAAAACAGGCTCAGGAC	ΝΑΙ	ΝΑΙ
KQAQDL	AAACAGGCTCAGGACTTA	ΝΑΙ	ΝΑΙ
RSYGIPF	AGAAGTTATGGAATTCCTTTT	ΝΑΙ	ΝΑΙ
AKTRQR	GCAAAGACAAGACAGAGA	ΝΑΙ	ΝΑΙ
KTRQRV	AAGACAAGACAGAGAGTG	ΝΑΙ	ΝΑΙ
TRQRVE	ACAAGACAGAGAGTGGAG	ΝΑΙ	ΝΑΙ
RQRVED	AGACAGAGAGTGGAGGAT	ΝΑΙ	ΝΑΙ
QRVEDA	CAGAGAGTGGAGGATGCT	ΝΑΙ	ΝΑΙ
RVEDAFY	AGAGTGGAGGATGCTTTTTAT	ΝΑΙ	ΝΑΙ
IRQYRL	ATCCGACAATACAGATTG	ΝΑΙ	ΝΑΙ
RQYRLK	CGACAATACAGATTGAAA	ΝΑΙ	ΝΑΙ
QYRLKK	CAATACAGATTGAAAAAA	ΝΑΙ	ΝΑΙ
RLKKIS	AGATTGAAAAAAATCAGC	ΝΑΙ	ΝΑΙ
KKISKE	AAAAAAATCAGCAAAGAA	ΟΧΙ	ΝΑΙ
ISKEEK	ATCAGCAAAGAAGAAAAG	ΟΧΙ	ΝΑΙ
SKEEKTP	AGCAAAGAAGAAAAGACTCCT	ΝΑΙ	ΝΑΙ
EEKTPG	GAAGAAAAGACTCCTGGC	ΝΑΙ	ΝΑΙ
EKTPGC	GAAAAGACTCCTGGCTGT	ΝΑΙ	ΝΑΙ
KTPGCV	AAGACTCCTGGCTGTGTG	ΝΑΙ	ΝΑΙ
TPGCVK	ACTCCTGGCTGTGTGAAA	ΝΑΙ	ΝΑΙ
PGCVKI	CCTGGCTGTGTGAAAATT	ΝΑΙ	ΝΑΙ
CVKIK	TGTGTGAAAATTAAA	ΟΧΙ	ΝΑΙ
KIKKC	AAAATTAAAAAATGC	ΟΧΙ	ΝΑΙ
IKKCII	ATTAAAAAATGCATTATA	ΝΑΙ	ΝΑΙ
KKCIIM	AAAAAATGCATTATAATG	ΝΑΙ	ΝΑΙ

Πίνακας 16 Αποτελέσματα για την μοναδικότητα των νουκλεοτιδίων στην πρωτεΐνη K-ras

N-RAS			
Core Unique Peptides	**Nucleotides**	**Unique στο μεταγράφημα**	**Unique σε μεταφραζόμενη πρωτεΐνη**
FAINNS	TTTGCCATCAATAATAGC	NAI	NAI
AINNSK	GCCATCAATAATAGCAAG	NAI	NAI
NNSKSF	AATAATAGCAAGTCATTT	NAI	NAI
NSKSFA	AATAGCAAGTCATTTGCG	NAI	NAI
SKSFAD	AGCAAGTCATTTGCGGAT	NAI	NAI
KSFADI	AAGTCATTTGCGGATATT	NAI	NAI
SFADIN	TCATTTGCGGATATTAAC	NAI	NAI
FADINL	TTTGCGGATATTAACCTC	NAI	NAI
ADINLY	GCGGATATTAACCTCTAC	NAI	NAI
DINLYR	GATATTAACCTCTACAGG	NAI	NAI
INLYRE	ATTAACCTCTACAGGGAG	NAI	NAI
NLYREQ	AACCTCTACAGGGAGCAG	NAI	NAI
LYREQI	CTCTACAGGGAGCAGATT	NAI	NAI
DDVPMVLVGNKCDLP	GATGATGTACCTATGGTGCTAGTG GGAAACAAGTGTGATTTGCCA	NAI	NAI
KCDLPT	AAGTGTGATTTGCCAACA	NAI	NAI
CDLPTR	TGTGATTTGCCAACAAGG	NAI	NAI
DLPTRT	GATTTGCCAACAAGGACA	NAI	NAI
LPTRTV	TTGCCAACAAGGACAGTT	NAI	NAI
PTRTVD	CCAACAAGGACAGTTGAT	NAI	NAI
TRTVDTK	ACAAGGACAGTTGATACAAAA	NAI	NAI
DTKQAH	GATACAAAACAAGCCCAC	NAI	NAI
KQAHEL	AAACAAGCCCACGAACTG	NAI	NAI
AHELAK	GCCCACGAACTGGCCAAG	NAI	NAI
HELAKS	CACGAACTGGCCAAGAGT	NAI	NAI
ELAKSY	GAACTGGCCAAGAGTTAC	NAI	NAI
LAKSYG	CTGGCCAAGAGTTACGGG	NAI	NAI
AKSYGI	GCCAAGAGTTACGGGATT	NAI	NAI
KSYGIPF	AAGAGTTACGGGATTCCATTC	NAI	NAI
FIETSAKTRQG	TTCATTGAAACCTCAGCC AAGACCAGACAGGGT	NAI	NAI
GVEDAFYTLVREIRQY	GGTGTTGAAGATGCTTTTTAC ACACTGGTAAGAGAAATACGC CAGTAC	NAI	NAI
IRQYRM	ATACGCCAGTACCGAATG	NAI	NAI
QYRMK	CGCCAGTACCGAATGAAA	NAI	NAI
YRMKKL	TACCGAATGAAAAAACTC	NAI	NAI
MKKLNS	ATGAAAAAACTCAACAGC	NAI	NAI
KLNSSD	AAACTCAACAGCAGTGAT	NAI	NAI
LNSSDD	CTCAACAGCAGTGATGAT	NAI	NAI
NSSDDG	AACAGCAGTGATGATGGG	NAI	NAI
SSDDGTQ	AGCAGTGATGATGGGACTCAG	NAI	NAI
DDGTQG	GATGATGGGACTCAGGGT	NAI	NAI
DGTQGC	GATGGGACTCAGGGTTGT	NAI	NAI
GTQGCM	GGGACTCAGGGTTGTATG	NAI	NAI
GCMGL	GGTTGTATGGGATTG	NAI	NAI
CMGLP	TGTATGGGATTGCCA	NAI	NAI
GLPCVV	GGATTGCCATGTGTGGTG	NAI	NAI
PCVVM	CCATGTGTGGTGATG	ΟΧΙ	NAI

Πίνακας 17 Αποτελέσματα για την μοναδικότητα των νουκλεοτιδίων στην πρωτεΐνη N-ras

H-RAS			
Core Unique Peptides	**Nucleotides**	**Unique στο μεταγράφημα**	**Unique σε μεταφραζόμενη πρωτεΐνη**
FEDIHQ	TTTGAGGACATCCACCAG	ΝΑΙ	ΝΑΙ
DIHQY	GACATCCACCAGTAC	ΌΧΙ	ΝΑΙ
IHQYR	ATCCACCAGTACAGG	ΌΧΙ	ΝΑΙ
HQYREQ	CACCAGTACAGGGAGCAG	ΝΑΙ	ΝΑΙ
QYREQI	CAGTACAGGGAGCAGATC	ΝΑΙ	ΝΑΙ
NKCDLA	AACAAGTGTGACCTGGCT	ΝΑΙ	ΝΑΙ
KCDLAA	AAGTGTGACCTGGCTGCA	ΝΑΙ	ΝΑΙ
CDLAAR	TGTGACCTGGCTGCACGC	ΝΑΙ	ΝΑΙ
DLAARTV	GACCTGGCTGCACGCACTGTG	ΝΑΙ	ΝΑΙ
LAARTVE	CTGGCTGCACGCACTGTGGAA	ΝΑΙ	ΝΑΙ
ARTVES	GCACGCACTGTGGAATCT	ΝΑΙ	ΝΑΙ
RTVESR	CGCACTGTGGAATCTCGG	ΝΑΙ	ΝΑΙ
TVESRQ	ACTGTGGAATCTCGGCAG	ΝΑΙ	ΝΑΙ
VESRQA	GTGGAATCTCGGCAGGCT	ΝΑΙ	ΝΑΙ
ESRQAQ	GAATCTCGGCAGGCTCAG	ΝΑΙ	ΝΑΙ
RQAQDL	CGGCAGGCTCAGGACCTC	ΝΑΙ	ΝΑΙ
RSYGIPY	CGAAGCTACGGCATCCCCTAC	ΝΑΙ	ΝΑΙ
YGIPYI	TACGGCATCCCCTACATC	ΝΑΙ	ΝΑΙ
GIPYIE	ATCCCCTACATCGAG	ΝΑΙ	ΝΑΙ
YIETSAKT	TACATCGAGACCTCGGCCAAGACC	ΝΑΙ	ΝΑΙ
REIRQH	CGTGAGATCCGGCAGCAC	ΝΑΙ	ΝΑΙ
EIRQHK	GAGATCCGGCAGCACAAG	ΝΑΙ	ΝΑΙ
IRQHKL	ATCCGGCAGCACAAGCTG	ΝΑΙ	ΝΑΙ
RQHKLR	CGGCAGCACAAGCTGCGG	ΝΑΙ	ΝΑΙ
QHKLRK	CAGCACAAGCTGCGGAAG	ΝΑΙ	ΝΑΙ
HKLRKL	CACAAGCTGCGGAAGCTG	ΝΑΙ	ΝΑΙ
KLRKLN	AAGCTGCGGAAGCTGAAC	ΝΑΙ	ΝΑΙ
LRKLNP	CTGCGGAAGCTGAACCCT	ΝΑΙ	ΝΑΙ
RKLNPP	CGGAAGCTGAACCCTCCT	ΝΑΙ	ΝΑΙ
KLNPPD	AAGCTGAACCCTCCTGAT	ΝΑΙ	ΝΑΙ
LNPPDE	CTGAACCCTCCTGATGAG	ΝΑΙ	ΝΑΙ
NPPDES	AACCCTCCTGATGAGAGT	ΝΑΙ	ΝΑΙ
PDESGP	CCTGATGAGAGTGGCCCC	ΝΑΙ	ΝΑΙ
DESGPG	GATGAGAGTGGCCCCGGC	ΝΑΙ	ΝΑΙ
GPGCM	GGCCCCGGCTGCATG	ΌΧΙ	ΝΑΙ
PGCMSC	CCCGGCTGCATGAGCTGC	ΝΑΙ	ΝΑΙ
GCMSCK	GGCTGCATGAGCTGCAAG	ΝΑΙ	ΝΑΙ
CMSCKC	TGCATGAGCTGCAAGTGT	ΝΑΙ	ΝΑΙ
MSCKCV	ATGAGCTGCAAGTGTGTG	ΝΑΙ	ΝΑΙ
SCKCVL	AGCTGCAAGTGTGTGCTC	ΝΑΙ	ΝΑΙ
CKCVLS	TGCAAGTGTGTGCTCTCC	ΝΑΙ	ΝΑΙ

Πίνακας 18 Αποτελέσματα για την μοναδικότητα των νουκλεοτιδίων στην πρωτεΐνη H-ras

Στο σύνολο των 127 μοναδικών πεπτιδίων για την οικογένεια πρωτεϊνών Ras (K-ras, N-ras και H-ras) υπάρχουν μόλις 10 ολιγονουκλεοτίδια τα οποία δεν είναι μοναδικά στο μεταγράφημα του ανθρώπου. Περεταίρω ανάλυση στα συγκεκριμένα 10 ολιγονουκλεοτίδια ανέδειξε πως παρόλο που δεν είναι μοναδικά σαν γραμμική αλληλουχία είναι μοναδικά ως προς το πεδίο ανάγνωσης (τριπλέτες) που θα ακολουθηθεί για να παραχθεί η αντίστοιχη μεταφραζόμενη πρωτεΐνη. Συνεπώς, στο σύνολο τους τα μοναδικά πεπτίδια ελαχίστου

μήκους των πρωτεϊνών της οικογένειας Ras προέρχονται από μοναδικά νουκλεοτίδια ακολουθώντας τον κανόνα του πεδίο ανάγνωσης σύμφωνα με το οποίο θα μεταφραστούν σε πεπτίδια (Πίνακας 16,17,18).

Στη συνέχεια της μελέτης διερευνήθηκε η αντίστροφη σχέση μοναδικότητας, δηλαδή αν μοναδικά νουκλεοτίδια κωδικοποιούν και μοναδικά πεπτίδια. Η ανάλυση, μέσω της χρήσης του αλγόριθμου blastn του NCBI, εντόπισε αλληλουχίες ολιγονουκλεοτιδίων που ενώ είναι μοναδικές στο ανθρώπινο μεταγράφημα δεν μεταφράζονται σε μοναδικά πεπτίδια. Χαρακτηριστικό παράδειγμα είναι το πεπτίδιο EEYSAM το οποίο χαρακτηριστικέ ως Family unique peptide στην οικογένεια RAS. Το συγκεκριμένο πεπτίδιο δεν είναι μοναδικό καθώς εντοπίζεται σε παραπάνω από μία πρωτεΐνη στο ανθρώπινο πρωτέωμα, παρόλα αυτά τα 3 διαφορετικά νουκλεοτίδια που το κωδικοποιούν είναι μοναδικά ως προς το μεταγράφημα του ανθρώπου (Πίνακας 19).

	K-Ras	N-RAS	H-RAS
Unique Nucleotides	**GAG-GAG-TAC AGT-GCA-ATG**	**GAA-GAG-TAC AGT-GCC-ATG**	**GAG-GAG-TAC AGC-GCC-ATG**
Peptide	**EEYSAM**		

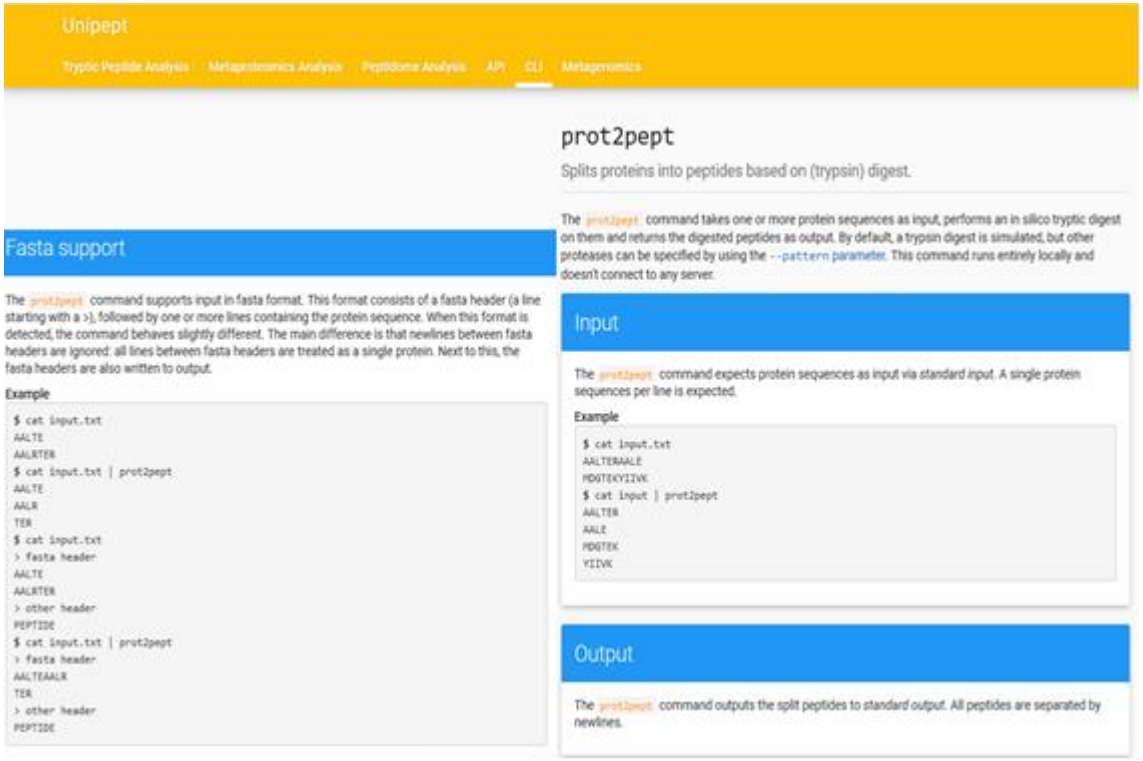
Πίνακας 19 Παράδειγμα για τον έλεγχο της μοναδικότητας πεπτιδίων που προέρχονται από μοναδικά νουκλεοτίδια.

4.3 Χρησιμότητα του ανθρώπινου Uniquome

4.3.1 Αναζήτηση μοναδικών πεπτιδίων CrUPs σε πεπτίδια παραγόμενα από την δράση της Θρυψίνης (tryptic digest peptides)

Η κύρια εφαρμογή της κατασκευής της βάσης δεδομένων του ανθρώπινου Uniquome είναι η δημιουργία μιας νέας προσέγγισης στην ταυτοποίηση των πρωτεϊνών. Όπως είναι γνωστό οι πιο εύχρηστες μέθοδοι για την ταυτοποίηση των πρωτεϊνών είναι αυτές που αξιοποιούν το πεπτιδικό αποτύπωμά τους (peptide finger-print) και αναλύουν την αμινοξική αλληλουχία των πεπτιδίων τους. Για να κοπούν οι πρωτεΐνες σε πεπτίδια και στη συνέχεια να γίνει η ταυτοποίηση χρησιμοποιούνται πρωτεολυτικά ένζυμα με το πλέον διαδεδομένο να είναι η θρυψίνη (trypsin).

Με σκοπό την εκμετάλλευση της βάσης δεδομένων των Uniquomes για την ταυτοποίηση πρωτεϊνών με την χρήση όσο των δυνατών λιγότερων πεπτιδίων γίνεται (ακόμα και με την χρήση ενός μόνο πεπτιδίου) χρησιμοποιήθηκαν εργαλεία και εφαρμογές από διάφορες βιολογικές βάσεις δεδομένων. Αρχικά, χρησιμοποιώντας τα εργαλεία της Βάσης δεδομένων Uniprot (Version 10/2019) απομονώθηκε το πρωτέωμα του ανθρώπου. Στη συνέχεια με την χρήση των αλγορίθμων της βάσης Unipept προσομοιώθηκε το πώς θα κόβονταν οι πρωτεΐνες που απομονώθηκαν σε μικρότερα πεπτίδια με θρυψίνη. Στην εικόνα 72, παρουσιάζεται το interface και τα εργαλεία της εφαρμογής Unipept τα οποία χρησιμοποιήθηκαν.



Εικόνα 72 Εντολές της εφαρμογής Unipept για το κόψιμο των πρωτεϊνών σε πεπτίδια με θρυψίνη.

Με την βοήθεια των εφαρμογών της Unipept παράχθηκε το αρχείο (Εικόνα 73) που περιέχει όλα τα πιθανά πεπτίδια που θα δημιουργηθούν εάν στο ανθρώπινο πρωτέωμα προστεθεί θρυψίνη.

```
>sp|Q9HBI6|CP4FB_HUMAN Cytochrome P450 4F11 OS=Homo sapiens OX=9606 GN=CYP4F11 PE=1 SV=3
MPQLSLSWLGLGPVAASPWLLLLLVGGSWLLAR
VLAWTYTFYDNCR
R
LQCFPQPPK
QNWFWGHQGLVTPTEEGMK
TLTQLVTTYPQGFK
LWLGPTFPLLILCHPDIIRPITSASAAVAPK
DMIFYGFLKPWLGDGLLLSGGDK
WSR
HR
R
MLTPAFHFNILKPYMK
IFNK
SVNIMHDK
WQR
LASEGSAR
LDMFEHISLMTLDSLQK
CVFSFESNCQEKPSEYIAAILELSAFVEK
R
NQQILLHTDFLYYLTPDGQR
FR
R
ACHLVHDFTDAVIQER
R
CTLPTQGIDDFLK
NK
AK
SK
TLDFIDVLLLSK
DEDGK
ELSDEDIR
AEADTFMFEGHDTTASGLSWVLYHLAK
HPEYQEQCR
QEVQELLK
DR
EPIEIEWDDLAQLPFLTMCIK
ESLR
LHPPVPVISR
CCTQDFVLPDGR
VIPK
GIVCLINIIGIHYNPTVWPDPEVYDPFR
FDQENIK
ER
SPLAFIPFSAGPR
NCIGQAFAMAEMK
VVLALTLLHFR
ILPTHTEPR
R
KPELILR
AEGGLWLR
VEPLGANSO
```

Εικόνα 73 Χρήση της θρυψίνης για το κόψιμο της πρωτεΐνης Q9HBI6 σε πεπτίδια.

Στην συνέχεια αναζητήθηκε αν τα πεπτίδια κομμένα με θρυψίνη (σε αυτά που είχαν μήκος > 3 αμινοξέα) περιείχαν έστω και ένα Core Unique Peptide. Από τον ορισμό των μοναδικών πεπτιδίων είναι γνωστό ότι αν ένα πεπτίδιο περιέχει στην αμινοξική του αλληλουχία έστω και ένα Core Unique Peptide τότε είναι μοναδικό πεπτίδιο. Με αυτή την διαδικασία ορίστηκαν τα μοναδικά πεπτίδια κομμένα με θρυψίνη (unique tryptic digest peptide).

Αφού καταγράφηκαν όλα τα μοναδικά πεπτίδια κομμένα με τρυψίνη σχετίστηκαν με τις πρωτεΐνες από τις οποίες προέρχονται και απομονώθηκαν οι πρωτεΐνες εκείνες που περιλαμβάνουν έστω ένα από αυτά τα πεπτίδια. Οι πρωτεΐνες αυτές μπορούν να ταυτοποιηθούν από ένα και μόνο πεπτίδιο (που έχει κοπεί με θρυψίνη). Τα αποτελέσματα έδειξαν ότι στο σύνολο των 20.430 πρωτεϊνών υπάρχουν 817.630 πεπτίδια κομμένα με θρυψίνη (που περιείχαν περισσότερα από 3 αμινοξέα) από τα οποία μοναδικά ήταν τα 563.613 (Πίνακας 20).

Species	Proteins (reviewed)	Tryptic digest generated Peptides	Tryptic Digest Unique Peptides
Human	**20.430**	**817.630**	**563.613**

Πίνακας 20 Μοναδικά πεπτίδια κομμένα με θρυψίνη

Τα αποτελέσματα για τις πρωτεΐνες οι οποίες μπορούν να ταυτοποιηθούν από ένα και μόνο πεπτίδιο κομμένο με θρυψίνη έδειξαν πώς από τις 20.282 πρωτεΐνες οι οποίες περιλαμβάνουν μοναδικά πεπτίδια οι 20.132 πρωτεΐνες μπορούν να ταυτοποιηθούν από ένα και μόνο πεπτίδιο, ενώ μόνο οι 150 δεν μπορούν να ταυτοποιηθούν από ένα πεπτίδιο (Πίνακας 21).

<table>
<tr><th>Proteins (reviewed)</th><th>Proteins with Unique Peptides</th><th>Proteins with Tryptic Digest Unique Peptides</th></tr>
<tr><td rowspan="3">20.430</td><td>20.282</td><td>20.132 (99%)</td></tr>
<tr><td>Proteins without Unique Peptides</td><td>Proteins without Tryptic Digest Unique Peptides</td></tr>
<tr><td>148</td><td>150</td></tr>
</table>

Πίνακας 21 Πρωτεΐνες που ταυτοποιούνται από ένα πεπτίδιο κομμένο με θρυψίνη

Με την χρήση της βάσης δεδομένων του ανθρώπινου Uniquome μπορεί να ταυτοποιηθεί το 99% των πρωτεϊνών του ανθρώπου χρησιμοποιώντας για την κάθε πρωτεΐνη ένα και μόνο πεπτίδιο ακολουθώντας την μέθοδο MS για την ταυτοποίηση αυτών των πρωτεϊνών.

4.3.2 Αναζήτηση μοναδικών πεπτιδίων του ανθρώπινου Uniquome σε βάσεις δεδομένων με πεπτίδια

Μια από τις πιο χρήσιμες εφαρμογές της βάσης δεδομένων των Uniquomes είναι η χρήση της για την αναζήτηση μοναδικών πεπτιδίων και σε άλλες βάσεις δεδομένων με αμινοξικές αλληλουχίες. Για τους σκοπούς αυτής της εφαρμογής αναζητήθηκαν και απομονώθηκαν πεπτίδια από άλλες βιολογικές βάσεις δεδομένων με πεπτίδια (όπως ανοσοπεπτίδια, αντιγονικά καρκινικά πεπτίδια κ.λπ.). Έπειτα ερευνήθηκε σε αυτά τα πεπτίδια αν περιλαμβάνουν μοναδικά πεπτίδια (από το Uniquome του ανθρώπου). Ακολουθούν τα παραδείγματα για τα :

- Ανοσοπεπτίδια – immune epitopes peptides
- Αντιγονικά καρκινικά πεπτίδια – Cancer Antigenic peptides

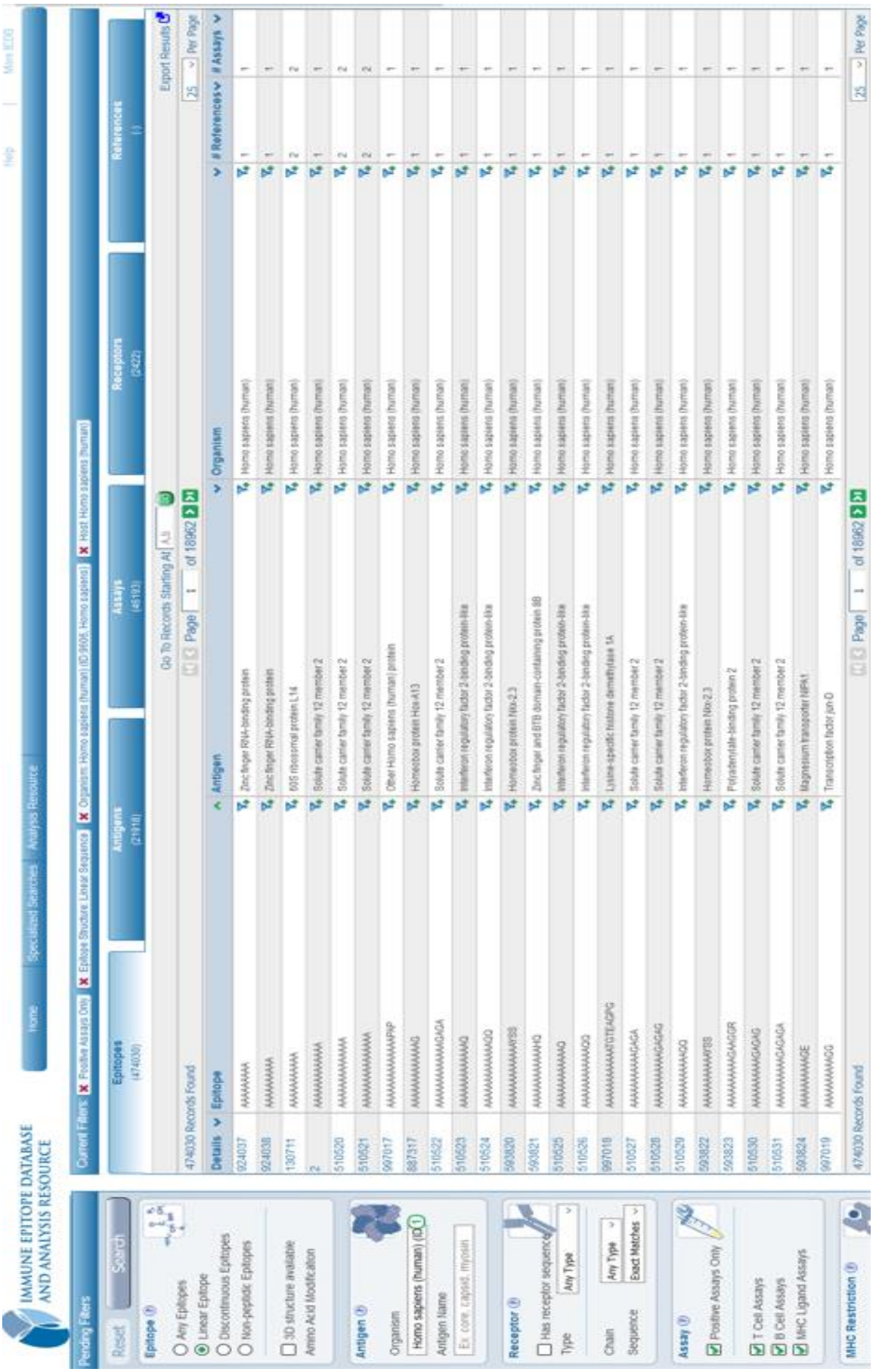
Ανοσοπεπτίδια – immune epitopes peptides

Χρησιμοποιώντας την Βάση δεδομένων Immune Epitope Database and analysis resource (https://www.iedb.org), απομονώθηκαν τα επιθυμητά ανοσοπεπτίδια. Πιο συγκεκριμένα για την καλύτερη κατανόηση των αποτελεσμάτων, απομονώθηκαν οι πεπτιδικοί επίτοποι με τα εξής χαρακτηριστικά (Εικόνα 74):

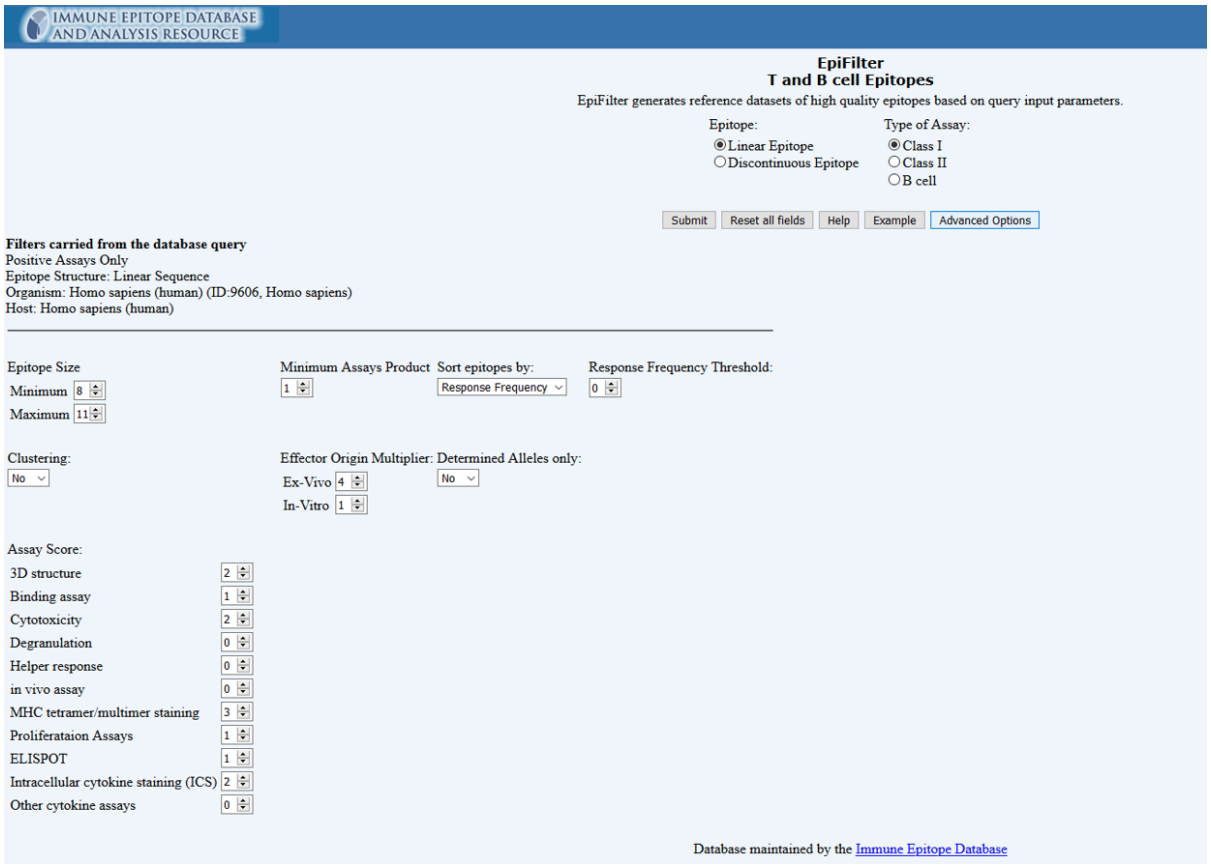
- Να είναι με τη μορφή γραμμικής αλληλουχίας αμινοξέων (Linear Sequence)
- Να προέρχονται από τον ανθρώπινο οργανισμό (*Homo sapiens*)
- Να έχουν δράση στον ανθρώπινο οργανισμό (ξενιστής*: Homo sapiens*)
- Η πειραματική μέθοδος που ακολουθήθηκε για την ταυτοποίηση τους ήταν : MHC class I, II και B Cell

Η αναζήτηση αυτή επέστρεψε έναν μεγάλο αριθμό από πεπτίδια πολλά μάλιστα από τα οποία ήταν και επαναλαμβανόμενα, για τον λόγο αυτό η αναζήτηση περιορίστηκε ακόμα περισσότερο, εξειδικεύοντας κι άλλο τα κριτήρια ανάλογα με την πειραματική μέθοδο που ακολουθήθηκε:

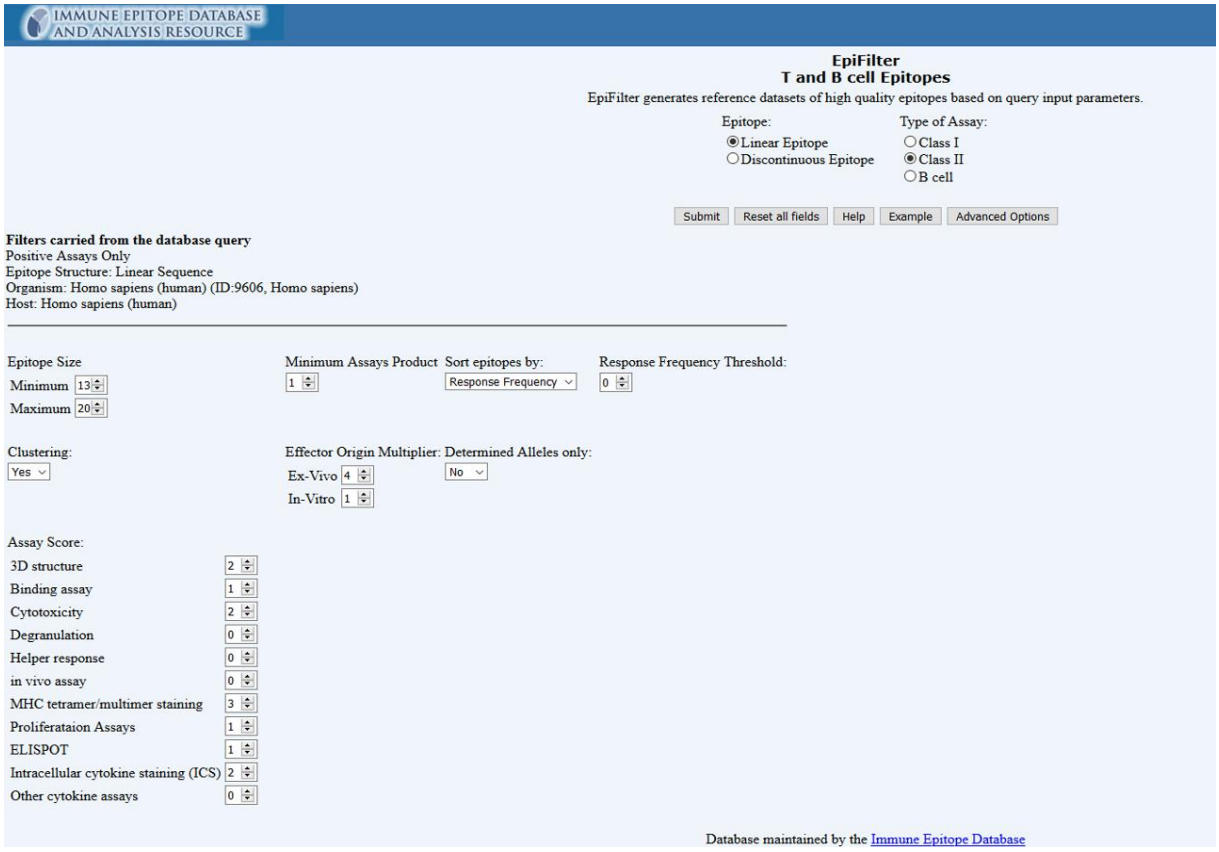
- Για την πειραματική μέθοδο MHC class I, επιλέχθηκαν τα πεπτίδια που έχουν μήκος από 8 έως 11 αμινοξέα, καθώς και τα επιθυμητά Assay score που φαίνονται στην εικόνα 75.
- Για την πειραματική μέθοδο MHC Class II, επιλέχθηκαν τα πεπτίδια που έχουν μήκος από 13 έως 20 αμινοξέα, καθώς και τα επιθυμητά Assay score που φαίνονται στην εικόνα 76.
- Για την πειραματική μέθοδο B Cell, επιλέχθηκαν τα πεπτίδια που έχουν μήκος από 5 έως 20 αμινοξέα, καθώς και τα επιθυμητά Assay score που φαίνονται στην εικόνα 77.



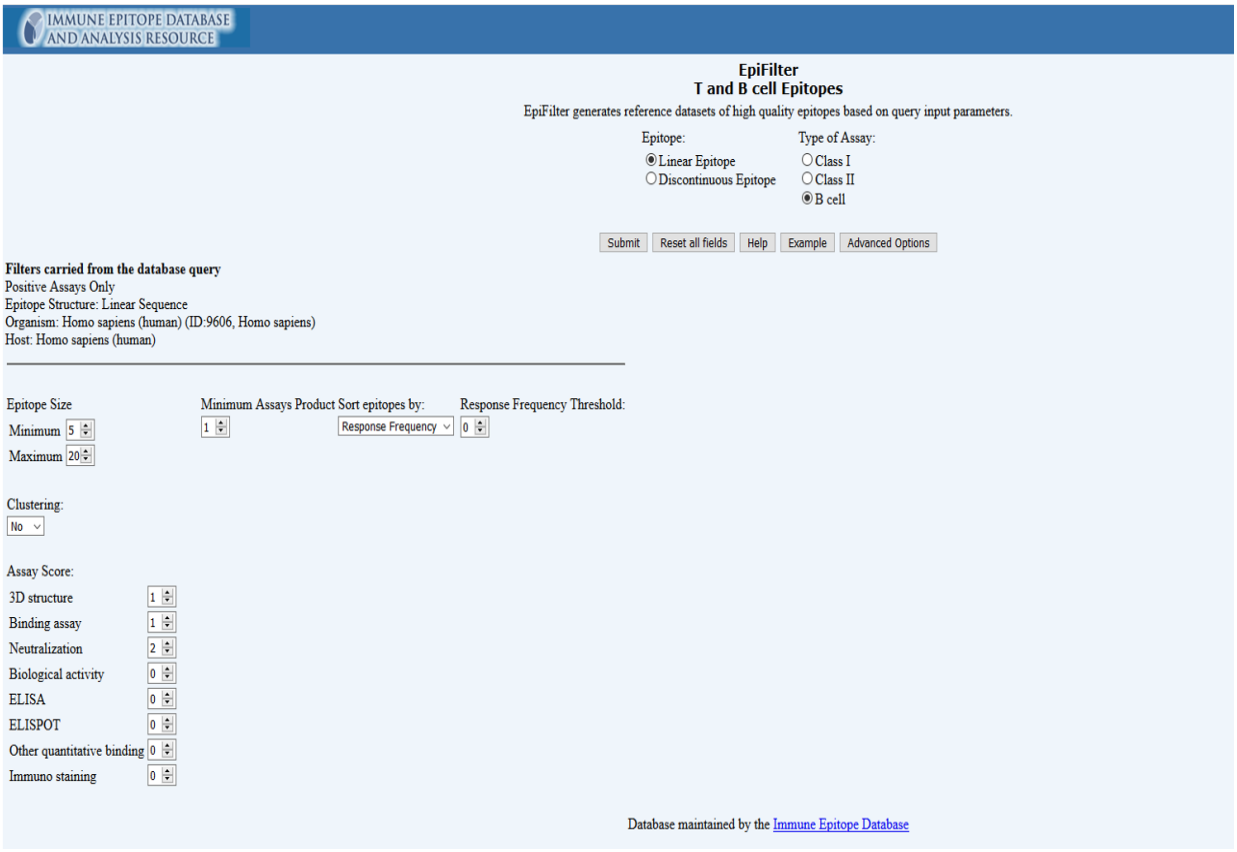
Εικόνα 74 Αναζήτηση στην βάση δεδομένων iedb πεπτιδικούς επίτοπους με τα επιλεγμένα κριτήρια αναζήτησης.



Εικόνα 75 Αναζήτηση επιτοπικών πεπτιδίων χρησιμοποιώντας την μέθοδο Class I και τα επιλεγμένα φίλτρα



Εικόνα 76 Αναζήτηση επιτοπικών πεπτιδίων χρησιμοποιώντας την μέθοδο Class II και τα επιλεγμένα φίλτρα



Εικόνα 77 Αναζήτηση επιτοπικών πεπτιδίων χρησιμοποιώντας την μέθοδο B-Cell και τα επιλεγμένα φίλτρα

Χρησιμοποιώντας τη βάση δεδομένων Immune Epitope Database and analysis resource (https://www.iedb.org) με τα παραπάνω κριτήρια απομονώθηκαν 3.709 ανοσοπεπτίδια χωρισμένα ανάλογα στην κατηγορία που ανήκουν:

- MHC class I: 963
- MHC class II: 428
- B-Cell: 2.318

Τα αποτελέσματα της ανάλυσης για τον έλεγχο αν τα επιτοπικά ανοσοπεπτίδια περιλαμβάνουν μοναδικά πεπτίδια (από το Uniquome του ανθρώπου) έδειξαν πως στο σύνολο των 3.709 ανοσοπεπτιδίων που απομονώθηκαν από την βάση δεδομένων Immune Epitope Database and analysis resource (https://www.iedb.org) τα 3.230 (87%) περιείχαν στην αμινοξική τους αλληλουχία τουλάχιστον ένα μοναδικό πεπτίδιο. Η ομάδα MHC Class II, είναι η ομάδα των επιτοπικών ανοσοπεπτιδίων της οποίας τα πεπτίδια που περιλαμβάνουν έστω ένα μοναδικό πεπτίδιο εμφανίζονται με το μεγαλύτερο ποσοστό της τάξης του 90% που αντιστοιχεί σε 387 πεπτίδια (Πίνακας 22).

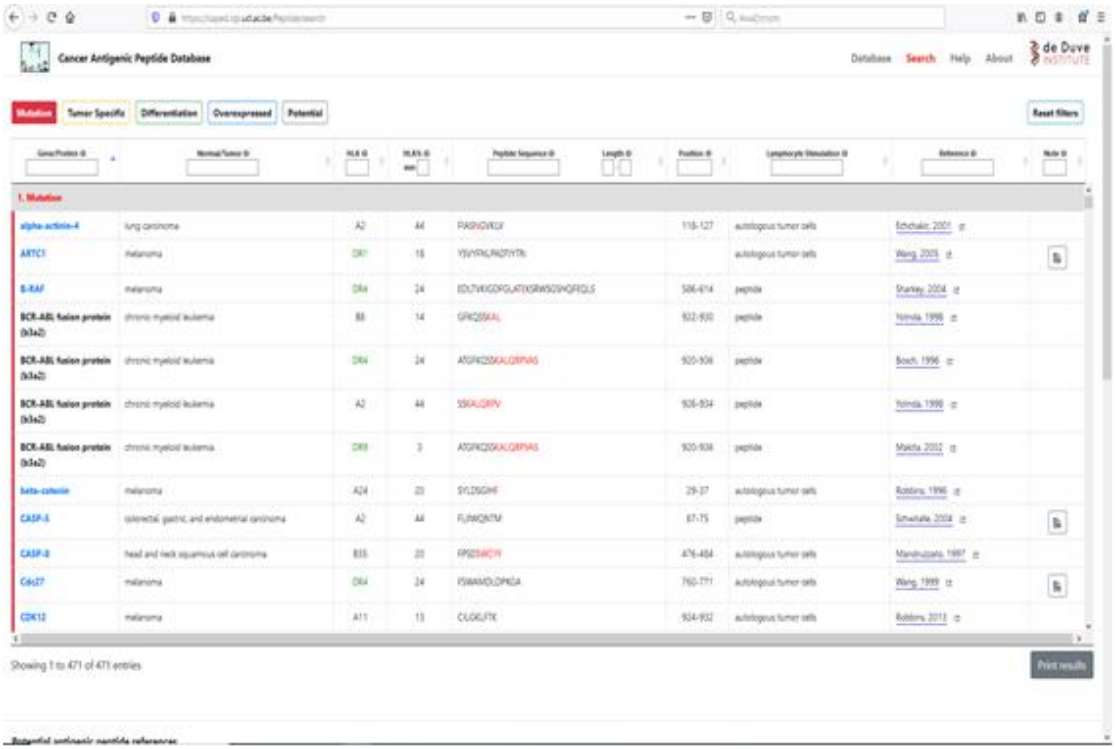
Type of Experiment	Immune Epitopes Peptides	Immune Epitopes Peptides contain Core Unique Peptides	%
MHC Class I	963	855	89%
MHC Class II	428	387	90%
B-Cell	2.318	1.988	86%
Total	**3.709**	**3.230**	**87%**

Πίνακας 22 Αποτελέσματα αναζήτησης μοναδικών πεπτιδίων του ανθρώπου σε Επιτοπικά Ανοσοπεπτίδια

Αντιγονικά καρκινικά πεπτίδια – Cancer Antigenic peptides

Χρησιμοποιώντας την βάση δεδομένων Cancer Antigenic Peptide Database (https://caped.icp.ucl.ac.be/), απομονώθηκαν 400 καρκινικά αντιγονικά πεπτίδια (Εικόνα 78) κατηγοριοποιημένα ως εξής:

- Mutation: 58
- Tumor-Specific: 130
- Differentiation: 85
- Overexpressed: 127



Εικόνα 78 Αναζήτηση στη βάση δεδομένων CAPB για καρκινικά αντιγονικά Πεπτίδια

Τα αποτελέσματα της ανάλυσης για το αν τα αντιγονικά καρκινικά πεπτίδια που απομονώθηκαν από την βάση δεδομένων Cancer Antigenic Peptide Database (https://caped.icp.ucl.ac.be/) είναι μοναδικά, έδειξαν πως στο σύνολο των 400 αντικαρκινικών πεπτιδίων τα 355 (ποσοστό 89%) περιλαμβάνουν έστω ένα μοναδικό πεπτίδιο (από το Uniquome του ανθρώπου). Σύμφωνα με την συγκεκριμένη ανάλυση η κατηγορία πεπτιδίων της οποίας τα πεπτίδια εμφανίζονται με τα μεγαλύτερα ποσοστά ως προς την περιεκτικότητα τους από μοναδικά πεπτίδια είναι η κατηγορία Differentiation της οποία από τα 85 αντικαρκινικά πεπτίδια που περιλαμβάνει τα 84 (ποσοστό 99%) περιλαμβάνουν έστω και ένα μοναδικό πεπτίδιο (Πίνακας 23).

Category	Cancer Antigenic Peptides	Cancer Antigenic Peptides contain Core Unique Peptides	%
Mutation	58	52	90%
Tumor-Specific	130	106	82%
Differentiation	85	84	99%
Overexpressed	127	113	89%
Total	**400**	**355**	**89%**

Πίνακας 23 Αποτελέσματα αναζήτησης μοναδικών πεπτιδίων του ανθρώπου σε Καρκινικά Αντιγονικά Πεπτίδια

4.3.3 Πρόβλεψη της Ανοσολογικής Απάντησης, της Ανοσολογικής Διαφυγής και της Παθογένειας του Ιού SARS-CoV-2, μέσω των Μοναδικών Πεπτιδικών Υπογραφών του ως προς το ανθρώπινο Πρωτέωμα

Η πανδημία του SARS-CoV-2 έχει απαιτήσει τον εντοπισμό περιοχών πεπτιδικής αλληλουχίας στο ιικό πρωτέωμα που είναι σε θέση να χρησιμεύουν ως αντιγονικές θέσεις - επίτοποι και στόχοι θεραπείας.

Στο υποκεφάλαιο αυτό θα παρουσιάσουμε την εφαρμογή των CrUPs για την δημιουργία μοναδικών πεπτιδίων ενός οργανισμού ως το πρωτέωμα ενός άλλου οργανισμού με σκοπό τη διαλεύκανση του μηχανισμού των αλληλεπιδράσεων ιού-ξενιστή. Η μέθοδος αυτή εφαρμόστηκε χαρτογραφώντας όχι τα CrUPs του ιού αυτού καθ' εαυτού, αλλά τα CrUPs του ιού που είναι μοναδικά έναντι ολόκληρου του πρωτεώματος του ξενιστή, δηλαδή εκείνα τα CrUPs του ιού τα οποία δεν υπάρχουν στο πρωτέωμα του ξενιστή. Με δεδομένο ότι ο ξενιστής του ιού SARS-CoV-2 είναι ο *Homo Sapiens* (Human), αναλύσαμε το SARS-CoV-2 πρωτέωμα για την ταυτοποίηση των CrUPs αυτού έναντι όλου του ανθρώπινου πρωτεώματος, με τη νέα αυτή κατηγορία πεπτιδίων να ονομάζεται πλέον C/H-CrUPs (Covid vs Human-Core Unique Peptides).

Η πρωτεωμική δομή και οργάνωση του ιού SARS-CoV-2 περιλαμβάνει 16 πρωτεΐνες, και έτσι για την ταυτοποίηση των C/H-CrUPs του ιού έναντι του ανθρώπινου πρωτεώματος κατασκευάσθηκαν *in silico* 16 υβριδικά πρωτεώματα, τα οποία περιείχαν τις 20.430 ταυτοποιημένες ανθρώπινες πρωτεΐνες συν 1 εκ των πρωτεϊνών του ιού (δηλαδή συνολικά 20.431 πρωτεΐνες), κάθε φορά. Στην συνέχεια, η εφαρμογή ανέλυσε βιοπληροφορικά τα 16 αυτά υβριδικά πρωτεώματα, με σκοπό την ταυτοποίηση των μοναδικών πεπτιδίων της κάθε πρωτεΐνης του ιού έναντι όλων των ανθρώπινων πρωτεϊνών.

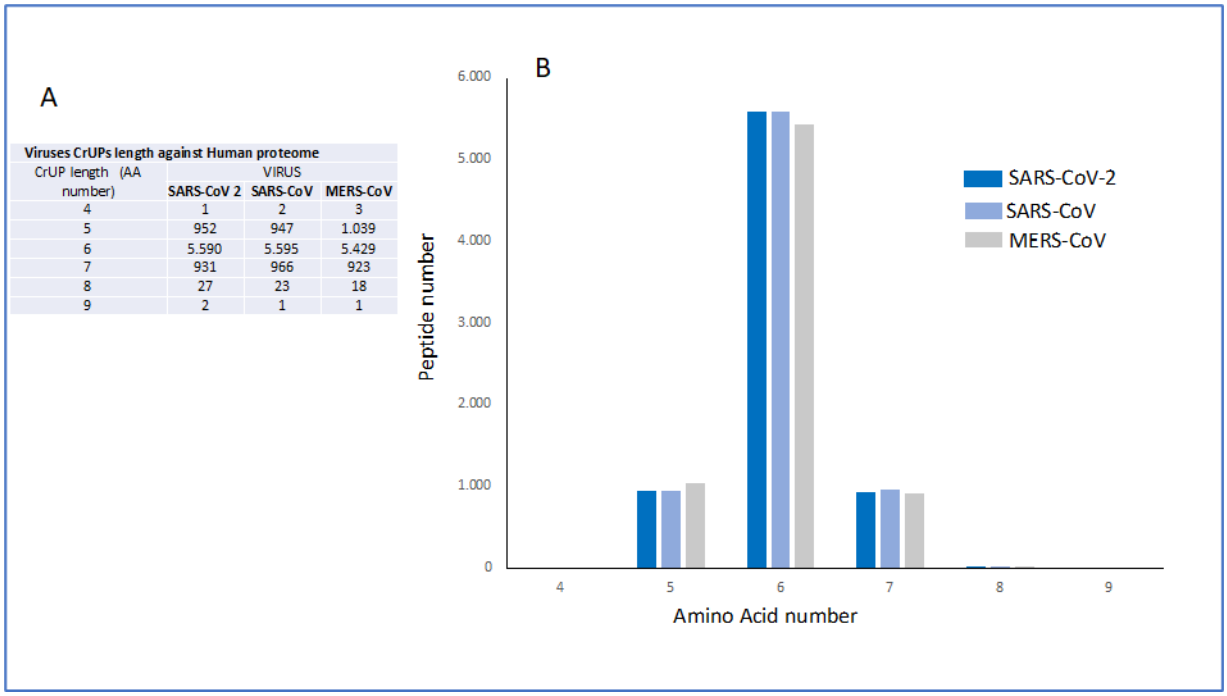
Διαπιστώθηκε, με έκπληξη, ότι ο SARS-CoV-2 περιλαμβάνει 7.503 C/H-CrUPs, με την πρωτεΐνη ακίδα SPIKE_SARS2 (P0DTC2) να ανιχνεύεται ως η πρωτεΐνη με την υψηλότερη πυκνότητα C/H-CrUP (Πίνακας 24).

Entry ID	Entry name	Protein name	Length (AA number)	C/H-CrUPs (number)	Density C/H-CrUPs
P0DTD1	R1AB_SARS2	Replicase polyprotein 1ab	7096	5334	75%
P0DTC1	R1A_SARS2	Replicase polyprotein 1a	4405	3294	75%
P0DTC2	SPIKE_SARS2	Spike glycoprotein	1273	987	**78%**
P0DTC9	NCAP_SARS2	Nucleoprotein	419	308	74%
P0DTC3	AP3A_SARS2	ORF3a protein	275	210	76%
P0DTC5	VME1_SARS2	Membrane protein	222	171	77%
P0DTC7	NS7A_SARS2	ORF7a protein	121	90	74%
P0DTC8	NS8_SARS2	ORF8 protein	121	82	68%
P0DTD2	ORF9B_SARS2	ORF9b protein	97	69	71%
P0DTD3	ORF9C_SARS2	Putative ORF9c protein	73	50	68%
P0DTC4	VEMP_SARS2	Envelope small membrane protein	75	48	64%
P0DTC6	NS6_SARS2	ORF6 protein	61	44	72%
P0DTG0	ORF3D_SARS2	Putative ORF3d protein	57	40	70%
P0DTD8	NS7B_SARS2	ORF7b protein	43	29	67%
P0DTG1	ORF3C_SARS2	ORF3c protein	41	23	56%
P0DTF1	ORF3B_SARS2	Putative ORF3b protein	22	15	68%

Πίνακας 24 Χαρτογράφηση των C/H-CrUPs του ιού SARS-CoV-2

Επεξεργασία και ανάλυση των C/H-CrUPs κατέδειξε το φάσμα μήκους αυτών από 4 έως 9 αμινοξέα, με μεγαλύτερα πεπτίδια να αδυνατούν να ανιχνευθούν στο -έναντι του ανθρώπου- (μοναδιαίο) πρωτέωμα του SARS-CoV-2. Συγκριτική μελέτη του C/H-CrUP χάρτη των μελών της ομάδας των β-κορονοϊών, η οποία περιλαμβάνει τους ιούς SARS-CoV-2, SARS-CoV και MERS-CoV, επιβεβαίωσε τις ισχυρές ομοιότητες μεταξύ αυτών στο

επίπεδο της δομής και αρχιτεκτονικής κατασκευής των CrUPs έναντι του ανθρώπινου πρωτεώματος (Εικόνα 79), με την πρωτεΐνη ακίδα (Spike) να εμφανίζει τη μεγαλύτερη C/H-CrUP πυκνότητα και στους τρείς τύπους ιών.



Viruses CrUPs length against Human proteome

CrUP length (AA number)	VIRUS		
	SARS-CoV 2	SARS-CoV	MERS-CoV
4	1	2	3
5	952	947	1.039
6	5.590	5.595	5.429
7	931	966	923
8	27	23	18
9	2	1	1

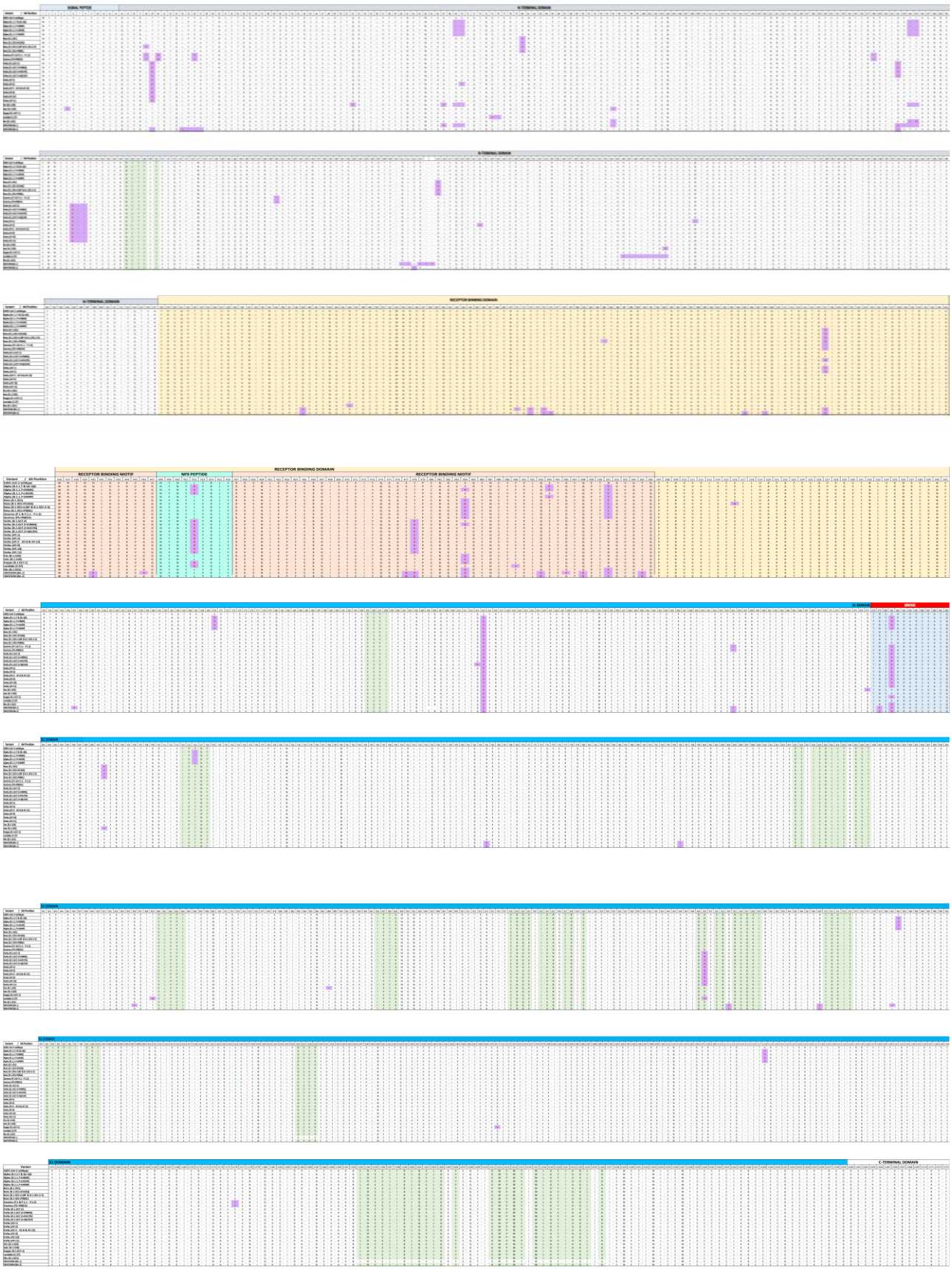
Εικόνα 79 Κατανομή μήκους αμινοξέων των C/H-CrUPs στους ιούς της ομάδας των β-κορώνα-ιών (SARS-CoV-2, SARS-CoV και MERS-CoV).

Στην παραπάνω εικόνα παρουσιάζονται:

(**A**) Ταυτοποίηση, καταγραφή και ομαδοποίηση των πεπτιδίων ανάλογα με το μήκος των αμινοξέων τους.

(**B**) Γραφική παρουσίαση του φάσματος μήκους αμινοξέων των C/H-CrUPs στην ομάδα των β-κορώνα-ιών.

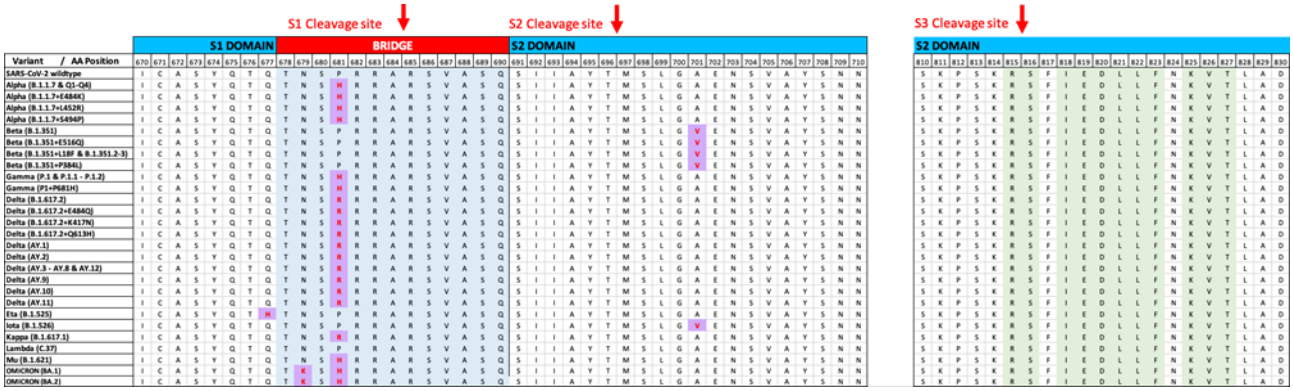
Λαμβάνοντας υπόψιν τα παραπάνω ευρήματα, η μελέτη επικεντρώθηκε περαιτέρω στην ανάλυση των C/H-CrUPs επί της Spike πρωτεΐνης (ακίδα) του ιού SARS-CoV-2 (SPIKE_SARS2, P0DTC2). Μέχρι στιγμής, έχουν χαρτογραφηθεί και χαρακτηριστεί διεξοδικά 10 κύριες παραλλαγές με προσαρμοστικές μεταλλάξεις και ταυτόχρονη μεγάλη εξάπλωση στους ανθρώπινους πληθυσμούς, που ονομάζονται από Άλφα (Alpha) έως Όμικρον (Omicron), αντίστοιχα. Για τη μηχανιστική διερεύνηση της συσχέτισης του προφίλ των μεταλλάξεων και της χωροθέτησης των C/H-CrUPs στην Spike πρωτεΐνη (ακίδα) του SARS-CoV-2, οι υπό-παραλλαγές μαζί με τη φυσικού-τύπου Spike πρωτεΐνη (ακίδα) ευθυγραμμίστηκαν κατάλληλα, στοιχήθηκαν και συγκρίθηκαν ενδελεχώς (Εικόνα 80).



Εικόνα 80 Ευθυγράμμιση της Spike πρωτεΐνης (ακίδα) του ιού SARS-CoV-2 των 26 κύριων υπο-παραλλαγών, μαζί με την Spike αλληλουχία φυσικού-τύπου.

Στην παραπάνω εικόνα, τα μωβ τετράγωνα επισημαίνουν τις σημειακές θέσεις των μεταλλάξεων στις υπό-παραλλαγές, ενώ το πράσινο χρώμα υποδηλώνει τα καθολικά συντηρημένα (ταυτόσημα) πεπτίδια των πρωτεϊνών “Spike”. Το κίτρινο χρώμα επισημαίνει τον τομέα δέσμευσης της Spike πρωτεΐνης (ακίδα) επί του ειδικού της υποδοχέα ACE2 στα κύτταρα-στόχους, ενώ το ροζ χρώμα υποδεικνύει το μοτίβο δέσμευσης στον υποδοχέα. Το κυανό χρώμα χαρακτηρίζει το πεπτίδιο NF9, ενώ το ανοιχτό-μπλε χρώμα επισημαίνει τη γέφυρα μεταξύ των τομέων S1 και S2. Τα κόκκινα βέλη υποδεικνύουν τα σημεία υδρολυτικής διάσπασης - σχάσης επί της Spike πρωτεΐνης (ακίδα). Με διαφορετικά χρώματα στην επάνω πλευρά της ευθυγράμμισης σημειώνονται οι διαφορετικοί τομείς της Spike πρωτεΐνης (ακίδα).

Στην εν λόγω πολλαπλή ευθυγράμμιση - στοίχιση αμινοξικής αλληλουχίας τοποθετήθηκαν όλες οι μεταλλάξεις που έχουν ανακοινωθεί ανά υπό-παραλλαγή και διαπιστώθηκε ότι η πλειονότητά τους συγκεντρώνεται στον τομέα S1 της Spike πρωτεΐνης (ακίδα), με δύο κρίσιμες μεταλλάξεις να ανιχνεύονται στην περιοχή γέφυρας S1-S2, και ειδικότερα στο αμινοξικό κατάλοιπο 681, που βρίσκεται κοντά στην πρώτη θέση διάσπασης της πρωτεΐνης από την πρωτεάση Φουρίνη (Furin), μεταξύ του 685ου και του 686ου αμινοξέος (Εικόνα 81).



Εικόνα 81 Ευθυγράμμιση (πολλαπλή στοίχιση) αμινοξικής αλληλουχίας της Spike πρωτεΐνης (ακίδα) του ιού SARS-CoV-2 γύρω από την περιοχή της γέφυρας (κόκκινο χρώμα), μεταξύ των τμημάτων - τομέων S1 και S2.

Στην παραπάνω εικόνα, το κόκκινο βέλος υποδεικνύει τη θέση διάσπασης (R685↓S) από την πρωτεάση Φουρίνη. Το μωβ χρώμα επισημαίνει τις σημειακές μεταλλάξεις γύρω από αυτήν τη θέση, ενώ το κόκκινο περίγραμμα υποδηλώνει τις παραλλαγές Delta και Kappa που φέρουν τη μετάλλαξη P681R.

Είναι αξιοσημείωτο, ότι όλες οι μεταλλάξεις που εξετάζονται εδώ αποδεικνύεται να δημιουργούν νέα CrUPs, έναντι του ανθρώπινου πρωτεώματος, σε σύγκριση με τη φυσικού-τύπου Spike πρωτεΐνη (ακίδα), υποδεικνύοντας, ως εκ τούτου, ότι τα στελέχη του

μεταλλαγμένου ιού εμφανίζουν νέες μοριακές ιδιότητες, και απαιτούν νέα διαγνωστική και θεραπευτική αντιμετώπιση. Αυτό αποτελεί ένα ιδιαίτερα σημαντικό εύρημα, δεδομένου ότι αυτά τα νέα C/H-CrUPs δεν ανιχνεύονται στο ανθρώπινο πρωτέωμα, αλλά παρατηρούνται αποκλειστικά και μόνο στα πρωτεώματα των παραλλαγών (και υπό-παραλλαγών) του ιού, δικαιολογώντας έτσι τη μεγάλη προσοχή που έχουν λάβει πρόσφατα οι παραλλαγές Alpha, Delta, Kappa, Lambda και Omicron του SARS-CoV-2 ιού, σε παγκόσμιο επίπεδο. Ο Πίνακας 25 παραθέτει όλα τα νέο-σχηματιζόμενα C/H-CrUPs από τις (μέχρι στιγμής) αναφερθείσες μεταλλάξεις στις παραλλαγές Alpha, Delta, Kappa και Lambda του SARS-CoV-2 κορώνα-ιού. Αυτές οι παραλλαγές και οι υπό-παραλλαγές τους περιλαμβάνουν 25 μεταλλάξεις, οι οποίες παράγουν 44 νέα CrUPs, έναντι του ανθρώπινου πρωτεώματος. Αυτά τα νέα C/H-CrUPs είναι δυνατό να οδηγήσουν στον σχηματισμό καινούργιων εγγενώς διαταραγμένων περιοχών (Intrinsically Disordered Regions: IDRs), καθώς και νέων μικρών γραμμικών μοτίβων (Short Linear Motifs: SLiMs) στις μεταλλαγμένες Spike πρωτεΐνες του SARS-CoV-2.

Ο μοριακός μηχανισμός της πρωτεολυτικής ενεργοποίησης της Spike πρωτεΐνης (ακίδα) έχει αποδειχθεί ότι παίζει κρίσιμο ρόλο: (α) στην επιλογή των ειδών του ξενιστή, (β) στη σύνδεση του ιού στον κυτταρικό υποδοχέα ACE2, (γ) στη σύντηξη ιού-κυττάρου και (δ) στην επιμόλυνση των ανθρώπινων πνευμονικών κυττάρων-στόχων από τον ιό. Η Spike πρωτεΐνη (ακίδα) περιέχει τρεις θέσεις πρωτεολυτικής διάσπασης - σχάσης: τις θέσεις R685↓S και R815↓S που χρησιμεύουν ως άμεσοι στόχοι της πρωτεάσης Φουρίνη (Furin), και τη θέση T696↓M που μπορεί να αναγνωριστεί από την πρωτεάση TMPRSS2. Η ανάλυση των νεο-σχηματιζόμενων, επαγόμενων από μετάλλαξη, C/H-CrUPs επί της Spike πρωτεΐνης (ακίδα) αποκαλύπτει ότι νέα πεπτίδια, λόγω των μεταλλάξεων, δημιουργούνται αποκλειστικά και μόνο γύρω από την κρίσιμη θέση διάσπασης R685↓S των δύο παθογόνων μεταλλάξεων P681H και P681R (Πίνακας 26 και Εικόνα 82).

Διαπιστώνεται, με ιδιαίτερο πραγματικά ενδιαφέρον, ότι μόνο η κρίσιμη μετάλλαξη P681R, στις παραλλαγές Delta και Kappa, μπορεί και δημιουργεί ένα νέο C/H-CrUP, το οποίον περιέχει και τη θέση διάσπασης R685↓S, στην οποίαν δρα πρωτεολυτικά η Φουρίνη. Φαίνεται, λοιπόν, ότι η αντικατάσταση της Προλίνης (P) με Αργινίνη (R), στην αμινοξική θέση 681, προκαλεί την απώλεια της μοναδικότητας της αλληλουχίας αμινοξέων που χαρακτηρίζει το φυσικού-τύπου C/H-CrUP "PRRARS↓V", ενώ μέσω σημαντικής σταθεροποίησης της θέσης R685↓S διευκολύνεται η διαδικασία υδρολυτικής διάσπασης - σχάσης (στοχευμένης πρωτεόλυσης) της Spike πρωτεΐνης (ακίδα) από την πρωτεάση Φουρίνη, και επακόλουθα επάγεται η ταχεία, επιτυχής και αποτελεσματική είσοδος του ιού στο κύτταρο-ξενιστή.

Mutation position	Mutation type	Variant	New C/H-CrUPs (first AA position)	New C/H-CrUPs (AA sequence)
19	T19R	Delta_P0DTC2	-	-
70	V70F	Delta_P0DTC2	69	HFSGTN
			70	FSGTNG
75 - 76	G75V&T76I	Lambda_P0DTC2	71	SGTNVI
			75	VIKRFD
222	A222V	Delta_P0DTC2	218	QGFSVL
258	W258L	Delta_P0DTC2	-	-
417	K417N	Delta_P0DTC2	413	GQTGNI
			414	QTGNIA
452	L452R	Delta_P0DTC2	449	YNYRY
		Kappa_P0DTC2		
		Alpha_P0DTC2		
	L452Q	Lambda_P0DTC2	448	NYNYQ
			449	YNYQY
478	T478K	Delta_P0DTC2	474	QAGSKP
			478	KPCNG
484	E484Q	Kappa_P0DTC2	481	NGVQG
			483	VQGFN
			484	QGFNC
	E484K	Alpha_P0DTC2	484	KGFNC
490	F490S	Lambda_P0DTC2	487	NCYSP
494	S494P	Alpha_P0DTC2	-	-
501	N501Y	Alpha_P0DTC2	498	QPTY
			499	PTYG
			500	TYGV
			501	YGVG
570	A570D	Alpha_P0DTC2	568	DIDDTT
614	D614G	Delta_P0DTC2	609	AVLYQG
		Kappa_P0DTC2		
		Alpha_P0DTC2		
		Lambda_P0DTC2		
		Delta_P0DTC2	610	VLYQGV
		Kappa_P0DTC2		
		Alpha_P0DTC2		
		Lambda_P0DTC2		
681	P681R	Delta_P0DTC2	680	SRRRARS
		Kappa_P0DTC2		
	P681H	Alpha_P0DTC2	677	QTNSH
			678	TNSHR

			680	SHRRAR
716	T716I	Alpha_P0DTC2	714	IPINF
859	T859N	Lambda_P0DTC2	855	FNGLNV
			857	GLNVLP
950	D950N	Delta_P0DTC2	946	GKLQN
			947	KLQNVV
			948	LQNVVN
			949	QNVVNQ
982	S982A	Alpha_P0DTC2	978	NDILAR
1071	Q1071H	Kappa_P0DTC2	1067	YVPAH
			1069	PAHEKN
			1071	HEKNF
1118	D1118H	Alpha_P0DTC2	1113	QIITTH
			1115	ITTHN
			1116	TTHNT
			1117	THNTF
			1118	HNTFV

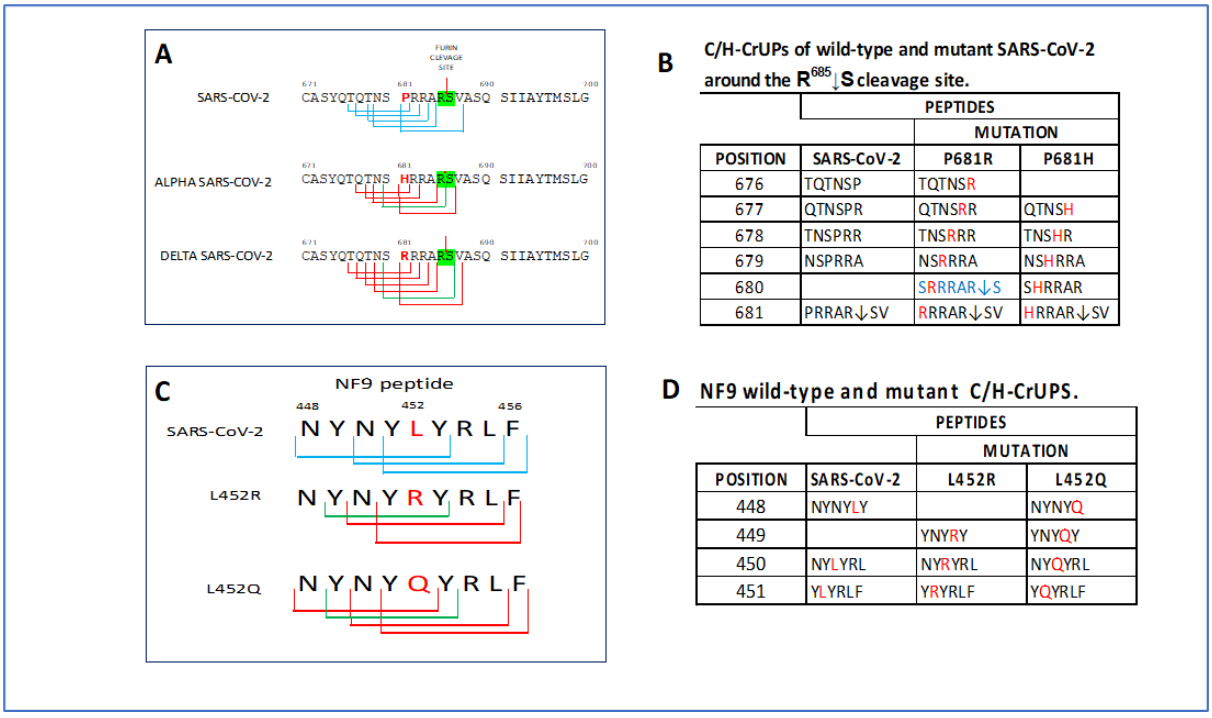
Πίνακας 25 Νέο-σχηματιζόμενα C/H-CrUPs της Spike πρωτεΐνης (ακίδα) του ιού SARS-CoV-2 στις υπό-παραλλαγές Alpha, Delta, Kappa και Lambda

Cleavage site	Mutation	Variant	New C/H-CrUPs (first AA position)	New C/H-CrUPs
R685↓S	P681R	Delta & Kappa	680	SRRRAR↓S
	P681H	Alpha & Gamma	677	QTNSH
			678	TNSHR
			680	SHRRAR
T696↓M	A701V	Beta	None	
R815↓S	None		None	

Πίνακας 26 Νεο-σχηματιζόμενα C/H-CrUPs, γύρω από τις θέσεις διάσπασης της Spike πρωτεΐνης (ακίδα) του ιού SARS-CoV-2

Ταυτοποίηση των νέων C/H-CrUPs που δημιουργούνται από τις μεταλλάξεις της Spike πρωτεΐνης (ακίδα) του ιού SARS-CoV-2. Πρώτη στήλη: Η θέση πρωτεολυτικής διάσπασης της Spike πρωτεΐνης (ακίδα) του SARS-CoV-2. Το σύμβολο «↓» υποδηλώνει την ακριβή θέση διάσπασης. Δεύτερη στήλη: Η μετάλλαξη που εντοπίζεται γύρω από τη θέση διάσπασης. Τρίτη στήλη: Οι παραλλαγές του ιού στις οποίες εμφανίζεται η μετάλλαξη. Τέταρτη στήλη: Η θέση στην αλληλουχία της Spike πρωτεΐνης (ακίδα) του SARS-CoV-2, στην οποία εμφανίζεται το πρώτο αμινοξύ του εκάστοτε C/H-CrUP. Πέμπτη στήλη: Η αλληλουχία του κάθε νέου C/H-CrUP. Το σύμβολο «↓» υποδεικνύει τη θέση διάσπασης εντός πεπτιδίου (Πίνακας 26).

Επιπλέον, λόγω των μεταλλαγμένων νεο-σχηματιζόμενων C/H-CrUPs, μπορούν και παράγονται νέα μικρά γραμμικά μοτίβα (Small Linear Motifs: SLiMs) στις παραλλαγές του ιού, όπως τα “SRRR”, “RRR”, “RRRAR” και “RRRARS”, τα οποία φαίνεται να λειτουργούν ως νέοι στόχοι πρωτεολυτικών ενζύμων διαφορετικών από τη Φουρίνη, επιτρέποντας έτσι αφενός την ισχυρότερη δέσμευση του μεταλλαγμένου ιού στον υποδοχέα ACE2 του ξενιστή, και αφετέρου την ταχύτερη είσοδο του ιού στο κύτταρο, η οποία οδηγεί σε μία συγκριτικά πιο γενικευμένη μόλυνση, καθώς και στη μαζική παραγωγή μεταλλαγμένου ιού. Αξίζει να σημειωθεί, ότι τα νέα SLiMs που περιέχονται στα νεο-δημιουργηθέντα C/H-CrUPs από τη μετάλλαξη P681R, στην παραλλαγή SARS-CoV-2 Delta, θα μπορούσαν επιπλέον να καθιστούν την Spike πρωτεΐνη (ακίδα) αντιγονικά αδύναμη ή ελαττωματική (ή ακόμη και ανεπαρκή), με αποτέλεσμα την απώλεια της ικανότητάς της να λειτουργεί ως στόχος (θεραπευτικών) αντισωμάτων, προωθώντας έτσι την ανοσολογική διαφυγή του ιού.



Εικόνα 82 Νέα C/H-CrUPs γύρω από τη θέση διάσπασης R685↓S και το πεπτίδιο NF9 της Spike πρωτεΐνης (SPIKE_SARS2, P0DTC2).

Στην παραπάνω εικόνα παρουσιάζονται:

(**A**) Η αλληλουχία των αμινοξέων της Spike πρωτεΐνης (ακίδα) μεταξύ των θέσεων 671 και 700, στο φυσικό-τύπο και στις παραλλαγές Alpha και Delta του ιού SARS-CoV-2. Σε κάθε παραλλαγή, επισημαίνονται τα αναγνωρισμένα C/H-CrUPs. Οι μπλε γραμμές υποδεικνύουν τα C/H-CrUPs που προέρχονται από την πρωτεΐνη φυσικού-τύπου γύρω από τη θέση πρωτεολυτικής διάσπασης R685↓S. Οι κόκκινες γραμμές υποδηλώνουν τα C/H-CrUPs που παράγονται από τις μεταλλάξεις P681H και P681R. Οι πράσινες γραμμές υποδεικνύουν τα

νεο-σχηματισθέντα C/H-CrUPs που δημιουργούνται από τις μεταλλάξεις P681H και P681R στις παραλλαγές Alpha και Delta, αντίστοιχα.

(**B**) Το σύνολο των C/H-CrUPs που δημιουργούνται γύρω από τη θέση διάσπασης R685↓S στη μορφή της φυσικού-τύπου και της μεταλλαγμένης Spike πρωτεΐνης (ακίδα). (**C**) Παρουσίαση της αλληλουχίας αμινοξέων του πεπτιδίου NF9 μεταξύ των θέσεων 448 και 456 στην Spike πρωτεΐνη (ακίδα) φυσικού-τύπου, καθώς και στις μεταλλάξεις L452R και L452Q. Οι μπλε γραμμές υποδεικνύουν τα C/H-CrUPs που ανήκουν στο πεπτίδιο NF9. Οι κόκκινες γραμμές υποδηλώνουν τα C/H-CrUPs που παράγονται από τις μεταλλάξεις L452R και L452Q. Οι πράσινες γραμμές υποδεικνύουν τα νέο-σχηματιζόμενα C/H-CrUPs από τις μεταλλάξεις.
D) Παρουσιάζεται το σύνολο των C/H-CrUPs που βρίσκονται στο πεπτίδιο NF9 στην Spike πρωτεΐνη (ακίδα) του ιού φυσικού-τύπου, καθώς και στις μεταλλαγμένες πρωτεϊνικές μορφές αυτού.

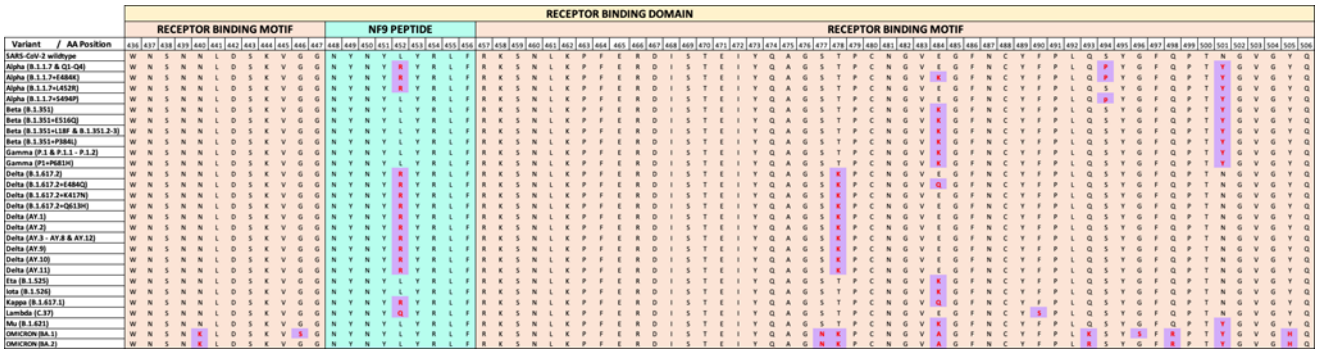
Μία ιδιαίτερα σημαντική περιοχή στο Receptor-Binding Motif (RBM) της Spike πρωτεΐνης (ακίδα) είναι το πεπτίδιο “NYNYLYRLF” (από τη θέση 448 έως τη θέση 456) (Εικόνα 80). Αυτό το πεπτίδιο είναι ιδιαίτερης σημασίας, καθώς εμφανίζεται εμπλουτισμένο με Τυροσίνη (Y), περιέχει δύο θέσεις επαφής (Y449 και Y453) με τον υποδοχέα ACE2 και είναι γνωστό ως πεπτίδιο NF9. Επιπλέον, το πεπτίδιο αυτό φαίνεται να επηρεάζει την αναγνώριση του αντιγόνου, καθώς είναι ένας ανοσο-κυρίαρχος επίτοπος των HLA*24:02 που αναγνωρίζεται από τα $CD8^+$ Τ-κύτταρα, με τη διέγερσή του να αυξάνει την παραγωγή κυτ(ταρ)οκίνων, όπως των IFNγ, TNFα και IL2. Μελέτη των C/H-CrUPs που σχετίζονται με το πεπτίδιο NF9 έδειξε την ύπαρξη τριών (3) C/H-CrUPs (Εικόνα 82Γ και Δ).

Ανάλυση μεταλλάξεων αποκάλυψε ότι στο πεπτίδιο NF9 η μετάλλαξη L452R εμφανίζεται στις παραλλαγές Alpha, Delta και Kappa, ενώ η μετάλλαξη L452Q παρατηρείται στην παραλλαγή Lambda (Εικόνα 80). Αυτές οι μεταλλάξεις ανιχνεύονται στο αμινοξύ που βρίσκεται στη θέση 5, ακριβώς στη μέση του πεπτιδίου, δημιουργώντας τρία (3) και τέσσερα (4) νέα C/H CrUPs, αντίστοιχα (Εικόνα 82Δ). Οι εν λόγω μεταλλάξεις προκαλούν δραματική επίδραση στη μοναδικότητα των CrUPs του NF9. Είναι εντυπωσιακό, ότι το C/H-CrUP μήκους 6 αμινοξέων “NYNYLY” χάνει τη μοναδικότητά του έναντι του ανθρώπινου πρωτεώματος, ενώ μόνο από τη μετάλλαξη L452Q δημιουργείται ένα νέο CrUP με μήκος 5 αμινοξέων (Εικόνα 82Γ και Δ). Η απώλεια της μοναδικότητας του συγκεκριμένου CrUP, το οποίον εντοπίζεται στην αρχή του πεπτιδίου NF9, φαίνεται να είναι απολύτως κρίσιμη, καθώς οδηγεί στην καταστροφή της αντιγονικής ικανότητας του πεπτιδίου NF9, αποφεύγοντας, ως εκ τούτου, την προκαλούμενη από το HLA-A24 ανοσία, και επάγοντας έτσι την ανοσολογική διαφυγή του ιού. Είναι ιδιαίτερα ενδιαφέρον, ότι η μετάλλαξη L452R (και άρα τα νέα C/H-CrUPs που δημιουργούνται από αυτήν) αυξάνουν τη μολυσματικότητα

του SARS-CoV-2, ενώ ταυτόχρονα ενισχύουν τις ηλεκτροστατικές αλληλεπιδράσεις αυτής της περιοχής επί της Spike πρωτεΐνης (ακίδα) με τον υποδοχέα ACE2, σταθεροποιώντας έτσι την πρόσδεση του ιού στον υποδοχέα του κυττάρου-στόχου.

Τα μέχρι πρότινος επιδημιολογικά δεδομένα έδειχναν ότι η κυρίαρχη και πλέον παθογόνος παραλλαγή του SARS-CoV-2 είναι η παραλλαγή Delta. Υπό το πρίσμα των προαναφερθέντων ευρημάτων μας, η ενισχυμένη παθογένεια της παραλλαγής Delta φαίνεται να είναι το αποτέλεσμα της ταυτόχρονης και συσσωρευτικής παρουσίας δύο κρίσιμων μεταλλάξεων, των L452R και P681R, στην εν λόγω παραλλαγή. Η μετάλλαξη L452R, μέσω της απώλειας της μοναδικότητας του πεπτιδίου NF9, προκαλεί ανοσολογική διαφυγή του ιού και ισχυρή σύνδεσή του με τον κυτταρικό υποδοχέα του, ενώ ταυτόχρονα η μετάλλαξη P681R φαίνεται να διευκολύνει τη διαδικασία διάσπασης της Spike πρωτεΐνης (ακίδα) από διαφορετικές πρωτεάσες, διευκολύνοντας έτσι την είσοδο του ιού στα κύτταρα του ξενιστή, προκαλώντας ταυτόχρονα γενικευμένη μόλυνση, και σε δεύτερο επίπεδο επάγοντας τη μαζική απελευθέρωση του ιού. Επομένως, η παραλλαγή Delta κερδίζει ένα σημαντικό πλεονέκτημα διαφυγής από το ανοσοποιητικό σύστημα, εμφανίζοντας αυξημένη μολυσματικότητα ως αποτέλεσμα της εισόδου του ιού στο κύτταρο-ξενιστή, καθώς και αύξηση του σχηματισμού του ιού και μαζική απελευθέρωσή του στον έξω-κυτταρικό χώρο.

Πρόσφατα, ταυτοποιήθηκε η παραλλαγή Omicron του SARS-CoV-2. Σε πρώτη φάση, στη Νότια Αφρική, ανιχνεύθηκε η υπό-παραλλαγή 1 (Omicron BA.1) και πολύ πρόσφατα χαρακτηρίστηκε η υπό-παραλλαγή 2 (Omicron BA.2). Η υπό-παραλλαγή 1 προσδιορίζεται από 30 σημειακές μεταλλάξεις, 3 μικρές ελλείψεις και 1 ένθεση, ενώ η υπο-παραλλαγή 2 τυποποιείται από 27 σημειακές μεταλλάξεις και 1 έλλειψη. Η δύο αυτές υπο-παραλλαγές φέρουν 21 κοινές μεταλλάξεις. Εξ' αυτών, 15 μεταλλάξεις τόσο της BA.1 όσο και της BA.2 υπο-παραλλαγής της Omicron ανευρίσκονται στην περιοχή δέσμευσης στον υποδοχέα RBD (Receptor-Binding Domain) (Εικόνα 83). Περαιτέρω ανάλυση αποκάλυψε ότι στο RBM (Receptor-Binding Motif) περιλαμβάνονται οι 10 από τις 15 μεταλλάξεις της υπο-παραλλαγής BA.1 και οι 8 από τις 15 της αντίστοιχης BA.2. Επιπρόσθετα, διαπιστώνεται ότι στο πεπτίδιο NF9 και στις δύο υπο-παραλλαγές της Omicron δεν ανιχνεύεται καμία μετάλλαξη, ενώ η πλειοψηφία των μεταλλάξεων συσσωρεύεται από τη θέση 477 έως τη θέση 506, προς το τέλος της RBM περιοχής. Αυτό έχει ως αποτέλεσμα την αναγνώριση του υψηλότερου αριθμού νεο-σχηματισθέντων C/H-CrUPs στην περιοχή RBD/RBM, σε σύγκριση με όλες τις προηγούμενες παραλλαγές του ιού, και κυρίως τις Alpha και Delta, οι οποίες αντιπροσωπεύουν δύο από τις πλέον κυρίαρχες SARS-CoV-2 παραλλαγές στον ανθρώπινο πληθυσμό.



Εικόνα 83 Παρουσίαση των μεταλλάξεων της Spike πρωτεΐνης (ακίδα) του SARS-CoV-2 στην περιοχή του RBM (Receptor-Binding Motif).

Φαίνεται λοιπόν ότι, σε αντίθεση με τις παραλλαγές Alpha και Delta, στο τέλος της περιοχής RBM των υπο-παραλλαγών BA.1 και BA.2 της Omicron, από τη θέση αμινοξέων 477 έως 506, εντοπίζονται 8 και 7 νέες μεταλλάξεις, με αποτέλεσμα την παραγωγή είκοσι οκτώ (28) και είκοσι πέντε (25) νέων C/H-CrUPs, αντίστοιχα. Το πιο σημαντικό εύρημα της παρούσας μελέτης είναι ότι στις υπο-παραλλαγές της Omicron δημιουργούνται νέα C/H-CrUPs που περιλαμβάνουν δύο ή τρία μεταλλαγμένα αμινοξέα, όπως τα πεπτίδια “QAGN*K*P”, “N*K*PCN”, “LK*SYS*F” και “K*SYS*FR*”, ως αποτέλεσμα της συσσώρευσης πολλαπλών μεταλλάξεων στις αμινοξικές θέσεις 440, 446, 477, 478 και 493-505. Είναι πραγματικά εντυπωσιακό, ότι αυτά τα νέα C/H-CrUPs που περιέχουν μεταλλαγμένα αμινοξέα δεν παρατηρούνται σε καμία άλλη (υπο-)παραλλαγή του ιού, μέχρι σήμερα (Πίνακας 27). Λαμβάνοντας υπόψιν τα πρόσφατα δεδομένα σχετικά με τη μολυσματικότητα του ιού, αυτή η πολύ-μεταλλαγμένη, νέα και κρίσιμη ομάδα των C/H-CrUPs φαίνεται να αλλάζει ριζικά τη δομή και τους επίτοπους της περιοχής RBM, στις υπό-παραλλαγές της Omicron, προκαλώντας σοβαρή μείωση της αντιγονικής ικανότητας της περιοχής, και επακόλουθη διευκόλυνση της ανοσολογικής διαφυγής του ιού.

Alpha variant

C/H-CrUP	Position	Mutation	New C/H-CrUPs	Position
GNYNYL	447	L452R	GNYNYR	447
NYNYLY	448			
			YNYRY	449
NYLYRL	450		NYRYRL	446
YLYRLF	451		YRYRLF	447
LYRLFR	452			
CNGVEG	480	E484K	CNGVKG	480
NGVEGF	481		NGVKGF	481
GVEGFN	482		GVKGFN	482
			KGFNC	484
YFPLQS	489	S494P N501Y	YFPLQP	489
FPLQSY	490		FPLQPY	490
PLQSYG	491		PLQPYG	491
QSYGF	493		QPYGF	493
SYGFQP	494		PYGFQP	494
GFQPTN	496			
FQPTNG	497			
QPTNGV	498			
			QPTY	498
PTNGVG	499		PTYG	499
TNGVGY	500		TYGV	500
NGVGYQ	501		YGVG	501

Delta variant

C/H-CrUP	Position	Mutation	New C/H-CrUPs	Position
PGQTGKI	412	K417N		
GQTGKIA	413		GQTGNI	413
			QTGNIA	414
TGKIAD	415		TGNIAD	415
GKIADY	416		GNIADY	416
GNYNYL	447	L452R	GNYNYR	447
NYNYLY	448			
			YNYRY	449
NYLYRL	450		NYRYRL	450
YLYRLF	451		YRYRLF	451
LYRLFR	452			
YQAGST	473	T478K	YQAGSK	473
			QAGSKP	474
AGSTPC	475		AGSKPC	475
			GSKPCN	476
STPCN	477		KPCNG	478

Omicron variant

C/H-CrUP	Position	Mutation	New C/H-CrUPs	Position
NLCPFG	334	G339D	NLCPFD	334
LCPFGE	335		LCPFDE	335
PFGEVF	337		PFDEV	337
FGEVFN	338		FDEVFN	338
GEVFNA	339		DEVFNA	339
VLYNSA	367	S371L S373P S375F	VLYNLAP	367
LYNSAS	368			
YNSASF	369		YNLAPF	369
NSASFST	370			
			LAPFFT	371
ASFSTF	372		APFFTF	372
SFSTFK	373			
			FFTFK	374
STFKC	375		FTFKCY	375
PGQTGKI	412	K417N		
GQTGKIA	413		GQTGNI	413
			QTGNIA	414
TGKIAD	415		TGNIAD	415
GKIADY	416		GNIADY	416
WNSNN	436	N440K G446S	WNSNKL	436
SNNLDS	438		SNKLDS	438
NNLDSK	439		NKLDSKV	439
			KLDSKVS	440
LDSKVG	441			
DSKVGG	442		DSKVSG	441
KVGGNY	444		KVSGNY	443
VGGNYN	445		VSGNYN	444
GGNYNY	446			
IYQAGS	472	S477N T478K		
YQAGST	473		YQAGN	473
			QAGNKP	474
AGSTPC	475			
STPCN	477		NKPCN	476
			KPCNG	478
CNGVEG	480	E484A	CNGVAG	480
NGVEGF	481		NGVAGF	481
GVEGFN	482		GVAGFN	482
			VAGFNC	483
CYFPLQ	488	Q493K G496S Q498R N501Y Y505H	CYFPLK	488
YFPLQS	489		YFPLKS	489
FPLQSY	490		FPLKSY	490
PLQSYG	491			
			LKSYSF	492
QSYGF	493		KSYSFR	493
SYGFQP	494		SYSFRP	494
YGFQPT	495		YSFRPT	495
GFQPTN	496			
FQPTNG	497		FRPTY	497
QPTNGV	498		RPTYGV	498
PTNGVG	499			
TNGVGY	500		TYGVGH	500
NGVGYQ	501			
GVGYQP	502		GVGHQ	502
VGYQPY	503		VGHQPY	503
GYQPYR	504			
YQPYRV	505		HQPYR	505

Πίνακας 27 Καταγραφή των C/H-CrUPs που ανήκουν στις παραλλαγές Alpha, Delta και Omicron, επί της περιοχής RBD (Receptor-Binding Domain) της Spike πρωτεΐνης (ακίδα) του ιού SARS-CoV-2

Είναι αξιοσημείωτο, ότι η περιοχή RBM περιέχει 11 από τα 12 σημεία επαφής της Spike πρωτεΐνης (ακίδα) του ιού SARS-CoV-2 με τον κυτταρικό υποδοχέα ACE2. Μεταξύ αυτών, 7 σημεία επαφής παραμένουν άθικτα, ενώ εντοπίζονται 4 μεταλλάξεις στις θέσεις Q493K, Q498R, N501Y και Y505H, τόσο στην υπο-παραλλαγή BA.1 όσο και στην αντίστοιχη BA.2, με αποτέλεσμα τη δημιουργία δεκαεπτά (17) νέων C/H-CrUPs. Από αυτές, η μετάλλαξη N501Y έχει δειχθεί να λειτουργεί ως ένας κύριος και καθοριστικός παράγοντας της αυξημένης μετάδοσης του ιού, λόγω της ισχυρότερης συγγένειας σύνδεσης της Spike πρωτεΐνης (ακίδα) με τον κυτταρικό υποδοχέα της ACE2. Τα ευρήματά μας υποδεικνύουν ότι η δέσμευση του ιού με τον υποδοχέα ACE2 επηρεάζεται σημαντικά από τα μεταλλαγμένα C/H-CrUPs, τα οποία πιθανότατα μπορούν και ενισχύουν την αλληλεπίδραση της Spike με την ACE2 πρωτεΐνη.

4.4 Επέκταση της βάσης δεδομένων Uniquome σε πρότυπους οργανισμούς

Η ανάλυση για την καταγραφή και διερεύνηση μοναδικών πεπτιδίων επεκτάθηκε (πέρα από το πρωτέωμα του ανθρώπου) και σε άλλους πρότυπους οργανισμούς μοντέλα (Πίνακας 28). Έτσι κατασκευάστηκε η βάση δεδομένων Uniquomes που περιλαμβάνει τα μοναδικά πεπτίδια των οργανισμών που μελετήθηκαν (τόσο τα μοναδικά πεπτίδια ελαχίστου μήκους όσο και τα σύνθετα μοναδικά πεπτίδια). Στη βάση δεδομένων των Uniquomes πέρα από την απλή καταγραφή των μοναδικών πεπτιδίων περιλαμβάνεται και όλη τη μετανάλυση που πραγματοποιήθηκε στου πρότυπους αυτούς οργανισμούς με σκοπό την καταγραφή των χαρακτηριστικών των μοναδικών πεπτιδίων του κάθε οργανισμού (Πίνακας 29).

Οργανισμός	Αριθμός Πρωτεϊνών
Homo Sapiens (Human)	20.430
Mus Musculus (Mouse)	17.023
Arabidopsis Thaliana (A. thaliana)	15.892
Rattus norvegicus (Rat)	8.072
Saccharomyces cerevisiae (Baker's yeast)	6.721
Bos taurus (Bovine)	6.008
Escherichia coli (E. Coli)	4.360
Caenorhabditis elegans (C. Elegans)	4.073
Drosophila melanogaster (fruit fly)	3.580
Danio rerio (Zebrafish)	3.115
Gallus gallus (Chicken)	2.291
Sus scrofa (Pig)	1.431
Oryctolagus cuniculus (Rabbit)	894
Zea mays (Maize)	831
Nicotiana tabacum (Common Tobacco)	476
Ovis Aries (Sheep)	458
Glycine max (Soybean)	409
Pisum sativum (Garden pea)	397
Triticum aestivum (Wheat)	379
Equus caballus (Horse)	288

Πίνακας 28 Οργανισμοί της βάσης δεδομένων των Uniquomes

Οργανισμοί	Πρωτεΐνες	Πρωτεΐνες με Μοναδικά πεπτίδια	Πρωτεΐνες χωρίς Μοναδικά Πεπτίδια	Πρωτεΐνες χωρίς Μοναδικά Πεπτίδια (%)	CrUP	CrUP ≥ 1	CrUP >1	CmUP	πυκνότητα CrU	πυκνότητα CmU	Unique κάλυψη
Homo Sapiens (Human)	20.430	20.282	148	0,72%	7.263.888	7.316.653	52.765	77.697	64%	0,68%	93%
Mus musculus (Mouse)	17.023	16.984	39	0,23%	6.460.608	6.489.302	28.694	60.839	67%	0,63%	96%
Arabidopsis Thaliana (A. thaliana)	15.892	15.779	113	0,71%	4.175.126	4.182.877	7.751	94.039	58%	1,30%	89%
Rattus norvegicus (Rat)	8.072	8.062	10	0,12%	2.769.118	2.775.436	6.318	27.914	67%	0,68%	96%
Saccharomyces cerevisiae (Baker's yeast)	6.721	6.528	193	2,87%	2.002.321	2.006.693	4.372	13.991	66%	0,46%	92%
Bos taurus (Bovine)	6.008	6.002	6	0,10%	1.632.435	1.636.675	4.240	13.995	69%	0,59%	96%
Escherichia coli (E. Coli)	4.360	4.290	70	1,61%	952.037	952.780	743	6.730	70%	0,49%	97%
Caenorhabditis elegans (C. Elegans)	4.073	4.064	9	0,22%	1.490.966	1.508.059	17.093	8.741	70%	0,41%	97%
Drosophila melanogaster (fruit fly)	3.580	3.553	27	0,75%	1.537.798	1.543.834	6.036	6.999	70%	0,32%	97%
Danio rerio (Zebrafish)	3.115	3.115	0	0,00%	1.006.003	1.007.401	1.398	9.750	68%	0,66%	95%
Gallus gallus (Chicken)	2.291	2.289	2	0,09%	761.295	764.140	2.845	7.215	68%	0,64%	95%
Sus scrofa (Pig)	1.431	1.428	3	0,21%	374.596	376.988	2.392	3.517	69%	0,65%	95%
Oryctolagus cuniculus (Rabbit)	894	894	0	0,00%	256.418	258.585	2.167	3.793	66%	0,97%	92%
Zea mays (Maize)	831	824	7	0,84%	145.650	146.470	820	3.565	52%	1,26%	76%
Nicotiana tabacum (Common Tobacco)	476	474	2	0,42%	77.182	77.493	311	1.952	51%	1,28%	75%
Ovis aries (Sheep)	458	457	1	0,22%	92.167	93.274	1.107	1.235	67%	0,89%	94%
Glycine max (Soybean)	409	403	6	1,47%	71.143	71.793	650	1.958	52%	1,43%	78%
Pisum sativum (Garden pea)	397	397	0	0,00%	75.138	75.742	604	1.436	59%	1,13%	85%
Triticum aestivum (Wheat)	379	376	3	0,79%	49.394	49.755	361	1.484	46%	1,37%	71%
Equus caballus (Horse)	288	287	1	0,35%	62.104	62.534	430	836	66%	0,88%	92%

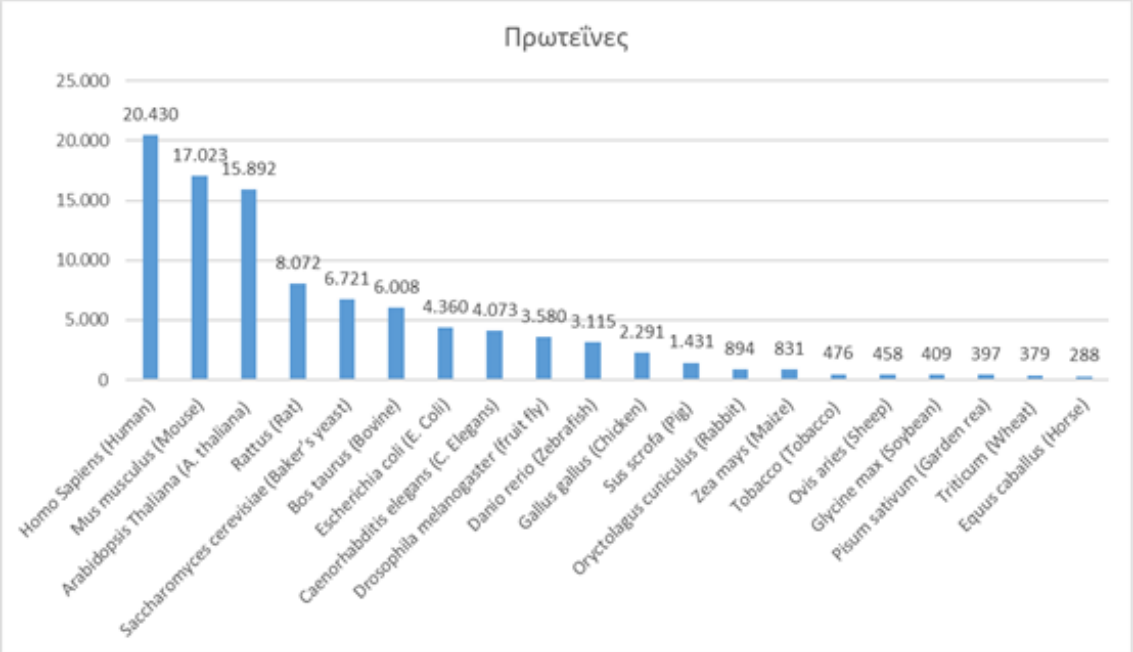
Πίνακας 29 Συγκεντρωτικός πίνακας με τα αποτελέσματα των πρότυπων οργανισμών της βάσης δεδομένων των Uniquomes

Για την καλύτερη κατανόηση των αποτελεσμάτων της βάσης δεδομένων των Uniquomes παρουσιάζονται τα παραπάνω χαρακτηριστικά σε ξεχωριστές αναλύσεις, ανάλογα το υπό μελέτη χαρακτηριστικό, με την κάθε ανάλυση να έχει πραγματοποιηθεί:

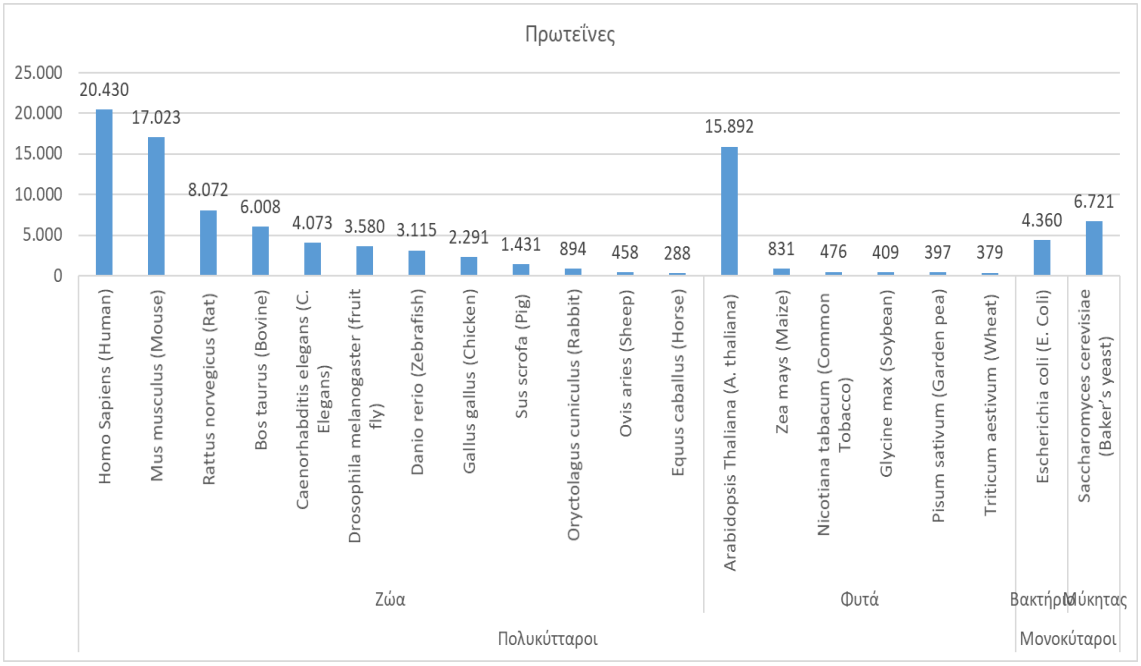
- Με τους οργανισμούς ταξινομημένους ως προς το μέγεθος του πρωτεώματος τους.
- Με τους οργανισμούς ταξινομημένους ως προς την κατηγορία οργανισμού στην οποία ανήκουν.

Η βάση δεδομένων των Uniquomes και οι Πρωτεΐνες

Αρχικά οι οργανισμοί ταξινομήθηκαν βάσει του μεγέθους του πρωτεώματος (σύμφωνα με την βάση δεδομένων Uniprot version 10/2019) και στην συνέχεια με βάση την κατηγορία που ανήκουν. ο οργανισμός του ανθρώπου είναι ο οργανισμός με τις περισσότερες reviewed πρωτεΐνες (20.430) και ακολουθούν οι οργανισμοί του Mouse και της *A. thaliana* (17.023 και 15.892 αντίστοιχα). Ο οργανισμός με τις λιγότερες reviewed πρωτεΐνες είναι αυτός του αλόγου που αποτελείται από 288 πρωτεΐνες (Εικόνα 84).Για τους πολυκύτταρους οργανισμούς αυτοί με το μεγαλύτερο πρωτέωμα είναι ο άνθρωπος για τα ζώα ενώ για τα φυτά ο οργανισμός της *A. thaliana*. Στην βάση δεδομένων των Uniquome περιλαμβάνονται και δύο μονοκύτταροι οργανισμοί με μεγαλύτερο σε αριθμό πρωτεϊνών τον μύκητα Baker's yeast (Εικόνα 85).



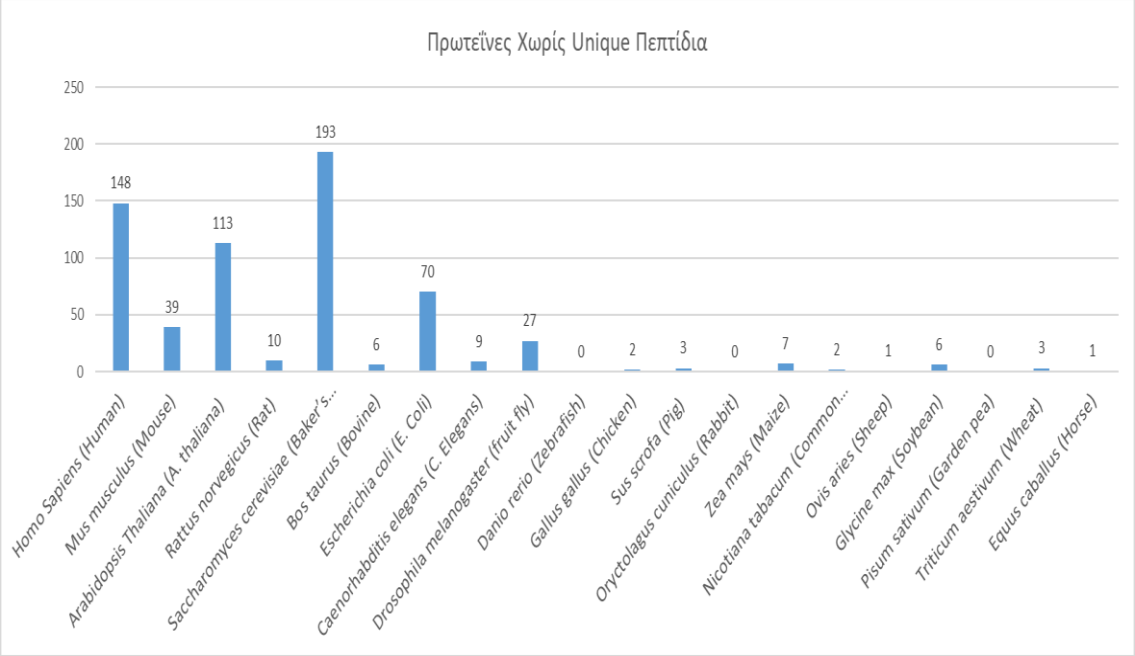
Εικόνα 84 Οι οργανισμοί της βάσης δεδομένων των Uniquomes ταξινομημένοι βάση του μεγέθους του πρωτεώματος τους



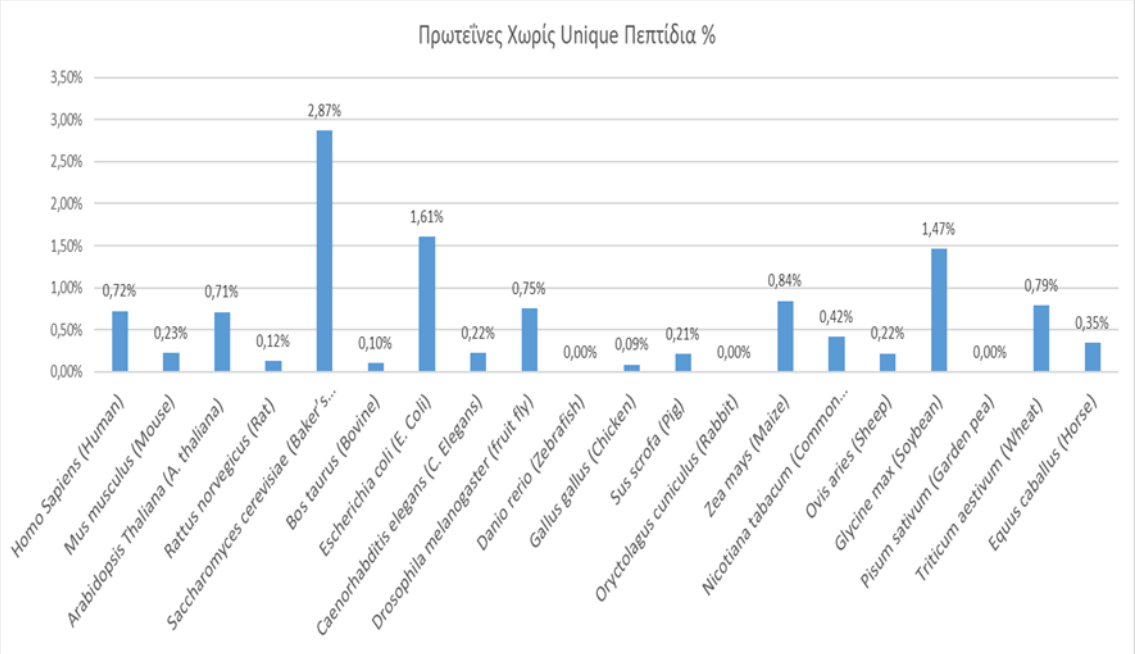
Εικόνα 85 Οι οργανισμοί της βάσης δεδομένων των Uniquomes ταξινομημένοι βάση την κατηγορία που ανήκουν και το μέγεθος του πρωτεώματος τους

Μια από τις σημαντικότερες αναλύσεις των χαρακτηριστικών των Uniquomes αφορά τις πρωτεΐνες οι οποίες δεν περιλαμβάνουν μοναδικά πεπτίδια. Τα αποτελέσματα της συγκεκριμένης ανάλυσης έδειξαν πως ο αριθμός των πρωτεϊνών οι οποίες δεν περιλαμβάνουν μοναδικά πεπτίδια δεν έχει σχέση με το μέγεθος του πρωτεώματος του κάθε οργανισμού (Εικόνα 86,87). Περεταίρω μελέτη ανάλογα με το είδος στο οποίο ανήκει ο κάθε οργανισμός και την πληροφορία για τις πρωτεΐνες του που δεν περιλαμβάνουν μοναδικά πεπτίδια ανέδειξε πως ο αριθμός των πρωτεϊνών που δεν εμφανίζουν μοναδικά πεπτίδια σχετίζεται άμεσα με το είδος του οργανισμού. Πιο συγκεκριμένα όπως παρατηρήθηκε οι μονοκύτταροι οργανισμοί έχουν αισθητά μεγαλύτερο ποσοστό από πρωτεΐνες που δεν περιλαμβάνουν μοναδικά πεπτίδια με τον Baker's Yeast να είναι ο οργανισμός με τις περισσότερες πρωτεΐνες χωρίς μοναδικά πεπτίδια αναλογικά με το μέγεθός του (193 πρωτεΐνες που αντιστοιχούν σε 2,87% ως προς το σύνολο του πρωτεώματος του). Παρατηρήθηκε επίσης ότι στου πολυκύτταρους οργανισμούς, τα φυτά έχουν μεγαλύτερο ποσοστό από πρωτεΐνες χωρίς μοναδικά πεπτίδια σε σχέση με τα Ζώα, με εξαίρεση τον οργανισμό του αρακά (garden Rea). Από του ζωικούς οργανισμούς, ο οργανισμός της Φρουτόμυγας (fruit fly) είναι αυτός με το μεγαλύτερο ποσοστό από πρωτεΐνες που δεν περιλαμβάνου μοναδικά πεπτίδια (27 πρωτεΐνες που αντιστοιχούν σε ποσοστό 0,75%) ενώ αντίθετα οι οργανισμοί του λαγού (rabbit) και του ψάρι ζέβρα (zebra fish) με το μικρότερο ποσοστό καθώς δεν έχουν καμία πρωτεΐνη που να μην περιλαμβάνει μοναδικά πεπτίδια. Στους οργανισμούς των φυτών ο οργανισμός της σόγιας (Soybean) είναι αυτός με το μεγαλύτερο ποσοστό από πρωτεΐνες χωρίς μοναδικά πεπτίδια (6 πρωτεΐνες που αντιστοιχούν σε ποσοστό 1,47%)

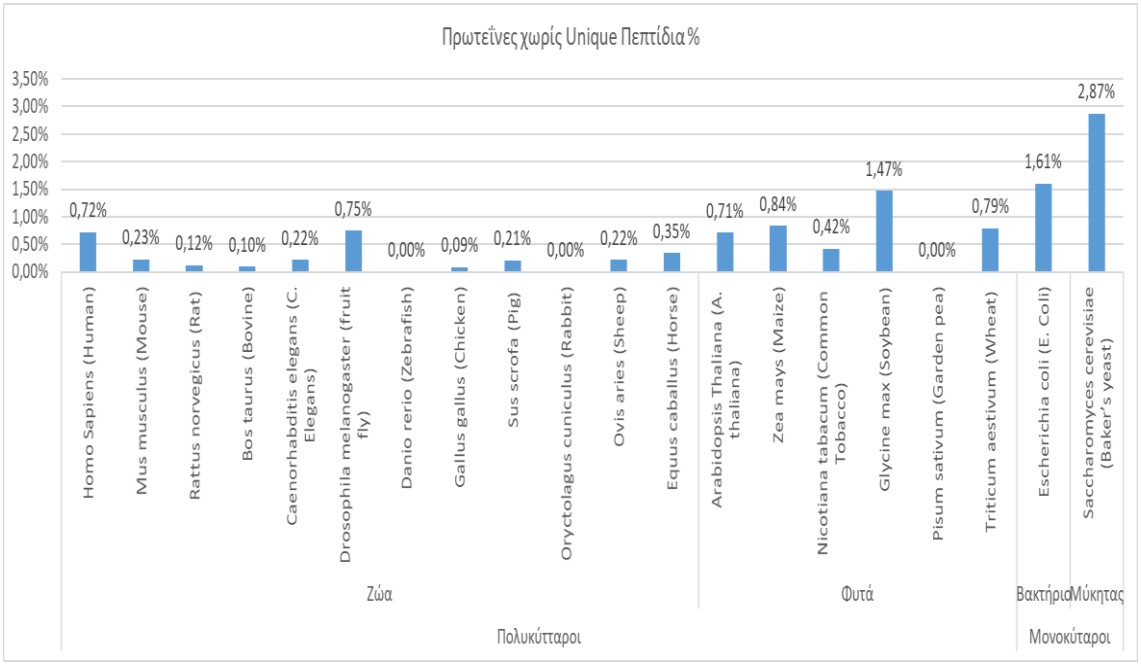
αντίθετα με τον οργανισμό του αρακά που όλες του οι πρωτεΐνες περιέχουν μοναδικά πεπτίδια (Εικόνα 88).



Εικόνα 86 Οι Οργανισμοί της βάσης δεδομένων των Uniquomes με τις πρωτεΐνες χωρίς Unique πεπτίδια ταξινομημένοι βάση του μεγέθους του πρωτεώματος τους



Εικόνα 87 Οι Οργανισμοί της βάσης δεδομένων των Uniquomes με τις πρωτεΐνες χωρίς Unique πεπτίδια (%) ταξινομημένοι βάση του μεγέθους του πρωτεώματος τους

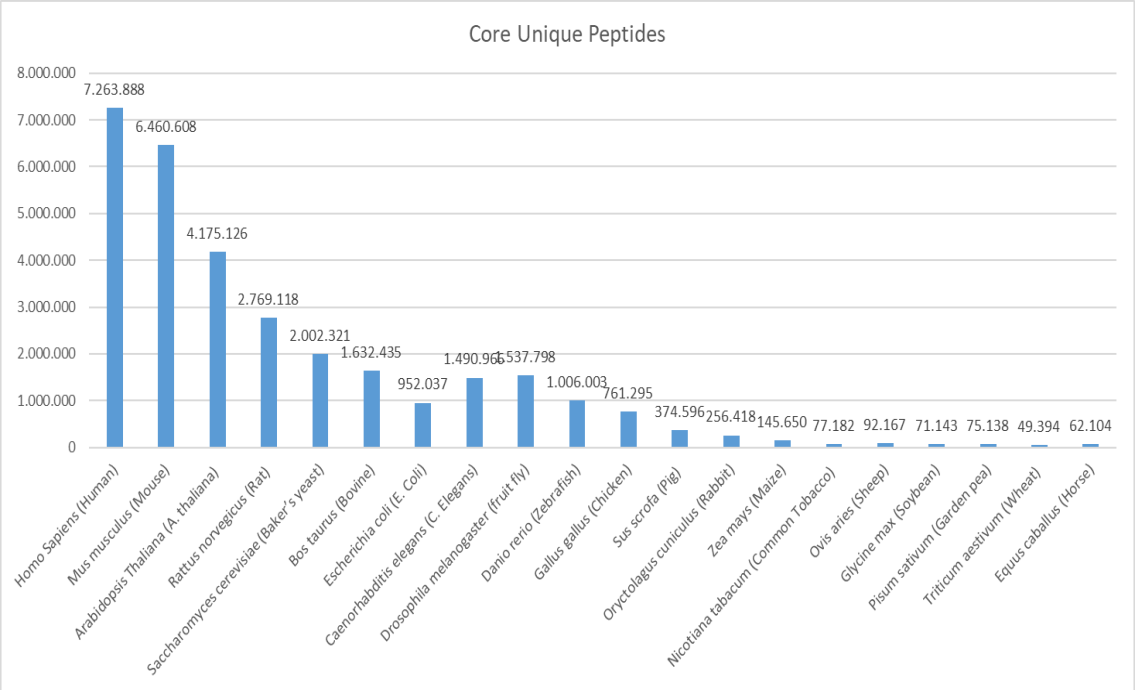


Εικόνα 88 Οι Οργανισμοί της βάσης δεδομένων των Uniquomes με τις πρωτεΐνες χωρίς Unique πεπτίδια (%) ταξινομημένοι βάση την κατηγορία που ανήκουν και το μέγεθος του πρωτεώματος τους

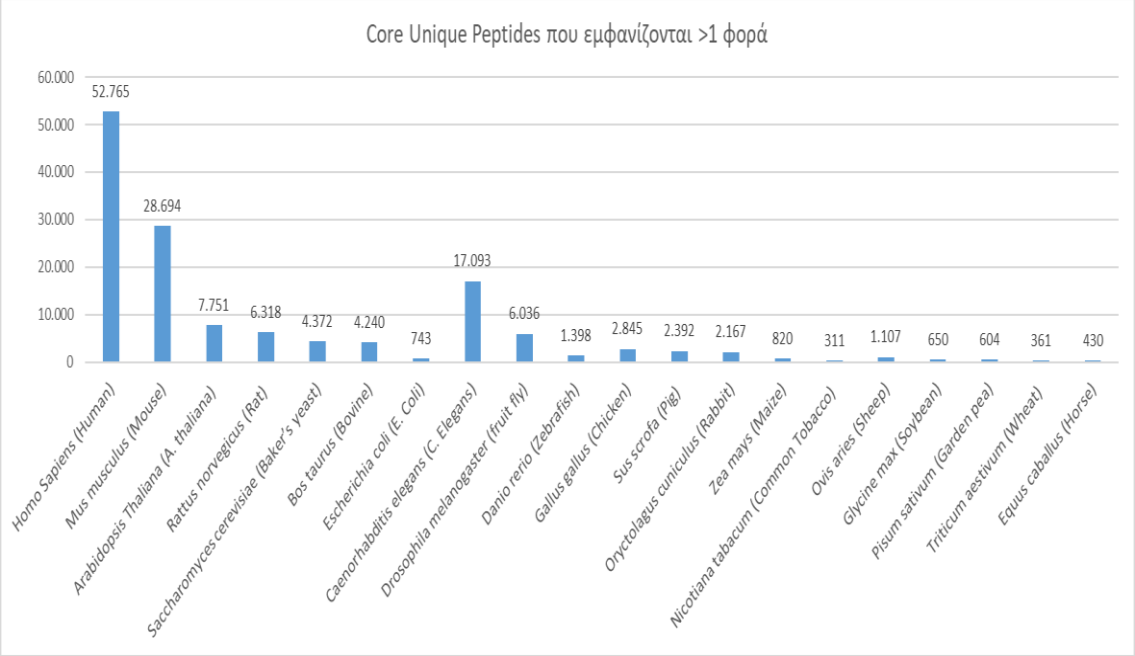
Η βάση δεδομένων των Uniquomes και τα μοναδικά πεπτίδια

Σε συνέχεια της σύγκρισης των Uniquomes των πρότυπων οργανισμών μελετήθηκε ο αριθμός των μοναδικών πεπτιδίων που περιλαμβάνονται σε κάθε οργανισμό. Αρχικά μελετήθηκαν τα μοναδικά πεπτίδια ελαχίστου μήκους τα οποία εμφανίζονται μόνο μια φορά στην ίδια πρωτεΐνη (υπάρχουν και μοναδικά πεπτίδια τα οποία εμφανίζονται περισσότερες φορές στην ίδια πρωτεΐνη όμως συνεπώς διατηρούν τη μοναδικότητα τους ως προς το συνολικό πρωτέωμα του εκάστοτε οργανισμού). Τα αποτελέσματα αυτής της ανάλυσης έδειξαν πως ο αριθμός από CrUPs του κάθε οργανισμού έχει άμεση σχέση με τον αριθμό από πρωτεΐνες που αυτός αποτελείται. Συγκεκριμένα, όσο μεγαλύτεροι σε αριθμό πρωτεϊνών είναι οι οργανισμοί τόσο τα μοναδικά πεπτίδια ελαχίστου μήκους που καταγράφηκαν στη βάση δεδομένων των Uniquomes είναι περισσότερα με εξαίρεση τον οργανισμό *E.Coli* (Εικόνα 89). Στη συνέχεια μελετήθηκαν οι οργανισμοί βάση τον αριθμό από μοναδικά πεπτίδια ελαχίστου μήκους που εμφανίζονται 2 ή περισσότερες φορές στην ίδια πρωτεΐνη. Σύμφωνα με τα αποτελέσματα αυτής της μελέτης παρατηρείται πως ο αριθμός των CrUPs που εμφανίζονται περισσότερες από μία φορές δεν εξαρτώνται ούτε από το μέγεθος αλλά και ούτε από το είδος του οργανισμού. Αναλυτικότερα στους πολυκύτταρους οργανισμούς ο οργανισμός του προβάτου (Sheep) από τα ζώα είναι αυτό που αναλογικά με τα CrUPs του έχει τον μεγαλύτερο αριθμό από μοναδικά πεπτίδια που εμφανίζονται περισσότερο από μια φορά στην ίδια πρωτεΐνη με 1.107 μοναδικά πεπτίδια ελαχίστου μήκους που

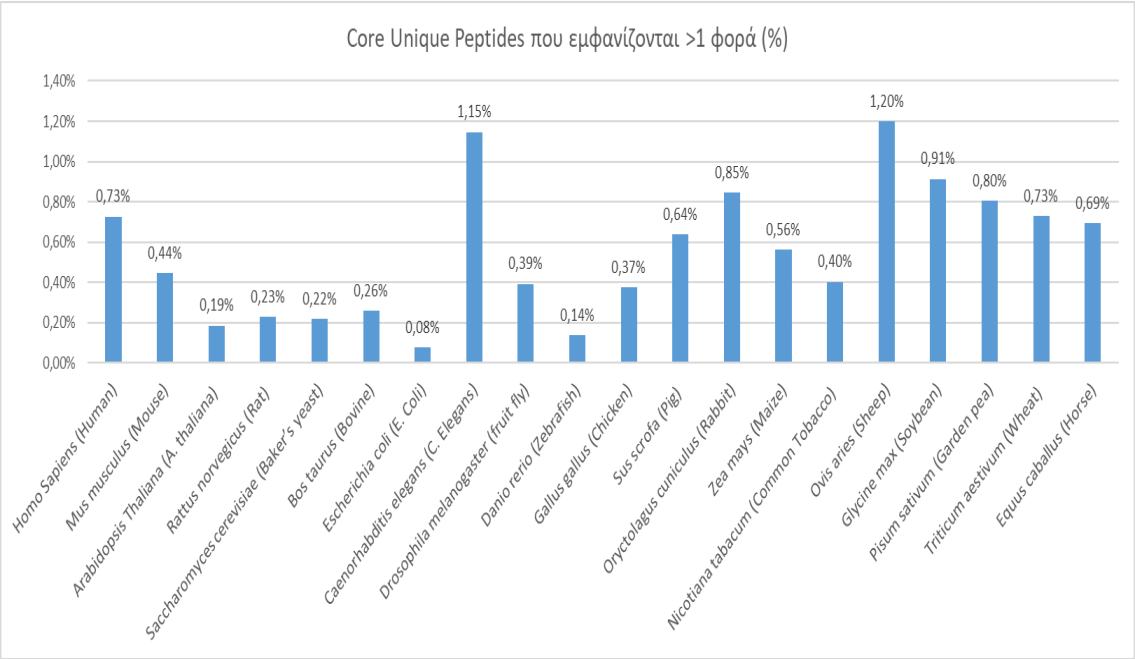
αντιστοιχούν στο 1,20% των συνολικών CrUPs ενώ αντίστοιχα ο οργανισμός της σόγιας (Soybean) με 650 μοναδικά πεπτίδια ελαχίστου μήκους που αντιστοιχούν στο 0,91% των συνολικών CrUPs για τα φυτά. Τέλος, παρατηρήθηκε πως οι μονοκύτταροι οργανισμοί έχουν αρκετό μικρό αριθμό από CrUPs που εμφανίζονται περισσότερες από μία φορές στην ίδια πρωτεΐνη (Εικόνα 90,91,92).



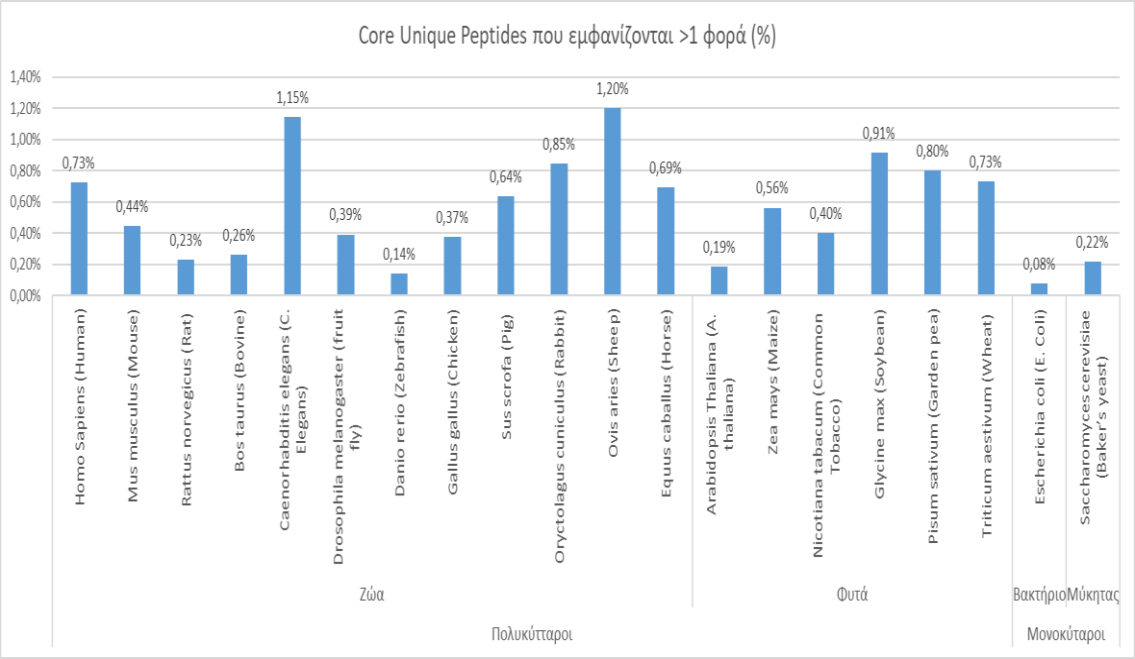
Εικόνα 89 Οι οργανισμοί της βάσης δεδομένων των Uniquomes με τα CrUPs που αποτελούνται ταξινομημένοι βάση του μεγέθους του πρωτεώματος τους



Εικόνα 90 Οι οργανισμοί της βάσης δεδομένων των Uniquomes και CrUPs που εμφανίζονται >1 ταξινομημένοι βάση του μεγέθους του πρωτεώματος τους



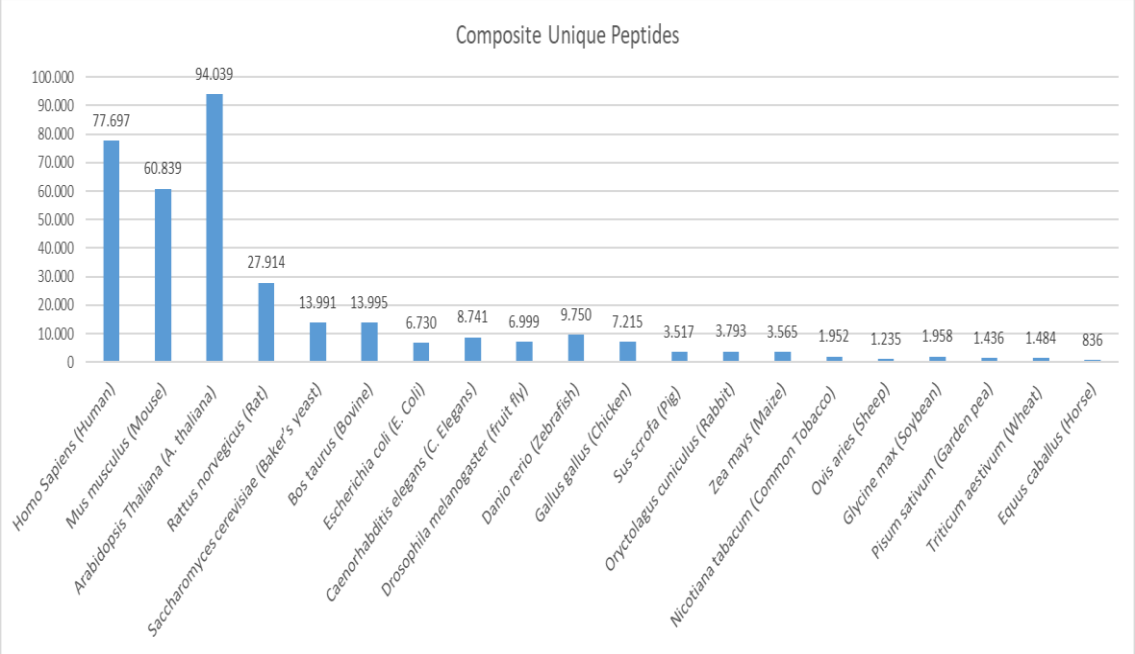
Εικόνα 91 Οι οργανισμοί της βάσης δεδομένων των Uniquomes και CrUPs που εμφανίζονται >1 (%) ταξινομημένοι βάση του μεγέθους του πρωτεώματος τους



Εικόνα 92 Οι οργανισμοί της βάσης δεδομένων των Uniquomes και CrUPs που εμφανίζονται >1 (%) ταξινομημένοι βάση την κατηγορία που ανήκουν και το μέγεθος του πρωτεώματος τους

Σε συνέχεια της ανάλυσης της βάσης δεδομένων των Uniquomes ως προς τα μοναδικά της πεπτίδια υπολογίστηκε ο αριθμός των σύνθετων μοναδικών πεπτιδίων ανά πρότυπο οργανισμό. Όπως στα CrUPs έτσι και στα CmUPs ο αριθμός πεπτιδίων του κάθε οργανισμού έχει άμεση σχέση με τον αριθμό από πρωτεΐνες που αυτός αποτελείται. Πιο συγκεκριμένα όσο μεγαλύτεροι σε αριθμό πρωτεϊνών ήταν οι οργανισμοί τόσο τα επικαλυπτόμενα μοναδικά πεπτίδια ελαχίστου μήκους που καταγράφηκαν στη βάση δεδομένων των Uniquomes ήταν περισσότερα. Εξαίρεση αποτελεί ο οργανισμός της

Arabidopsis thaliana που αναλογικά με τους άλλους οργανισμούς, έχει μεγαλύτερο αριθμό CmUPs από αυτό που αναμενόταν (Εικόνα 93).



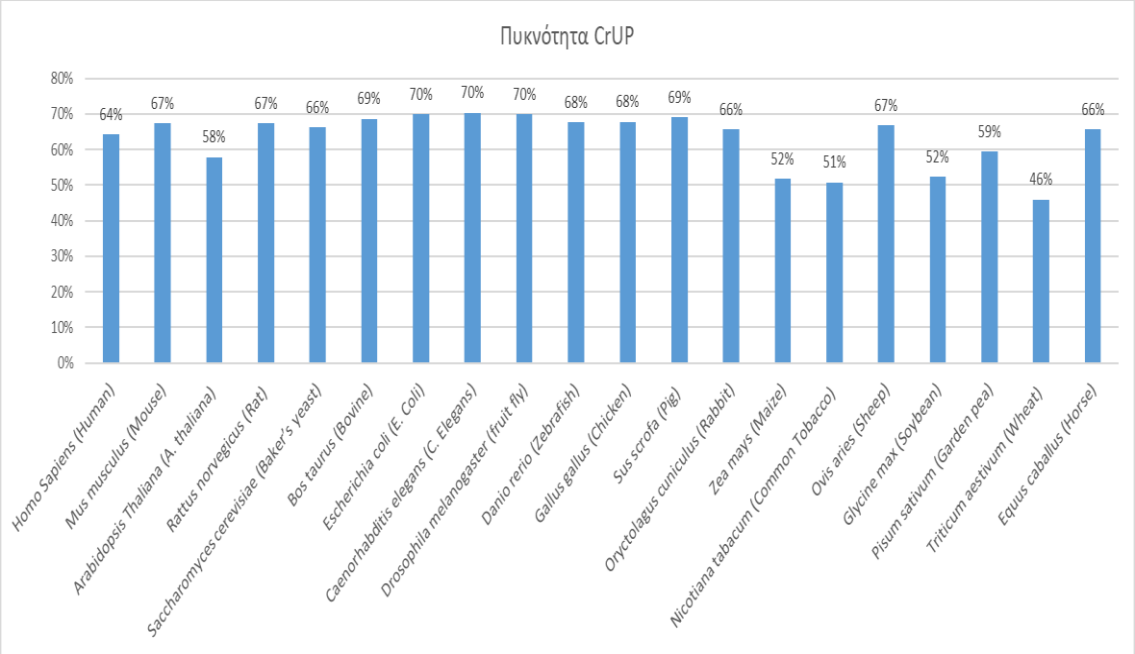
Εικόνα 93 Οι οργανισμοί της βάσης δεδομένων των Uniquomes με τα CmUPs που αποτελούνται ταξινομημένοι βάση του μεγέθους του πρωτεώματος τους

Η βάση δεδομένων των Uniquomes, πυκνότητα και μοναδική κάλυψη

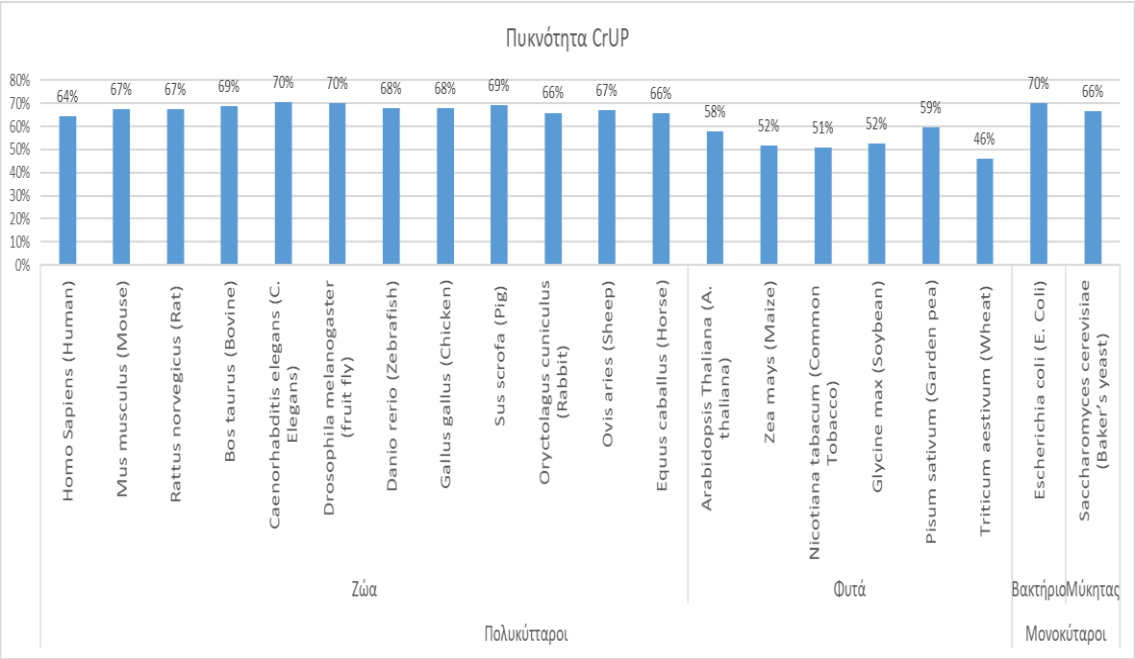
Για την καλύτερη προσέγγιση των μοναδικών πεπτιδίων ως προς το πλήθος τους σε κάθε οργανισμό της βάσης δεδομένων των Uniquomes ανάλογα με το μέγεθος τους πρωτεώματος χρησιμοποιήθηκαν οι εξής ορισμοί:

- **Πυκνότητα Core Unique Peptide οργανισμού**: Είναι το σύνολο των CrUP ενός οργανισμού ως προς το σύνολο των αμινοξέων του. Με τον όρο αυτό υπολογίζουμε την πυκνότητα (%) από CrUP που συναντάμε σε έναν οργανισμό σε σχέση με το πρωτέωμα του.
- **Πυκνότητα Composite Unique Peptide οργανισμού**: Είναι το σύνολο των CmUP ενός οργανισμού ως προς το σύνολο των αμινοξέων του. Με τον όρο αυτό υπολογίζουμε την πυκνότητα (%) από CmUP που συναντάμε σε έναν οργανισμό σε σχέση με το πρωτέωμα του.
- **Unique Κάλυψη οργανισμού:** Είναι το σύνολο των αμινοξέων ενός οργανισμού που εμπεριέχονται στο σχηματισμό των μοναδικών πεπτιδίων ως προς το σύνολο των αμινοξέων του. Με τον όρο αυτό υπολογίζουμε την κάλυψη (%) από αμινοξέα που συμμετέχουν στο σχηματισμό μοναδικών πεπτιδίων.

Αρχικά. το χαρακτηριστικό των μοναδικών πεπτιδίων για τους πρότυπους οργανισμούς που αναλύθηκε ήταν η πυκνότητα των μοναδικών πεπτιδίων ελαχίστου μήκους. Τα αποτελέσματα της ανάλυσης έδειξαν πώς η πυκνότητα των CrUP στους οργανισμούς της βάσης δεδομένων των Uniquomes εξαρτάται από το είδος του οργανισμού που ανήκουν τα μοναδικά πεπτίδια. Αναλυτικότερα στους πολυκυττάρους οργανισμούς παρατηρείται πως οι ζωικοί οργανισμοί εμφανίζονται με μεγαλύτερη πυκνότητα από μοναδικά πεπτίδια σε σχέση με τους φυτικούς οργανισμούς. Πιο συγκεκριμένα για τα ζώα, οι οργανισμοί της Φρουτόμυγας (fruit fly) και του Καινοραβδίτη (*C. elegans*) εμφανίζονται με την μεγαλύτερη πυκνότητα (που αντιστοιχεί στο ποσοστό 70%) από μοναδικά πεπτίδια ελαχίστου μήκους στο σύνολο του πρωτεώματος του εκάστοτε οργανισμού ενώ αντίθετα ο οργανισμός του ανθρώπου εμφανίζεται με την μικρότερη πυκνότητα (που αντιστοιχεί στο ποσοστό 64% από μοναδικά πεπτίδια ελαχίστου μήκους).Στους φυτικούς οργανισμούς την μεγαλύτερη πυκνότητα (59%) από μοναδικά πεπτίδια ελαχίστου μήκους την έχει οργανισμός του αρακά (garden Rea) ενώ την χαμηλότερη πυκνότητα (46%) ο οργανισμός του σιταριού (Seat). Τέλος οι μονοκύτταροι οργανισμοί ακολουθούν τα ποσοστά πυκνότητας που συναντάμε στους ζωικούς οργανισμούς με το βακτήριο *E.coli* να εμφανίζεται με την μεγαλύτερη πυκνότητα (70%) από μοναδικά πεπτίδια ελαχίστου μήκους (Εικόνα 94, εικόνα 95).

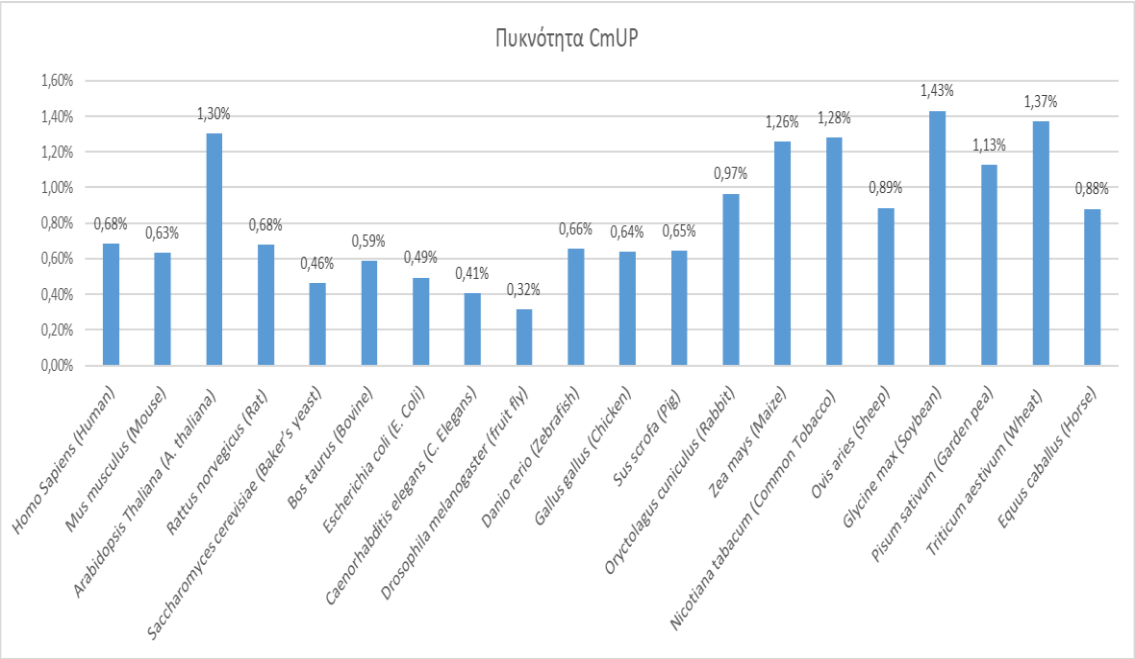


Εικόνα 94 Πυκνότητα από CrUPs για τους οργανισμούς της βάσης δεδομένων των Uniquomes ταξινομημένοι βάση του μεγέθους του πρωτεώματος τους

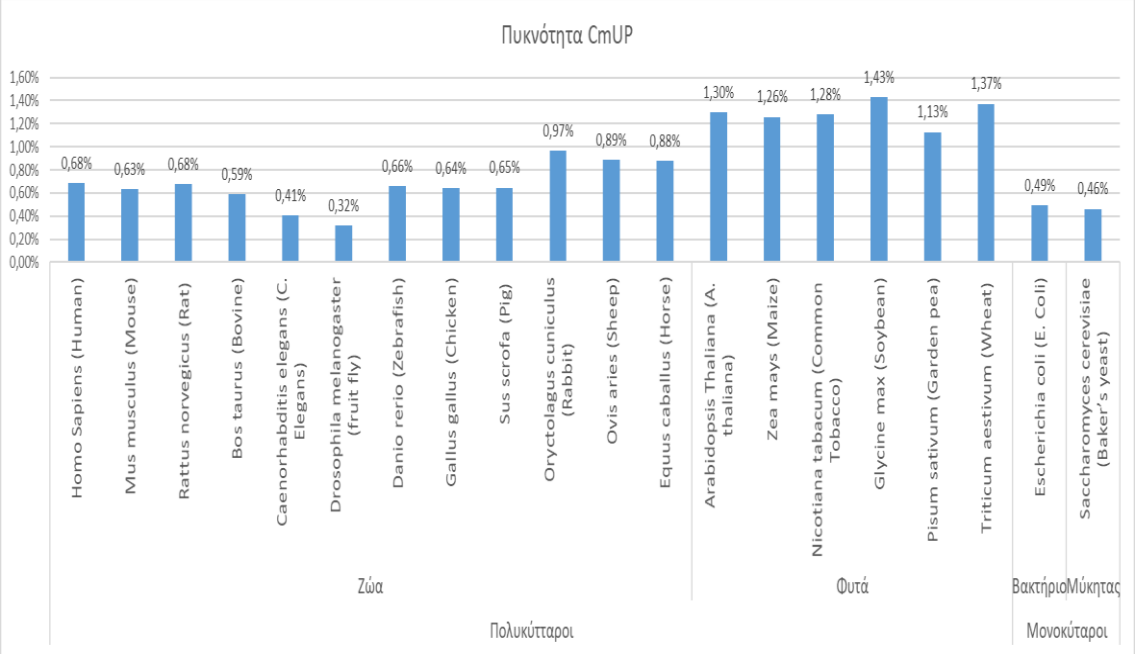


Εικόνα 95 Πυκνότητα από CrUPs για τους οργανισμούς της βάσης δεδομένων των Uniquomes ταξινομημένοι βάση την κατηγορία που ανήκουν και το μέγεθος του πρωτεώματος τους

Σε συνέχεια της ανάλυσης της πυκνότητα των μοναδικών πεπτιδίων για τους πρότυπους οργανισμούς της βάσης δεδομένων Uniquome αναλύθηκε η πυκνότητα από σύνθετα μοναδικά πεπτίδια των οργανισμών. Η ανάλυση αυτή ανέδειξε πώς η πυκνότητα από σύνθετα μοναδικά πεπτίδια των οργανισμών επηρεάζεται από το είδος στο οποίο ανήκουν. Πιο συγκεκριμένα στους πολυκύτταρους οργανισμούς παρατηρείται ότι στους φυτικούς οργανισμούς η πυκνότητα των σύνθετων μοναδικών πεπτιδίων είναι υψηλότερη σε σχέση με την πυκνότητα των ζωικών οργανισμών. Στους ζωικούς οργανισμούς ο Λαγός (Rabbit) εμφανίζεται με την μεγαλύτερη πυκνότητα σύνθετων μοναδικών πεπτιδίων (0,97%) ενώ ο οργανισμός της Φρουτόμυγας (fruit fly) με την χαμηλότερη (0,32%). Για τους φυτικούς οργανισμούς τα αποτελέσματα της πυκνότητας των σύνθετων μοναδικών πεπτιδίων έδειξαν πως το μεγαλύτερο ποσοστό εμφανίζεται στον οργανισμού της Σόγιας (1,43%) ενώ το χαμηλότερο στον οργανισμό του αρακά (1,13%). Όπως και στην ανάλυση της πυκνότητας των μοναδικών πεπτιδίων ελαχίστου μήκους που παρατηρήθηκε πως οι μονοκύτταροι οργανισμού ακολουθούν τα ποσοστά των ζωικών οργανισμών κάτι ανάλογο παρατηρείται και στο χαρακτηριστικό της πυκνότητας των σύνθετων μοναδικών πεπτιδίων για τους μονοκύτταρους οργανισμούς. Έτσι, το βακτήριο *E.coli* εμφανίζεται με την μεγαλύτερη πυκνότητα από σύνθετα μοναδικά πεπτίδια (0,49%) για τους μονοκύτταρους οργανισμούς (Εικόνα 96, εικόνα 97).



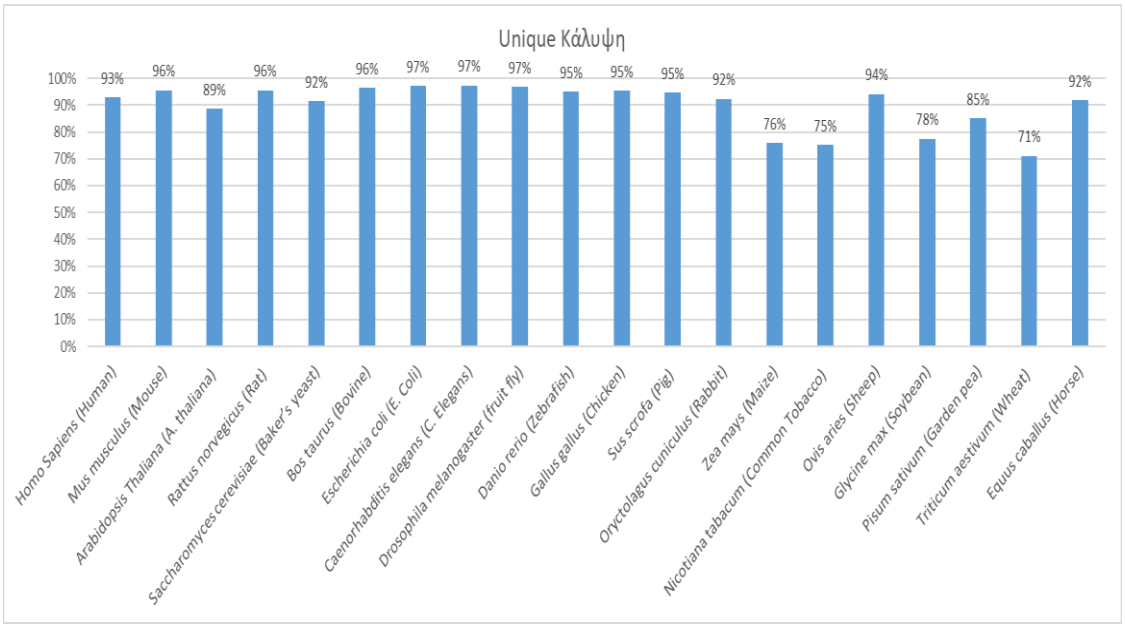
Εικόνα 96 Πυκνότητα από CmUP για τους οργανισμούς της βάσης δεδομένων των Uniquomes ταξινομημένοι βάση του μεγέθους του πρωτεώματος τους



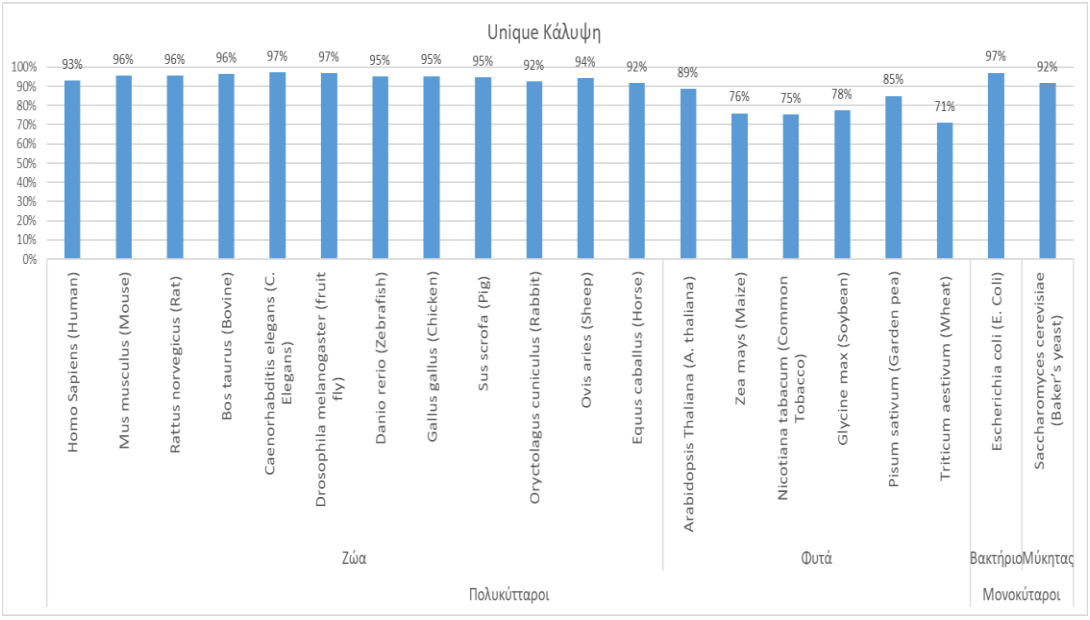
Εικόνα 97 Πυκνότητα από CmUP για τους οργανισμούς της βάσης δεδομένων των Uniquomes ταξινομημένοι βάση την κατηγορία που ανήκουν και το μέγεθος του πρωτεώματος τους

Η ανάλυση των χαρακτηριστικών των μοναδικών πεπτιδίων για τους πρότυπους οργανισμούς της βάσης δεδομένων Uniquome επεκτάθηκε και στην μοναδική κάλυψη. Τα αποτελέσματα ανέδειξαν πως όπως η πυκνότητα των μοναδικών πεπτιδίων έτσι και η μοναδική κάλυψη εξαρτάται από το είδος του υπό μελέτη οργανισμού. Πιο συγκεκριμένα παρατηρείται πως στους φυτικούς οργανισμούς η μοναδική κάλυψη είναι χαμηλότερη σε σχέση με τους ζωικούς οργανισμούς, το βακτήριο και τον μύκητα. Αναλυτικότερα στο σύνολο των οργανισμών, οι οργανισμοί Καινοραβδίτης (*C. elegans*),

Φρουτόμυγα (Fruit fly) καθώς και το βακτήριο E. coli εμφανίζονται με τη μεγαλύτερη μοναδική κάλυψη (97%) ενώ αντίθετα με την χαμηλότερη μοναδική κάλυψη (71%) εμφανίζεται ο οργανισμός του Σίτου (Wheat). Από τους φυτικούς οργανισμούς ξεχωρίζει ο οργανισμός της A.thaliana που εμφανίζεται με αρκετά μεγαλύτερη μοναδική κάλυψη (85%) σε σχέση με τους υπόλοιπους οργανισμούς του είδους της (Εικόνα 98,99).



Εικόνα 98 Μοναδική κάλυψη για τους οργανισμούς της βάσης δεδομένων των Uniquomes ταξινομημένοι βάση του μεγέθους του πρωτεώματος τους



Εικόνα 99 Μοναδική κάλυψη για τους οργανισμούς της βάσης δεδομένων των Uniquomes ταξινομημένοι βάση την κατηγορία που ανήκουν και το μέγεθος του πρωτεώματος τους

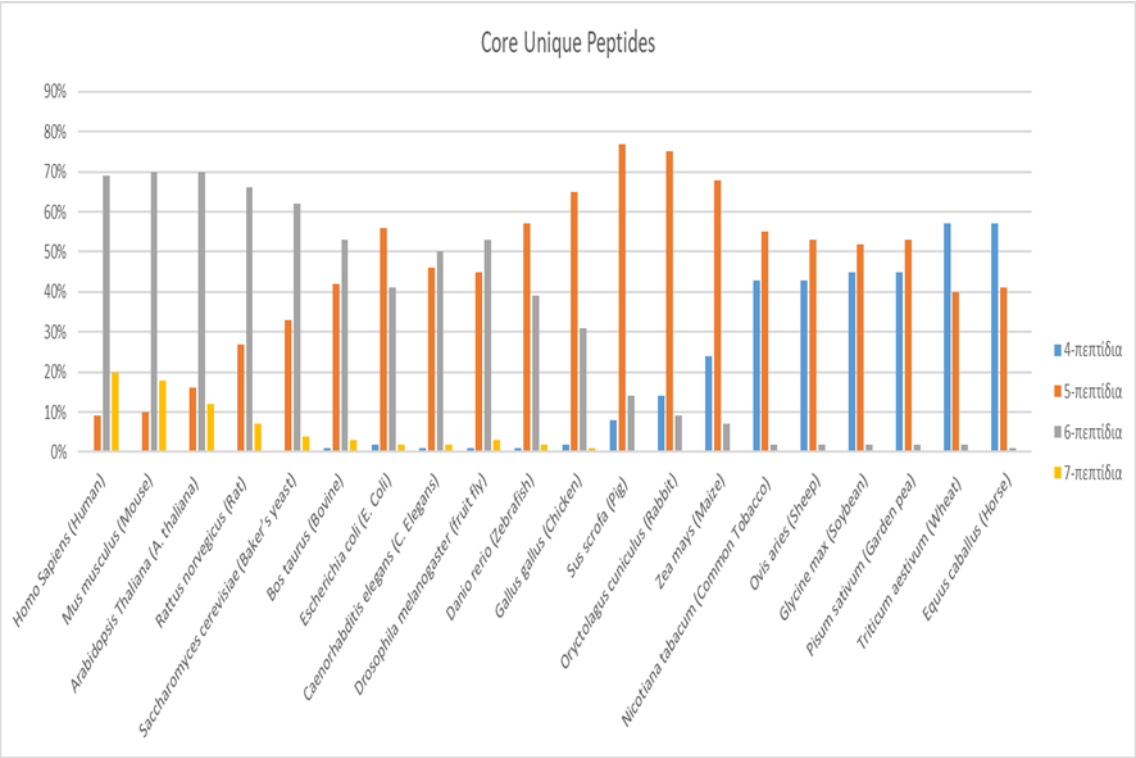
Η βάση δεδομένων των Uniquomes και το μήκος των μοναδικών πεπτιδίων

Σε συνέχεια της μελέτης και της σύγκρισης των χαρακτηριστικών στους πρότυπους οργανισμούς της βάσης δεδομένων Uniquome μελετήθηκαν τα μοναδικά πεπτίδια ως προς το μήκος τους. Αρχικά μελετήθηκαν τα μοναδικά πεπτίδια ελαχίστου μήκους ως προς το μήκος τους ανάλογα τον οργανισμό στον οποίο ανήκουν (Πίνακας 30). Για τους σκοπούς της σύγκρισης η μελέτη αυτή παρουσιάζεται σε ποσοστά για την κάθε ομάδα μήκους των πεπτιδίων ανάλογα τον υπό μελέτη οργανισμό (δηλαδή ο αριθμός μοναδικών πεπτιδίων ελαχίστου μήκους της κάθε ομάδας προς τον συνολικό αριθμό από μοναδικά πεπτίδια ελαχίστου μήκους του οργανισμού). Η ανάλυση αυτή έδειξε πως το μεγαλύτερο ποσοστό των μοναδικών πεπτιδίων ελαχίστου μήκους αποτελείται από 4, 5, 6 και 7 αμινοξέα για όλους τους πρότυπους οργανισμούς. Σε αντίθεση με τα προηγούμενα χαρακτηριστικά που συγκρίθηκαν στα Uniquome των πρότυπων οργανισμών τα συγκεκριμένα αποτελέσματα της ανάλυσης έδειξαν πως το πλήθος αυτών των ομάδων από μοναδικά πεπτίδια ελαχίστου μήκους εξαρτάται αποκλειστικά και μόνο από το μέγεθος του πρωτεώματος του του εκάστοτε οργανισμού. Πιο συγκεκριμένα όσο μεγαλύτερος σε μέγεθος πρωτεώματος είναι ο οργανισμός τόσο λιγότερα (αναλογικά) είναι τα μοναδικά 4-πεπτίδια και 5-πεπτίδια ελαχίστου μήκους που έχει ο αντίστοιχος οργανισμός. Αντιθέτως στους μικρότερος σε μέγεθος πρωτεώματος οργανισμούς τα μοναδικά 4-πεπτίδια και 5-πεπτίδια ελαχίστου μήκους είναι αυτά που συναντάμε σε μεγαλύτερο ποσοστό. Ταυτόχρονα το ακριβώς αντίθετο ισχύει για τα μοναδικά 6-πεπτίδια και 7-πεπτίδια ελαχίστου μήκους. Δηλαδή όσο μεγαλύτερος σε μέγεθος πρωτεώματος είναι ο οργανισμός τόσο περισσότερα (αναλογικά) μοναδικά 6-πεπτίδια και 7-πεπτίδια ελαχίστου μήκους έχει, ενώ αντίθετα στους μικρότερος σε μέγεθος πρωτεώματος οργανισμούς τα μοναδικά 6-πεπτίδια και 7-πεπτίδια ελαχίστου μήκους είναι αυτά που εμφανίζονται με μικρότερο ποσοστό.

Στα μοναδικά πεπτίδια ελαχίστου μήκους που αποτελούνται από 4 αμινοξέα ο οργανισμός του Αλόγου (Horse) και του Σίτου (Wheat) εμφανίζονται με τα μεγαλύτερα ποσοστά από μοναδικά πεπτίδια ελαχίστου μήκους. Ο οργανισμός του Γουρουνιού (Pig) εμφανίζεται με το μεγαλύτερο ποσοστό στην ομάδα των μοναδικών πεπτιδίων με μήκος 5 αμινοξέων. Το μεγαλύτερο ποσοστό για τα μοναδικά πεπτίδια ελαχίστου μήκους 6 αμινοξέων συναντάμε στους οργανισμούς της *A.thaliana* και του Ποντικού (Mouse). Τέλος ο οργανισμός του ανθρώπου εμφανίζεται με το μεγαλύτερο ποσοστό έναντι των υπόλοιπων πρότυπων οργανισμών στα μοναδικά πεπτίδια ελαχίστου μήκους που αποτελούνται από 7 αμινοξέα (Εικόνα 100).

Οργανισμός	4CrU Πεπτίδια	5CrU Πεπτίδια	6CrU Πεπτίδια	7CrU Πεπτίδια
Human	0%	9%	69%	20%
Mouse	0%	10%	70%	18%
A. thaliana	0%	16%	70%	12%
Rat	0%	27%	66%	7%
Baker's yeast	0%	33%	62%	4%
Bovine	1%	42%	53%	3%
E. coli	2%	56%	41%	2%
C. Elegans	1%	46%	50%	2%
Fruit fly	1%	45%	53%	3%
Zebrafish	1%	57%	39%	2%
Chicken	2%	65%	31%	1%
Pig	8%	77%	14%	0%
Rabbit	14%	75%	9%	0%
Maize	24%	68%	7%	0%
Tobacco	43%	55%	2%	0%
Sheep	43%	53%	2%	0%
Soybean	45%	52%	2%	0%
Garden pea	45%	53%	2%	0%
Wheat	57%	40%	2%	0%
Horse	57%	41%	1%	0%

Πίνακας 30 Ποσοστό εμφάνισης των CrUP που αποτελούνται από 4, 5, 6 και 7 αμινοξέα στους οργανισμούς της βάσης των Uniquomes ταξινομημένοι βάση το μέγεθος τους πρωτεώματος τους



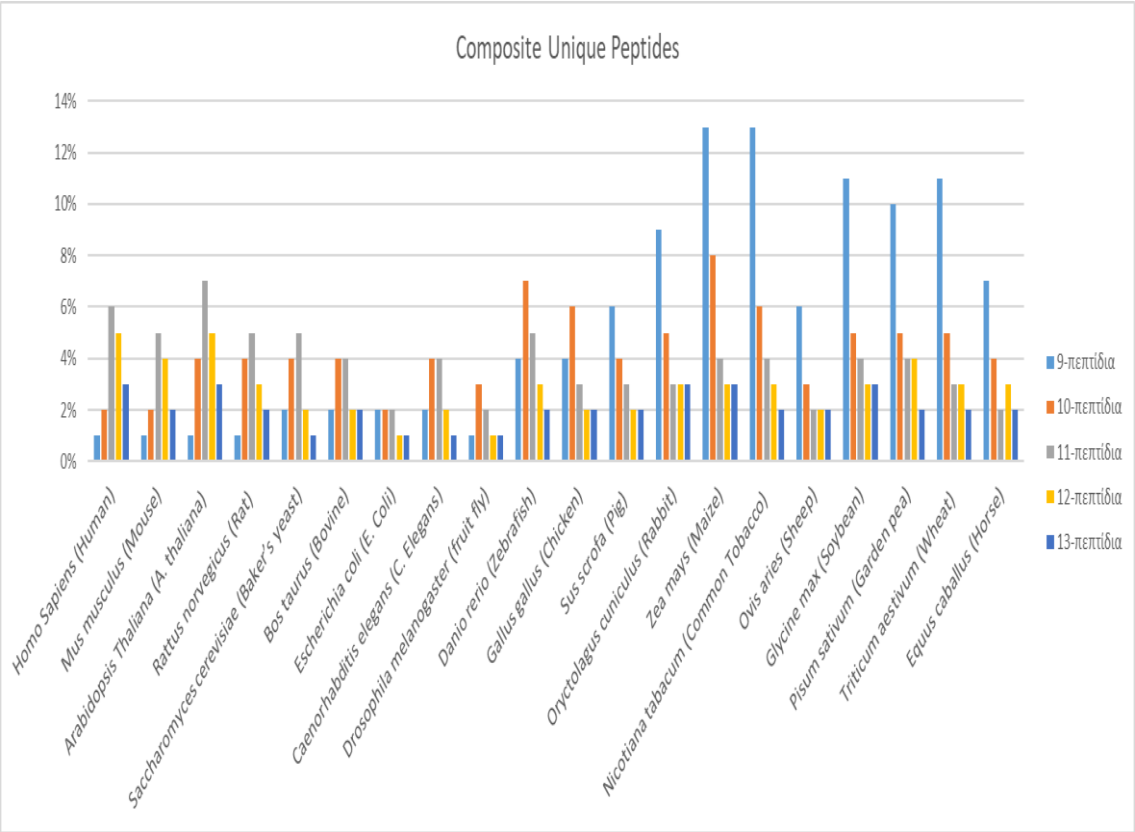
Εικόνα 100 Ποσοστό εμφάνισης μοναδικών 4,5,6 και 7-πεπτιδίων ελαχίστου μήκους στους οργανισμούς της βάσης των Uniquomes ταξινομημένοι βάση το μέγεθος τους πρωτεώματος τους

Σε συνέχεια της μελέτης του μήκους των μοναδικών πεπτιδίων στους πρότυπους οργανισμούς της βάσης δεδομένων των Uniquomes μελετήθηκε το μήκος των σύνθετων μοναδικών πεπτιδίων. Όπως και στην μελέτη του μήκους των μοναδικών πεπτιδίων ελαχίστου μήκος για τους σκοπούς της σύγκρισης έτσι και για την μελέτη των σύνθετων μοναδικών πεπτιδίων τα αποτελέσματα παρουσιάζονται σε ποσοστά για την κάθε ομάδα μήκους των πεπτιδίων ανάλογα τον υπό μελέτη οργανισμό. Σε αντίθεση με τα μοναδικά πεπτίδια ελαχίστου μήκους, στα σύνθετα μοναδικά πεπτίδια το μήκος τους δεν εξαρτάται από το μέγεθος του πρωτεώματος του οργανισμού στον οποίον ανήκουν. Πιο συγκεκριμένα στα CmUP το μεγαλύτερο ποσοστό πεπτιδίων αποτελείται από 9,10,11,12 και 13 αμινοξέα χωρίς όμως αυτό να εξαρτάται από το μέγεθος σε πρωτεΐνες του οργανισμού. Η ομάδα των σύνθετων μοναδικών πεπτιδίων με μήκος 9 αμινοξέων εμφανίζεται με τα μεγαλύτερα ποσοστά στους οργανισμούς του αλόγου (Horse) και του αρακά (garden Rea). Το άλογο είναι επίσης ο οργανισμός που εμφανίζει το μεγαλύτερο ποσοστό και για τα σύνθετα μοναδικά πεπτίδια με μήκος 10 αμινοξέων. Στα σύνθετα μοναδικά πεπτίδια με μήκος 11 αμινοξέων ο άνθρωπος είναι ο οργανισμός που εμφανίζεται με τα μεγαλύτερα ποσοστά. Τέλος για τις ομάδες πεπτιδίων που αποτελούνται από 12 και 13 αμινοξέα δεν υπάρχει κάποιος οργανισμός που να ξεχωρίζει ως προς τα ποσοστά εμφάνισης των συγκεκριμένων ομάδων καθώς όλοι οι οργανισμοί

εμφανίζονται με ποσοστά από 2% έως 5% και από 1% έως 3% αντίστοιχα (Πίνακας 31, Εικόνα 101).

Οργανισμός	9 CmU Πεπτίδια	10 CmU Πεπτίδια	11 CmU Πεπτίδια	12 CmU Πεπτίδια	13 CmU Πεπτίδια
Human	1%	4%	7%	5%	3%
Mouse	2%	4%	4%	2%	2%
A. thaliana	2%	4%	4%	2%	1%
Rat	4%	7%	5%	3%	2%
Baker's yeast	1%	3%	2%	1%	1%
Bovine	7%	4%	2%	3%	2%
E. coli	2%	2%	2%	1%	1%
C. Elegans	4%	6%	3%	2%	2%
Fruit fly	11%	5%	4%	3%	3%
Zebrafish	1%	2%	6%	5%	3%
Chicken	1%	2%	5%	4%	2%
Pig	9%	5%	3%	3%	3%
Rabbit	6%	3%	2%	2%	2%
Maize	10%	5%	4%	4%	2%
Tobacco	1%	4%	5%	3%	2%
Sheep	2%	4%	5%	2%	1%
Soybean	6%	4%	3%	2%	2%
Garden pea	13%	6%	4%	3%	2%
Wheat	11%	5%	3%	3%	2%
Horse	13%	8%	4%	3%	3%

Πίνακας 31 Ποσοστό εμφάνισης των CmUP που αποτελούνται από 9, 10, 11, 12 και 13 αμινοξέα στους οργανισμούς της βάσης δεδομένων των Uniquomes ταξινομημένοι βάση το μέγεθος τους πρωτεώματος τους



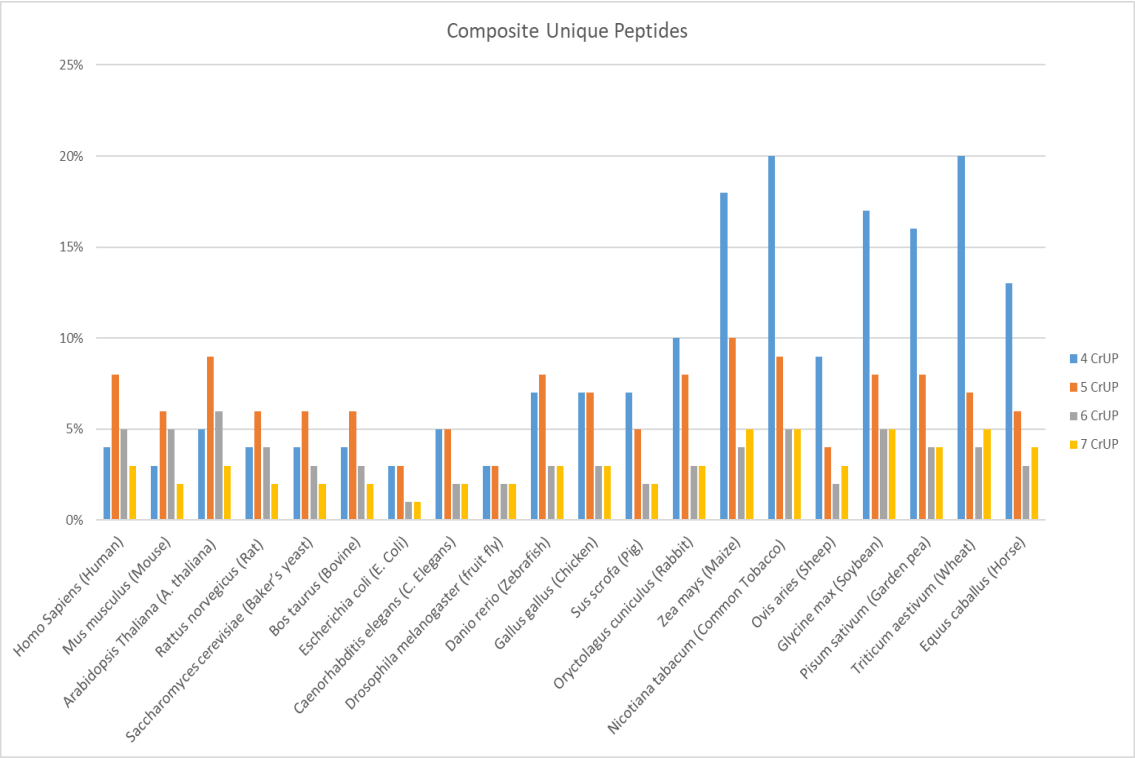
Εικόνα 101 Ποσοστό εμφάνισης σύνθετων μοναδικών 9,10,11,12 και 13-πεπτιδίων στους οργανισμούς της βάσης δεδομένων των Uniquomes ταξινομημένοι βάση το μέγεθος τους πρωτεώματος τους

Η βάση δεδομένων των Uniquomes και ο σχηματισμός σύνθετων μοναδικών πεπτιδίων

Τέλος σε συνέχεια της ανάλυσης και της σύγκρισης των χαρακτηριστικών της βάσης δεδομένων των Uniquomes μελετήθηκαν τα σύνθετα μοναδικά πεπτίδια ως προς τον αριθμό από μοναδικά πεπτίδια ελαχίστου μήκος που αποτελούνται. Τα αποτελέσματα της ανάλυσης έδειξαν ότι το πλήθος των μοναδικών πεπτιδίων ελαχίστου μήκους που συντελούν για τον σχηματισμό των σύνθετων μοναδικών πεπτιδίων έχει σχέση με το μέγεθος του πρωτεώματος στον οργανισμό που ανήκει. Πιο συγκεκριμένα σε μικρότερους σε μέγεθος πρωτεώματος οργανισμούς το μεγαλύτερο ποσοστό των σύνθετων μοναδικών πεπτιδίων αποτελείται από 4 μοναδικά πεπτίδια ελαχίστου μήκους. Στους μεσαίους και μεγάλους βάση πρωτεώματος οργανισμούς όμως παρατηρείται πως υπάρχει μια σχεδόν κανονική κατανομή ως προς το πλήθος των μοναδικών πεπτιδίων ελαχίστου μήκους που συμβάλουν για την σύνθεση των σύνθετων μοναδικών πεπτιδίων (Πίνακας 32, Εικόνα 102).

Οργανισμός	CmU από 4 CrU Πεπτίδια	CmU από 5 CrU Πεπτίδια	CmU από 6 CrU Πεπτίδια	CmU από 7 CrU Πεπτίδια
Human	4%	8%	5%	3%
Mouse	3%	6%	5%	2%
A. thaliana	5%	9%	6%	3%
Rat	4%	6%	4%	2%
Baker's yeast	4%	6%	3%	2%
Bovine	4%	6%	3%	2%
E. coli	3%	3%	1%	1%
C. Elegans	5%	5%	2%	2%
Fruit fly	3%	3%	2%	2%
Zebrafish	7%	8%	3%	3%
Chicken	7%	7%	3%	3%
Pig	7%	5%	2%	2%
Rabbit	10%	8%	3%	3%
Maize	18%	10%	4%	5%
Tobacco	20%	9%	5%	5%
Sheep	9%	4%	2%	3%
Soybean	17%	8%	5%	5%
Garden pea	16%	8%	4%	4%
Wheat	20%	7%	4%	5%
Horse	13%	6%	3%	4%

Πίνακας 32 Ποσοστό εμφάνισης των CmUP που αποτελούνται από 4, 5, 6 και 7 CrUP στους οργανισμούς της βάσης δεδομένων των Uniquomes ταξινομημένοι βάση το μέγεθος τους πρωτεώματος τους



Εικόνα 102 Ποσοστό εμφάνισης σύνθετων μοναδικών πεπτιδίων που αποτελούνται από 4, 5, 6 και 7 μοναδικά πεπτίδια ελαχίστου μήκους στους οργανισμούς της βάσης δεδομένων των Uniquomes ταξινομημένοι βάση το μέγεθος τους πρωτεώματος τους

5. Συζήτηση

Η Πρωτεωμική είναι ένα σύνολο πολύπλοκων μεθόδων και τεχνολογιών που αποσκοπεί στην ταυτοποίηση, καταγραφή και μελέτη του ολικού πρωτεϊνικού περιεχομένου ενός βιολογικού υλικού. Περιλαμβάνει το διαχωρισμό των πρωτεϊνών, την ανάλυση με φασματομετρία μάζας [83], την ταυτοποίησή τους με τη χρήση εργαλείων βιοπληροφορικής και, τέλος, τη συστηματική εισαγωγή των αποτελεσμάτων σε βάσεις δεδομένων. Οι πλέον εύχρηστες μέθοδοι για την ταυτοποίηση των πρωτεϊνών είναι αυτές που αξιοποιούν το πεπτιδικό αποτύπωμά τους (peptide finger-print) [84] και αναλύουν την αμινοξική αλληλουχία των πεπτιδίων τους. Για την ασφαλή ταυτοποίηση μίας πρωτεΐνης, με πιθανότητα σφάλματος «P» μικρότερη του 0,05, απαιτείται η ανάλυση τουλάχιστον δύο πεπτιδίων ανά πρωτεΐνη. Η παραπάνω γνώση σε συνδυασμό με το ότι πολλά από τα πεπτίδια που ταυτοποιούνται από το φασματογράφο μάζας δεν οδηγούν τελικά σε ασφαλή χαρακτηρισμό μίας πρωτεΐνης και απορρίπτονται κατά τη βιοπληροφορική επεξεργασία οδήγησε στην ανάγκη ανάπτυξης νέων προσεγγίσεων για την ταυτοποίηση των πρωτεϊνών ενός υπό μελέτη κυτταρικού πληθυσμού. Οι μέχρι τώρα προσεγγίσεις βασίζονται στην θεωρία ότι η αμινοξική αλληλουχία κάθε πρωτεΐνης περιλαμβάνει τουλάχιστον δύο πεπτίδια που είναι μοναδικά (Unique) ως προς την πρωτεΐνη, με αποτέλεσμα να τη χαρακτηρίζει διαφορικά και μονοσήμαντα [67].

Σχετικές μελέτες για την μοναδικότητα των πεπτιδίων στις πρωτεΐνες βασίστηκαν στην ανάλυση του μοριακού βάρους των πεπτιδίων (peptide mass finger printing) [67-69]. Τέτοιου είδους προσεγγίσεις όμως είχαν να αντιμετωπίσουν το πρόβλημα με τα πεπτίδια που παρότι είχαν το ίδιο μοριακό βάρος, η αμινοξική τους αλληλουχία ήταν διαφορετική. Λόγω των ισοβαρών πεπτιδίων, η ανάλυση για την ταυτοποίηση των πρωτεϊνών χρησιμοποιώντας το μοριακό βάρος των πεπτιδίων οδηγήθηκε σύντομα στην αδυναμία επιτυχούς ταυτοποίησης των πρωτεϊνών. Συγκεκριμένα τα λανθασμένα αποτελέσματα οφείλονταν στην αδυναμία να ταυτοποιηθεί το σωστό μοναδικό πεπτίδιο στις περιπτώσεις που υπήρχαν περισσότερα του ενός πεπτιδίου με το ίδιο μοριακό βάρος με συνέπεια την ταυτοποίηση πρωτεϊνών με μεγάλη πιθανότητα λάθους.

Για να υπερκεράσουμε τα μειονεκτήματα της προηγούμενης προσέγγισης θεωρήθηκε ότι η ανάλυση των πρωτεϊνών με βάση την γραμμική αμινοξική αλληλουχία των πεπτιδίων τους θα προσέφερε μεγαλύτερη ακρίβεια και αυξημένη πιθανότητα στην ταυτοποίηση των πρωτεϊνών [85]. Με βάση την παραπάνω υπόθεση σχεδιάστηκε η μεθοδολογική, εμπεριστατωμένη και λεπτομερής ανάλυση του πρωτεώματος με βάση την αμινοξική αλληλουχία των πεπτιδίων των πρωτεϊνών. Με βάση αυτή τη προσέγγιση η κάθε πρωτεΐνη αναλύθηκε σε πεπτίδια μήκους από 4 έως 100 αμινοξέων. Από την

ανάλυση διαπιστώθηκε ότι σε κάθε πρωτεΐνη περιέχονταν μοναδικής αλληλουχίας πεπτίδια χαρακτηριστικά για την εν λόγω πρωτεΐνη. Έτσι σαν αποτέλεσμα αυτής της προσέγγισης, αναδείχθηκαν δύο νέες οντότητες μοναδικών πεπτιδίων και εισάχθηκαν οι όροι μοναδικό πεπτίδιο (unique peptide), μοναδικό πεπτίδιο ελάχιστου μήκους (core unique peptide) και σύνθετο μοναδικό πεπτίδιο (composite unique peptide). Τα μοναδικά πεπτίδια ελαχίστου μήκους (CrUPs) είναι τα ελάχιστου μήκους πεπτίδια των οποίων η αμινοξική της αλληλουχία εμφανίζεται σε μία μόνο πρωτεΐνη ενώ τα σύνθετα μοναδικά πεπτίδια (CmUPs) είναι τα πεπτίδια τα οποία συνθέτονται από την ένωση δύο ή περισσότερων μοναδικών πεπτιδίων ελαχίστου μήκους, όταν το ένα επικαλύπτει το άλλο. Τέλος, εισήχθη για πρώτη φορά ο όρος του Uniquome που περιλαμβάνει το σύνολο των μοναδικών πεπτιδίων (ελάχιστου μήκους και σύνθετων πεπτιδίων), του προς μελέτη οργανισμού. Συγκεκριμένα για τον άνθρωπο το human Uniquome περιλαμβάνει το σύνολο των μοναδικών πεπτιδίων CrUP και CmUP του ανθρώπινου πρωτεώματος. Η μεθοδολογία της δημιουργίας του και η ανάλυση του αποτελούν το αντικείμενο της παρούσας διατριβής.

Η μελέτη του human Uniquome, αποκάλυψε μια σειρά ιδιαίτερων χαρακτηριστικών και ιδιοτήτων τόσο των CrUPs όσο και των CmUPs. Συγκεκριμένα η ανάλυση του μήκους των μοναδικών πεπτιδίων έδειξε ότι η ομάδα με το μεγαλύτερο αριθμό πεπτιδίων είναι τα μοναδικά πεπτίδια ελαχίστου μήκους (CrUPs) με μέγεθος 6 αμινοξέων είναι (5.015.527), ενώ αντίστοιχα για τα σύνθετα μοναδικά πεπτίδια (CmUPs) η ανάλυση έδειξε ότι τα πεπτίδια μεγέθους 11 αμινοξέων αποτελούν την πολυπληθέστερη ομάδα (4.676 πεπτίδια) (Εικόνα 39,40,44 και Πίνακας 5). Αναφορικά με τον αριθμό των μοναδικών πεπτίδια ελαχίστου μήκους τα οποία συμμετέχουν για στο σχηματισμό των σύνθετων μοναδικών πεπτιδίων η ανάλυση ανέδειξε πως υπάρχουν CmUPs που συνθέτονται από 2 μέχρι και 25.055 CrUPs, Τα CmUPs που συνθέτονται από 5 CrUPs αποτελούν τον μεγαλύτερο αριθμό πεπτιδίων (6.103 CmUPs) (Πίνακας 6 και Εικόνα 48). Ένα άλλο χαρακτηριστικό των μοναδικών πεπτιδίων που μελετήθηκε εκτενώς είναι η θέση εμφάνισή της μέσα της πρωτεΐνες. Τα αποτελέσματα έδειξαν πως τα CrUPs ανιχνεύονται σε όλο το μήκος της αλληλουχίας της πρωτεΐνης, εύρημα που υποδηλώνει ότι δεν υπάρχει προτιμητέα θέση/περιοχή ως της την εμφάνιση της CrUP μέσα στην αμινοξική αλληλουχία της πρωτεΐνης, ενώ αντίθετα διαπιστώθηκε ότι τα CmUPs ανιχνεύονται της αρχικές θέσεις των πρωτεϊνών (Εικόνα 41 και 45). Η ιδιαιτερότητα αυτή των CmUPs ως προς την εμφάνισή τους στις αρχικές θέσεις των πρωτεϊνών οφείλεται στο ότι η πλειοψηφία των πρωτεϊνών έχουν μεγάλο αριθμό από επικαλυπτόμενα CrUPs κατά μήκος ολόκληρης της αλληλουχίας της, με το πρώτο να ξεκινάει στις αρχικές θέσεις τής εκάστοτε πρωτεΐνης. Συνεπώς οι πρωτεΐνες αυτές

αποτελούνται από τουλάχιστον ένα CmUP που η θέση εμφάνισής του εντοπίζεται στις αρχικές θέσεις τους. Ένα τέτοιο παράδειγμα είναι η ανάλυση της ανθρώπινης πρωτεΐνης MYC ως της τα μοναδικά της πεπτίδια (Εικόνα 46).

Η πρωτεΐνη MYC είναι μία ογκοπρωτεΐνη [86] η οποία αποτελείται από 439 αμινοξέα. Σύμφωνα με την ανάλυσή με βάση τη μεθοδολογία που αναπτύχθηκε περιλαμβάνει 314 CrUPs τα οποία σχηματίζουν 3 CmUPs. Η κατανομή των μοναδικών της πεπτιδίων φαίνεται στην Εικόνα 46 με τα CrUPs της πρωτεΐνης MYC να είναι ομοιόμορφα κατανεμημένα στην αμινοξική αλληλουχία της πρωτεΐνης ξεκινώντας από τις αρχικές θέσεις της πρωτεΐνης. Αυτό έχει ως συνέπεια το πρώτο από τα τρία CmUPs να ξεκινάει στις αρχικές θέσεις της πρωτεΐνης, ενώ το αμέσως επόμενο CmUP ξεκινά από την θέση 60 και το τρίτο CmUP να ξεκινά από την θέση 136 έως το τέλος της πρωτεΐνης. Η πλειοψηφία των πρωτεϊνών του ανθρώπου, ακολουθεί περίπου το ίδιο μοτίβο ως της την θέση εμφάνισης των μοναδικών πεπτιδίων στην αμινοξική τους αλληλουχία με αποτέλεσμα η συντριπτική πλειοψηφία των CmUPs (70%) να εμφανίζεται στις αρχικές θέσεις των πρωτεϊνών.

Ένα από τα ενδιαφέροντα χαρακτηριστικά που ανέδειξε η ανάλυση του human Uniquome είναι ότι 148 πρωτεΐνες, από τις 20.430 reviewed πρωτεΐνες που αναλύθηκαν, (0,72%) δεν περιέχουν κανένα μοναδικό πεπτίδιο ελαχίστου μήκους (μεγέθους από 4 έως 100 αμινοξέα) καθώς και κανένα σύνθετο μοναδικό πεπτίδιο. Οι 148 αυτές πρωτεΐνες διαπιστώθηκε ότι ανήκουν σε 51 οικογένειες πρωτεϊνών, ενώ 12 πρωτεΐνες (9%) από αυτές τις 148 πρωτεΐνες ανήκουν σε τρείς οικογένειες πρωτεϊνών.

Η οικογένεια G-protein coupled receptor 1 (GPCR) είναι η οικογένεια με τον μεγαλύτερο αριθμό πρωτεϊνών στο ανθρώπινο πρωτέωμα [87]. Η οικογένεια GPCR αποτελείται από 724 πρωτεΐνες οι οποίες αποτελούνται από 137.815 μοναδικά πεπτίδια ελαχίστου μήκους και 3.081 σύνθετα μοναδικά πεπτίδια (Πίνακας 8). Από τις 724 πρωτεΐνες της συγκεκριμένης οικογένειας υπάρχουν μόλις 5 πρωτεΐνες (εικόνα 53) οι οποίες δεν περιλαμβάνουν κανένα μοναδικό πεπτίδιο οι οποίες είναι:

- OPSG Medium-wave-sensitive opsin 1 (P04001)
- OPSG2 Medium-wave-sensitive opsin 2 (P0DN77)
- OPSG3 Medium-wave-sensitive opsin 3 (P0DN78)
- Neuropeptide Y receptor type 4 (P50391) as
- Neuropeptide Y receptor type 4-2 (P0DQD5)

Ανάλυση των παραπάνω πρωτεϊνών μέσω αμινοξικής ευθυγράμμισης έδειξε ότι οι πρωτεΐνες P04001, P0DN77 και P0DN78 εμφανίζουν ταυτόσημη ομολογία, όπως επίσης και οι πρωτεΐνες P50391 και P0DQD5 (Εικόνα 36). Αποτέλεσμα της ταυτοσημίας είναι η

απουσία μοναδικών πεπτιδίων στις πρωτεΐνες αυτές. Οι υπόλοιπες δύο οικογένειες πρωτεϊνών οι οποίες περιλαμβάνουν μέλη τα οποία δεν έχουν μοναδικά πεπτίδια είναι οι οικογένειες Peptidase C19 και Peptidase S1 (Εικόνα 53, Πίνακας 8). Η οικογένεια των Peptidase C19 περιλαμβάνει 5 πρωτεΐνες που δεν έχουν μοναδικά πεπτίδια, ενώ η οικογένεια των Peptidase S1 περιλαμβάνει 2 πρωτεΐνες. Οι οικογένειες αυτές ανήκουν στην υπεροικογένεια των πεπτιδασών, των οποίων ο ρόλος είναι να αποικοδομούν πρωτεΐνες σε πεπτίδια με σκοπό την απενεργοποίηση τους [88]. Η οικογένεια των πρωτεϊνών Peptidase C19 αποτελείται από 76 πρωτεΐνες, 5 εκ των οποίων δεν περιλαμβάνουν κανένα μοναδικό πεπτίδιο:

- U17LK_ human Ubiquitin carboxyl-terminal hydrolase 17-like protein 20 (D6RJB6)
- UL17C_human Ubiquitin carboxyl-terminal hydrolase 17-like protein 12 (C9JPN9)
- U17LO_ human Ubiquitin carboxyl-terminal hydrolase 17-like protein 24 (Q0WX57)
- U17LL_ human Ubiquitin carboxyl-terminal hydrolase 17-like protein 21 (D6R901)
- U17LB_ human Ubiquitin carboxyl-terminal hydrolase 17-like protein 11 (C9JVI0)

ενώ η οικογένεια Peptidase S1 αποτελείται από 119 πρωτεΐνες με μόλις 2 από αυτές να μην περιλαμβάνουν κανένα μοναδικό πεπτίδιο:

- TRYB1_ human Tryptase alpha/beta 1 (Q15661)
- TRYB2_ human Tryptase beta-2 (P20231)

Έπειτα από αμινοξική ευθυγράμμιση διαπιστώθηκε, ότι οι 5 πρωτεΐνες της ομάδας Peptidase C19 έχουν ομολογία μεταξύ τους μεγαλύτερη από 99%, ενώ οι 2 πρωτεΐνες της ομάδας Peptidase S1 έχουν ομολογία 100%. (Εικόνα 37, Εικόνα 38). Η ανάλυση υποδεικνύει ότι οι πρωτεΐνες που δεν περιλαμβάνουν μοναδικά πεπτίδια είναι πρωτεΐνες που ανήκουν στην ίδια οικογένεια ενώ η αμινοξική τους αλληλουχία εμφανίζεται πανομοιότυπα και σε κάποια άλλη πρωτεΐνη της οικογένειας, δηλαδή πρωτεΐνες που ουσιαστικά είναι ισομορφές με ποσοστό ομολογίας μεγαλύτερο του 99%. Γίνεται λοιπόν φανερό από τα αποτελέσματα, ότι μέσω του Uniquome, reviewed πρωτεΐνες (θεωρημένες) με 100% ταυτόσημη ομολογία μπορούν να θεωρηθούν ως η ίδια πρωτεΐνη. Το εύρημα αυτό μπορεί να οδηγήσει σε αναθεώρηση της σχετικής βάσης δεδομένων, καθώς και σε αναθεώρηση του ρόλου των πρωτεϊνών αυτών στα βιολογικά συστήματα.

Κατανομή των 148 αυτών πρωτεϊνών στα χρωμοσώματα του Ανθρώπου έδειξε ότι 51 από αυτές (35%) ανήκουν στα χρωμοσώματα 13 (6 πρωτεΐνες 1,92% των πρωτεϊνών του χρωμοσώματος), 21 (10 πρωτεΐνες 4,52% των πρωτεϊνών του

χρωμοσώματος), X (27 πρωτεΐνες 3,40% των πρωτεϊνών του χρωμοσώματος) και Y(6 πρωτεΐνες 16,22% των πρωτεϊνών του χρωμοσώματος) (Εικόνα 49,50 και Πίνακας 7). Η βιολογική σημασία του παραπάνω ευρήματος γίνεται κατανοητή παρακάτω συνδυαζόμενη με περεταίρω ευρήματα της ανάλυσης.

Ανάλυση των μοναδικών πεπτιδίων (CrUPs και CmUPs) ανά οικογένεια πρωτεϊνών στο human Uniquome έδειξε ότι υπάρχουν 3 οικογένειες (G-protein coupled receptor 1 family, Krueppel C2H2-type zinc-finger protein family, Protein kinase superfamily) οι οποίες έχουν τον μεγαλύτερο αριθμό μοναδικών πεπτιδίων ελαχίστου μήκους (137.815, 170.255, 265.716 αντίστοιχα) και σύνθετων μοναδικών πεπτιδίων (3.081, 2.952, 3.553, αντίστοιχα) γεγονός το οποίο οφείλεται στο μεγάλο αριθμό των πρωτεϊνών που περιλαμβάνουν οι συγκεκριμένες ομάδες (724, 544, 490 πρωτεΐνες αντίστοιχα) (εικόνα 54 και 59, Πίνακας 8). Αντικειμενικότερη ανάλυση της σχέσης των CrUP και των CmUP ανά οικογένεια πρωτεϊνών επιτυγχάνεται μέσω της ανάλυσης της πυκνότητας των μοναδικών πεπτιδίων ανά οικογένεια. Διαπιστώθηκε ότι η πλειοψηφία των οικογενειών των πρωτεϊνών εμφανίζουν στατιστικά παρόμοια πυκνότητα CrUPs (Εικόνα 51). Η οικογένεια των πρωτεϊνών MHC class1, εμφανίζεται με την μικρότερη πυκνότητα από μοναδικά πεπτίδια ελαχίστου μήκους (10%) καθώς και με την μικρότερη κάλυψη από μοναδικά πεπτίδια (37%) (Εικόνα 55,61 και Πίνακας 8).

Οι πρωτεΐνες της οικογένειας των πρωτεϊνών Major histocompatibility complex class I (MHC I) παίζουν καθοριστικό ρόλο στην ανάπτυξη της προσαρμοστικής ανοσολογικής απάντησης [89]. Η οικογένεια MHC I αποτελείται από 85 πρωτεΐνες οι οποίες περιλαμβάνουν έστω και ένα μοναδικό πεπτίδιο. Πιο συγκεκριμένα, στις πρωτεΐνες της ομάδας MHC I ανιχνεύθηκαν 3.012 CrUPs και 291 CmUPs, αριθμοί που οδηγούν σε χαμηλά ποσοστά μοναδικής κάλυψης καθώς και χαμηλή πυκνότητα μοναδικών πεπτιδίων (10% και 37%, αντίστοιχα). Περαιτέρω ανάλυση που πραγματοποιήθηκε σε τρείς ομάδες πρωτεϊνών που ανήκουν στην οικογένεια MHC I έδειξε ότι οι πρωτεΐνες που ανήκουν σε αυτές τις ομάδες παρουσιάζουν αμινοξικές αλληλουχίες με μεγάλα ποσοστό ομοιότητας.

Αναλυτικότερα, η ομάδα των πρωτεϊνών που περιλαμβάνει της πρωτεΐνες:

- MICB_ human MHC class I polypeptide-related sequence B (Q29980)
- MICA_ human MHC class I polypeptide-related sequence A (Q29983)

Έπειτα από αμινοξική ευθυγράμμιση των ακολουθιών τους (Εικόνα 56) παρατηρήθηκε ότι εμφανίζουν 82,51% ομολογίας. Αντίστοιχα η ομάδα πρωτεϊνών :

- ULBP1_ human UL16-binding protein 1 (Q9BZM6)

- ULBP2_ human UL16-binding protein 2 (Q9BZM5)
- ULBP3_ human UL16-binding protein 3 (Q9BZM4)
- ULBP5_ human UL-16 binding protein 5 (Q6H3X3)
- ULBP6_ human UL16-binding protein 6 (Q5VY80)

εμφανίζεται με ποσοστό ομοιότητας από 54,32% έως 94,31% ανά ζεύγη πρωτεϊνών έπειτα από ευθυγράμμιση που πραγματοποιήθηκε στις αμινοξικές αλληλουχίες των πρωτεϊνών (Εικόνα 57). Τέλος η ομάδα πρωτεϊνών:

- HLAA_ human HLA class I histocompatibility antigen, A alpha chain (P04439)
- HLAE_ human HLA class I histocompatibility antigen, alpha chain E (P13747)
- HLAF_ human HLA class I histocompatibility antigen, alpha chain F (P30511)
- HLAG_ human HLA class I histocompatibility antigen, alpha chain G (P17693)
- HLAH_ human Putative HLA class I histocompatibility antigen, alpha chain H (P01893)

έπειτα από αμινοξική ευθυγράμμιση των ακολουθιών τους (Εικόνα 58) διαπιστώθηκε ότι έχουν ποσοστά ομολογίας από 72,54% έως 85,64% ανά ζεύγη πρωτεϊνών. Από τις παραπάνω αναλύσεις στις ομάδες των πρωτεϊνών της οικογένειας MHC I που πραγματοποιήθηκαν, υποδεικνύεται ότι ό λόγος για τον οποίο η συγκεκριμένη οικογένεια πρωτεϊνών εμφανίζεται με χαμηλά ποσοστά πυκνότητας από CrUPs, CmUPs καθώς και μοναδικής κάλυψης οφείλεται στο ότι τα μέλη της εμφανίζουν μεγάλα ποσοστά ομολογίας στις μεταξύ τους αμινοξικές αλληλουχίες.

Τα αποτελέσματα των μοναδικών πεπτιδίων της οικογένεια πρωτεϊνών RAS αναλύθηκαν εκτενώς λόγω των ιδιαίτερων χαρακτηριστικών της. Οι πρωτεΐνες RAS ανήκουν στην υπεροικογένεια των μικρών GTP-άσων και εναλλάσσονται μεταξύ μιας ανενεργής (που προσδένει το GDP) και μιας ενεργής (που προσδένει το GTP) μορφής. Οι μεταλλάξεις του γονιδίου οδηγούν συνήθως σε πρωτεΐνη που βρίσκεται συνεχώς στην ενεργή μορφή και ενεργοποιεί σηματοδοτικά μονοπάτια που ελέγχουν τον πολλαπλασιασμό, την απόπτωση και τις κυτταρικές λειτουργίες [90,91]. Η οικογένεια των πρωτεϊνών Ras περιλαμβάνει 3 πρωτεΐνες, τις K-Ras, N-Ras και την H-Ras, οι οποίες εμφανίζονται στο ανθρώπινο πρωτέωμα με τον ίδιο αριθμό αμινοξέων (189 αμινοξέα). Οι πρωτεΐνες αυτές αποτελούνται από μικρό αριθμό CrUPs (41 οι K-RAS, H-RAS και 45 η N-RAS) καθώς και μικρό αριθμό από CmUPs (6 η K-RAS, 2 η N-RAS και 4 η H-RAS) με αποτέλεσμα να εμφανίζονται με χαμηλό ποσοστό πυκνότητας μοναδικών πεπτιδίων και μοναδικής κάλυψης (Πίνακας 9). Ενδελεχής ανάλυση της οικογένειας ως προς τα χαρακτηριστικά των μοναδικών της πεπτιδίων έδειξε ότι η πλειονότητα των CrUPs και των CmUPs εμφανίζεται στο 2^{ο} μισό των πρωτεϊνών (Εικόνα 63). Έπειτα από αμινοξική

ευθυγράμμιση των ακολουθιών των τριών πρωτεϊνών παρατηρήθηκε πως στο 1ο μισό τους οι πρωτεΐνες αυτές έχουν πανομοιότυπη αμινοξική αλληλουχία με αποτέλεσμα να μην δύναται να σχηματίσουν μοναδικά πεπτίδια (Εικόνα 62).

Περαιτέρω, πραγματοποιήθηκε λεπτομερής ανάλυση της αμινοξικής σύστασης των πεπτιδίων στις κοινές θέσης εμφάνισης των CrUPs στις τρείς πρωτεΐνες της οικογένειας RAS (Πίνακας 10). Η ανάλυση ανέδειξε πως τα πεπτίδια που εμφανίζονται στις θέσεις 64 και 95 της πρωτεΐνες έχουν καθοριστικό ρόλο στον χαρακτηρισμό της μοναδικότητας αυτών των πρωτεϊνών. Αναλυτικότερα στην θέση 95 εμφανίζεται το μοτίβο XYREQI όπου ανάλογα με την πρωτεΐνη το X είναι H στην K-RAS, L στην N-RAS και Q στην H-RAS, ενώ για κάθε μία από αυτές τις τρείς πρωτεΐνες το σχηματιζόμενο πεπτίδιο είναι CrUP. Το μοτίβο XYREQI που συναντάμε στην συγκεκριμένη θέση χαρακτηρίζει μοναδικά την κάθε πρωτεΐνη ανάλογα με το αμινοξύ της πρώτης θέσης του (H,L ή Q). Είναι ενδιαφέρον ότι στην θέση 64 και στις 3 πρωτεΐνες της οικογένειας RAS συναντάτε το 6-πεπτίδιο EEYSAM. Το πεπτίδιο αυτό δεν εμφανίζεται σε καμία άλλη ανθρώπινή πρωτεΐνη πέραν των τριών πρωτεϊνών της οικογένειας RAS (K-RAS, N-RAS, H-RAS). Έτσι λοιπόν το πεπτίδιο αυτό φαίνεται ότι μπορεί να χαρακτηρίσει μοναδικά την οικογένεια RAS και να χαρακτηριστεί ως Family Unique Peptide για αυτή την οικογένεια.

Φαίνεται λοιπόν από τα παραπάνω ότι μπορεί να υπάρξει μία υποκατηγορία μοναδικών πεπτιδίων τα «Family Unique Peptides», η εμφάνιση των οποίων σε μία πρωτεΐνη μπορεί να ταξινομήσει την εν λόγω πρωτεΐνη ως μέλος της συγκεκριμένης ομάδας (οικογένειας), υποδηλώνοντας ότι μέσω της προσέγγισης των unique πεπτιδίων, «ορφανές» πρωτεΐνες στο ανθρώπινο πρωτέωμα, δηλαδή πρωτεΐνες που μέχρι σήμερα δεν έχει καταστεί δυνατό να ταξινομηθούν σε κάποια οικογένεια, μέσω των «Family Unique Peptides» μπορούν να καταταγούν σε κάποια από τις ήδη υπάρχουσες οικογένειες πρωτεϊνών.

Για την κατανόηση του γενικότερου ρόλου των μοναδικών πεπτιδίων (CrUPs και CmUPs), διερευνήθηκε η κατανομή τους στα χρωμοσώματα του Ανθρώπου. Η μελέτη της πυκνότητας των μοναδικών πεπτιδίων ελαχίστου μήκους στα χρωμοσώματα έδειξε ότι το χρωμόσωμα 19 και Y εμφανίζουν την χαμηλότερη πυκνότητα (59% και 19% αντίστοιχα) σε σχέση με τα υπόλοιπα χρωμοσώματα, ενώ αντίθετα το Μιτοχονδριακό χρωμόσωμα εμφανίζει αρκετά υψηλή πυκνότητα CrUPs (74%) (Εικόνα 51, Πίνακας 7). Τέλος, όσον αφορά την μοναδική κάλυψη των μοναδικών πεπτιδίων και την κατανομή της στα χρωμοσώματα παρατηρήθηκε παρόμοια κατανομή. Συγκεκριμένα, διαπιστώθηκε ότι το χρωμόσωμα Y εμφάνισε χαμηλό ποσοστό μοναδικής κάλυψης (35%) σε σχέση με τα υπόλοιπα χρωμοσώματα ενώ αντίθετα το Μιτοχονδριακό χρωμόσωμα εμφάνισε πλήρη μοναδική κάλυψη (ποσοστό 100%) (Εικόνα 52, Πίνακας 7).

Από τα αποτελέσματα διαπιστώνεται ότι από το σύνολο των ανθρώπινων χρωμοσωμάτων, τα χρωμοσώματα 13, 19, 21, X, Y και το Μιτοχονδριακό χρωμόσωμα εμφανίζουν ιδιαίτερα χαρακτηριστικά όσον αφορά το Uniquome. Συγκεκριμένα, τα χρωμοσώματα 19, Y εμφανίζουν τα χαμηλότερα ποσοστά μοναδικών πεπτιδίων αναλογικά με τα υπόλοιπα χρωμοσώματα ενώ το Μιτοχονδριακό χρωμόσωμα έχει 100% κάλυψη. Αντίθετα τα χρωμοσώματα 13, 21, X και Y περιλαμβάνουν μεγάλο αριθμό από πρωτεΐνες που δεν περιλαμβάνουν μοναδικά πεπτίδια. Τα μιτοχόνδρια είναι από τα πιο αρχαία και καλοδιατηρημένα ενδομεβρανικά συστήματα στα ευκαριωτικά κύτταρα [92] γεγονός που φαίνεται να σχετίζεται με τη μέγιστη μοναδικότητα των πεπτιδίων του. Αναφορικά με τα χρωμοσώματα που εντοπίστηκαν με χαμηλά χαρακτηριστικά μοναδικότητας καθώς και με μεγάλό ποσοστό από πρωτεΐνες που δεν εμφανίζουν μοναδικά πεπτίδια, επισημαίνεται πως κάποια από αυτά είναι υπεύθυνα για τις χρωμοσωμικές ανωμαλίες που έχουν καταγραφεί στον άνθρωπο καθώς και για μεταθέσεις χρωμοσωμάτων που οδηγούν σε κακοήθεις εξαλλαγές. Αναλυτικότερα το χρωμόσωμα 21 είναι υπεύθυνο για το σύνδρομο Down γνωστό ως τρισωμία 21 [93]. Τα φυλετικά χρωμοσώματα (X,Y) ευθύνονται τόσο για το σύνδρομο Turner (X0) [94], όσο και για τις τρισωμίες που αφορούν το Klinefelter σύνδρομο (XXY) [95], καθώς και το σύνδρομο Jacobs (XYY) [96]. Τέλος το χρωμόσωμα 13 ενοχοποιείται για το σύνδρομο Patau γνωστό με το όνομα τρισωμία 13 [97]. Το εύρημα είναι υπό διερεύνηση καθώς απαιτούνται πολύ περισσότερα στοιχεία τόσο από το γονιδίωμα όσο και από το πρωτέωμα για να γίνει πλήρως κατανοητό και ξεφεύγουν από το αντικείμενο της παρούσας διατριβής.

Για την διερεύνηση της σημαντικότητας των χαρακτηριστικών του ανθρώπινου Uniquome κατασκευάστηκε ένα τεχνητό πρωτέωμα ως εξομοίωση του ανθρώπινου πρωτεώματος βασιζόμενο στις μεθόδους Monte Carlo [81,82]. Οι παραδοχές στις οποίες βασίστηκε η κατασκευή του προσομοιωμένου πρωτεώματος εστιάζουν στον αριθμό, το μήκος, καθώς και την σύσταση από αμινοξέα των ανθρώπινων πρωτεϊνών. Πιο συγκεκριμένα κατασκευάστηκαν ίδιου αριθμού πρωτεΐνες (20.430) με τις θεωρημένες πρωτεΐνες του ανθρώπου που αναλύθηκαν. Η κάθε μία από τις προσομοιωμένες πρωτεΐνες είχε τον ίδιο αριθμό αμινοξέων με κάθε αναλυμένη πρωτεΐνη και επιπλέον σε κάθε θέση της η πιθανότητα εμφάνισης του κάθε αμινοξέος ισούται με την πιθανότητα εμφάνισης του στο σύνολο των ανθρώπινων πρωτεϊνών. Από την παραπάνω ανάλυση διαπιστώθηκε ότι τα ευρήματα σχετικά με τα πεπτίδια του ανθρώπινου Uniquome, είναι χαρακτηριστικά καθόσον δεν επαληθεύονται σε ένα τεχνητά εξομοιωμένο ανθρώπινο πρωτέωμα. Αναλυτικότερα, διαπιστώθηκε ότι στο προσομοιωμένο πρωτέωμα τα CrUPs είναι κατά 17% περισσότερα (8.542.941) σε σχέση με το πρωτέωμα του ανθρώπου (7.263.888) (Πίνακας 15). Αντιθέτως τα CmUPs στο προσομοιωμένο πρωτέωμα είναι

περίπου το ¼ των πεπτιδίων (20.461) σε σχέση με το πρωτέωμα του ανθρώπου (77.697) (Πίνακας 15). Η σημαντικότερη διαφορά παρατηρήθηκε στον αριθμό πρωτεϊνών οι οποίες δεν περιλαμβάνουν μοναδικά πεπτίδια. Συγκεκριμένα στο προσομοιωμένο ανθρώπινο πρωτέωμα διαπιστώθηκε ότι μόλις 4 πρωτεΐνες δεν περιέχουν μοναδικά πεπτίδια ελαχίστου μήκους αντίθετα με το πρωτέωμα του ανθρώπου στο οποίο υπάρχουν 148 τέτοιες πρωτεΐνες (Πίνακας 15). Περαιτέρω στο προσομοιωμένο πρωτέωμα το μήκος των μοναδικών πεπτιδίων ελαχίστου μήκους κυμαίνεται μεταξύ 4 και 9 αμινοξέων (Εικόνα 64, Πίνακας 12) εν αντιθέσει με το μήκος των μοναδικών πεπτιδίων ελαχίστου μήκους στο ανθρώπινο Uniquome που κυμαίνεται από 4 έως 100 αμινοξέα (Εικόνα 39). Αντίστοιχα στα σύνθετα μοναδικά πεπτίδια του ανθρώπινου πρωτεώματος η πλειοψηφία των πεπτιδίων έχει μήκος από 9 έως 11 αμινοξέα (Εικόνα 44), ενώ στο προσομοιωμένο πρωτέωμα η πλειοψηφία των σύνθετων μοναδικών πεπτιδίων έχει μήκος μεγαλύτερο από 100 αμινοξέα (Εικόνα 68, Πίνακας 13). Τέλος, εν αντιθέσει με το ανθρώπινο πρωτέωμα, διαπιστώθηκε ότι στο προσομοιωμένο πρωτέωμα όσο αυξάνεται το μέγεθος των πρωτεϊνών τόσο αυξάνεται και ο αριθμός από μοναδικά πεπτίδια ελαχίστου μήκους (Εικόνα 66). Τα παραπάνω ευρήματα υποδεικνύουν ότι τα χαρακτηριστικά που διαπιστώθηκαν στο ανθρώπινο Uniquome, δεν είναι τυχαία αλλά συνοδεύονται από σημαντικές βιολογικές δράσεις που δεν ήταν πλήρως γνωστές μέχρι τώρα.

Για την κατανόηση της σημαντικότητας και της μοναδικότητας της βιολογικής δράσης του ανθρώπινου Uniquome, διερευνήθηκε επιπλέον η υπόθεση εάν τα CrUPs προέρχονται από μοναδικά νουκλεοτίδια στο ανθρώπινο μεταγράφημα. Με το δεδομένο ότι κατά την μετάφραση δεν είναι μονοσήμαντη η σχέση αμινοξέος – τριάδας νουκλεοτιδίων, η διερεύνηση της μοναδικότητας των ολιγονουκλεοτιδίων από την οποία παράγονται τα CrUPs, βασίστηκε στη θέση του CrUP στο mRNA (position related) εφαρμόζοντας τον αλγόριθμο blastn [98] από τα εργαλεία του NCBI [80] και όχι στην δημιουργία ολιγονουκλεοτιδίου βασιζόμενου στην αλληλουχία του CrUP και στην διαπίστωση μοναδικότητας του στο ανθρώπινο μεταγράφημα (sequence related). Η παραπάνω προσέγγιση εφαρμόστηκε για το σύνολο των CrUPs και προέκυψαν πολύ ενδιαφέροντα ευρήματα σχετικά με την μονοσήμαντη σχέση CrUP → ολιγονουκλεοτιδίου. Για την κατανόηση της σχέσης και την ομοιομορφία της ανάλυσης παρουσιάζεται η παραπάνω προσέγγιση στην οικογένεια πρωτεϊνών RAS. Όπως αναφέρθηκε, η οικογένεια RAS (K-ras, N-ras, H-ras) αποτελείται από 127 μοναδικά πεπτίδια ελαχίστου μήκους (41, 45 και 41 αντίστοιχα). Με το δεδομένο ότι ένα πεπτίδιο μπορεί να παράγεται από περισσότερα του ενός ολιγονουκλεοτιδίου εξετάστηκε εάν η ιδιότητα της μοναδικότητας των πεπτιδίων ελαχίστου μήκους επεκτείνεται και σε μοναδικά ολιγονουκλεοτίδια. Η ανάλυση έδειξε ότι στο σύνολο των 127 μοναδικών πεπτιδίων για

την οικογένεια πρωτεϊνών Ras (K-ras, N-ras και H-ras) υπάρχουν 10 ολιγονουκλεοτίδια τα οποία φαίνεται ότι δεν είναι μοναδικά στο ανθρώπινο μεταγράφημα. Ανάλυση των συγκεκριμένων 10 ολιγονουκλεοτιδίων έδειξε ότι παρόλο που αυτά δεν είναι μοναδικά σαν γραμμική αλληλουχία στο μεταγράφημα είναι μοναδικά ως το πεδίο ανάγνωσης (τριπλέτες) που θα ακολουθηθεί για να παραχθεί η αντίστοιχη μεταφραζόμενη πρωτεΐνη. Δηλαδή τα συγκεκριμένα 10 ολιγονουκλεοτίδια ως μεταφραζόμενη αλληλουχία είναι μοναδικά παρόλο που ως γραμμική αλληλουχία βρίσκονται και σε άλλα σημεία του μεταγραφήματος. Επιβεβαιώνοντας λοιπόν την παραπάνω υπόθεση, στο σύνολο τα μοναδικά πεπτίδια ελαχίστου μήκους των πρωτεϊνών τόσο της οικογένειας RAS όσο και γενικότερα, προέρχονται από μοναδικά ολιγονουκλεοτίδια ακολουθώντας τον κανόνα του πεδίου ανάγνωσης σύμφωνα με το οποίο θα μεταφραστούν σε πεπτίδια (Πίνακας 16,17,18).

Η διερεύνηση της αντίστροφης σχέσης μοναδικότητας ολιγονουκλεοτιδίου → μοναδικότητας πεπτιδίου εντόπισε αλληλουχίες ολιγονουκλεοτιδίων που ενώ ήταν μοναδικές στο ανθρώπινο μεταγράφημα δεν μεταφράζονται σε μοναδικά πεπτίδια. Όπως βρέθηκε το πεπτίδιο EEYSAM ενώ δεν είναι CrUP καθόσον βρίσκεται αποκλειστικά και στις τρείς πρωτεΐνες της οικογένειας RAS και το χαρακτηρίσαμε ως Family Unique Peptide κωδικοποιείται από τρία μοναδικά ολιγονουκλεοτίδια στο ανθρώπινο μεταγράφημα (Πίνακας 19). Το γεγονός αυτό φαίνεται να οφείλεται στην εκφύλιση του γενετικού κώδικα [11,99] καθώς ένα αμινοξύ μπορεί να κωδικοποιείται με περισσότερες από μία τριάδες νουκλεοτιδίων (τριπλέτες). Αυτό υποδεικνύει ότι είναι πιθανόν σημειακές μεταλλάξεις στα μοναδικά ολιγονουκλεοτίδια τελικά δίνουν το ίδιο CrUP και κατ' επέκταση την ίδια πρωτεΐνη δηλαδή κάποιες σημειακές μεταλλάξεις στα μοναδικά νουκλεοτίδια δεν προκαλούν σημαντικές μεταλλάξεις στην τελική αμινοξική αλληλουχία της πρωτεΐνης.

Τα αποτελέσματα της σύνθεσης του ανθρώπινου Uniquome έδειξαν ότι το 99% των θεωρημένων ανθρώπινων πρωτεϊνών περιλαμβάνουν στην αμινοξική τους αλληλουχία τουλάχιστον ένα μοναδικό πεπτίδιο ελαχίστου μήκους, το οποίο τελικά χαρακτηρίζει την αντίστοιχη πρωτεΐνη. Δηλαδή, στο 99% των πρωτεϊνών υπάρχει αμφιμονοσήμαντη σχέση CrUP ↔ Πρωτεΐνη, που σημαίνει ότι κάθε CrUP ανήκει σε μία μόνο πρωτεΐνη και κάθε πρωτεΐνη περιέχει τουλάχιστον ένα CrUP. Το εύρημα αυτό υποδεικνύει ότι η ταυτοποίηση ενός CrUP ταυτοποιεί και την αντίστοιχη πρωτεΐνη. Κατ' επέκταση στην πρωτεωμική ανάλυση η εφαρμογή του παραπάνω ευρήματος υποδεικνύει ότι μία πρωτεΐνη μπορεί να ταυτοποιηθεί με φασματομετρία μάζας [83] από ένα και μόνο πεπτίδιο αρκεί αυτό να είναι CrUP.

Η πειραματική διαδικασία για την ταυτοποίηση των πρωτεϊνών με φασματομετρία μάζας περιλαμβάνει ανάλυση του πεπτιδικού αποτυπώματος τους (peptide finger-print)

[84] και ανάλυση της πεπτιδικής αλληλουχίας (peptide sequence) των πεπτιδίων τους [100]. Πριν την ανάλυση των πεπτιδίων στο φασματογράφο μάζας οι πρωτεΐνες επωάζονται με πρωτεολυτικά ένζυμα και αποικοδομούνται σε πεπτίδια ανάλογα με το ένζυμο [37,101,102]. Το πιο ευρέως χρησιμοποιούμενο ένζυμο στην πρωτεωμική είναι η θρυψίνη (trypsin) [37,103]. Το ένζυμο αυτό αποικοδομεί τις πρωτεΐνες κόβοντας την αλληλουχία τους στην καρβοξυλική περιοχή της λυσίνης (K) ή της αργινίνης (R). Συνδυάζοντας λοιπόν τα παραπάνω, φαίνεται ότι εάν με φασματομετρία μάζας ταυτοποιηθεί ένα CrUP που περιέχεται σε ένα πεπτίδιο προερχόμενο από την επώαση της πρωτεΐνης με την θρυψίνη μπορεί με ασφάλεια να ταυτοποιηθεί και η εν λόγω πρωτεΐνη. Έτσι φτιάχτηκε ένας αλγόριθμος μέσω του οποίου οι 20.282 πρωτεΐνες που περιέχουν μοναδικά πεπτίδια αποικοδομήθηκαν σε πεπτίδια προερχόμενα από την επώαση με θρυψίνη που περιείχαν τα CrUPs των πρωτεϊνών (Πίνακας 20,21). Η βάση αυτή πεπτιδίων που ονομάστηκε Tryptic Digest Unique Peptides είναι ιδιαίτερα χρήσιμη στην ταυτοποίηση πρωτεϊνών με φασματοσκοπία μάζας, καθόσον μέχρι σήμερα η ασφαλής ταυτοποίηση μιας πρωτεΐνης με MS απαιτεί την ταυτοποίηση 2 τουλάχιστον πεπτιδίων ενώ με τη χρήση της παραπάνω βάσης δεδομένων μια πρωτεΐνη ταυτοποιείται ασφαλώς από ένα και μόνο πεπτίδιο. Μία πρώτη εφαρμογή της παραπάνω βάσης σε αποτελέσματα φασματοσκοπίας μάζας αύξησε τον λόγο ταυτοποίησης (identification rate) κατά 15 έως 20%, καθόσον πεπτίδια τα οποία δεν έχουν χρησιμοποιηθεί για την ταυτοποίηση των πρωτεϊνών επαναξιολογούνται και οδηγούν σε ταυτοποίηση επιπλέον πρωτεϊνών. Το ολοκληρωμένο αυτό σύστημα που βασίζεται στην βάση δεδομένων Tryptic Digest Unique Peptides είναι υπό δημιουργία. Το σύστημα αυτό θα οδηγήσει τόσο στην ταυτοποίηση πρωτεϊνών με μεγαλύτερη ακρίβεια όσο και στην ταχύτατη και ασφαλή ταυτοποίηση βιοδεικτών [104] από ένα και μόνο πεπτίδιο υποδεικνύοντας και μία διαγνωστική εφαρμογή του ανθρώπινου Uniquome.

Για την διερεύνηση περαιτέρω εφαρμογών του ανθρώπινου Uniquome όπως αυτό δημιουργήθηκε και αναλύθηκε στην παρούσα διατριβή εξετάστηκαν βάσεις δεδομένων [105,106] με ανοσοπεπτίδια (immune epitopes peptides) [77] και αντιγονικά καρκινικά πεπτίδια (Cancer Antigenic peptides) [78]. Τα ανοσοπεπτίδια είναι πεπτίδια τα οποία χρησιμοποιούνται για την ενεργοποίηση του ανοσοποιητικού συστήματος και την παραγωγή αντισωμάτων. Τα ανοσοπεπτίδια [107] χρησιμοποιούνται σαν αντιγονικοί επίτοποι για την δράση των αντισωμάτων, και την πρόσδεση τους στην πρωτεΐνη στόχο. Από την ανάλυση μας διαπιστώθηκε ότι η συντριπτική πλειοψηφία (87%) των ανοσοπεπτιδίων που αναλύθηκαν εμπεριείχε στην αλληλουχία τους έστω και ένα μοναδικό πεπτίδιο ελαχίστου μήκους της αντίστοιχης πρωτεΐνης (Πίνακας 22). Το ίδιο παρατηρήθηκε και στα αντιγονικά καρκινικά πεπτίδια, τα οποία χρησιμοποιούνται σήμερα ευρέως στη ανοσοθεραπεία μορφών καρκίνου. Διαπιστώθηκε ότι το 89% των

αντιγονικών καρκινικών πεπτιδίων περιείχε CrUP της αντίστοιχης ογκοπρωτείνης (Πίνακας 23). Τα παραπάνω υποδεικνύουν ότι το ανθρώπινο Uniquome μπορεί με ασφάλεια να χρησιμοποιηθεί για την πρόβλεψη και δημιουργία νέων ανοσοπεπτιδίων και αντιγονικών πεπτιδίων τόσο για τον καρκίνο όσο και για άλλες παθολογικές καταστάσεις. Ο σχεδιασμός νέων αντιγονικών πεπτιδίων μπορεί με ασφάλεια μέσω του Uniquome να επεκταθεί και σε μεταλλαγμένες πρωτεΐνες ή σε πρωτεΐνες που δεν έχουν ακόμα σχετιστεί με παθολογικές καταστάσεις.

Στο πλαίσιο αυτό έγινε εκτεταμένη χρήση της προσέγγισης των CrUPs για την πρόβλεψη της Ανοσολογικής Απάντησης, της Ανοσολογικής Διαφυγής και της Παθογένειας του ιού SARS-CoV-2, μέσω των Μοναδικών Πεπτιδικών Υπογραφών του ως προς το Ανθρώπινο Πρωτέωμα [108-111]. Ακολουθώντας την λογική των μοναδικών πεπτιδίων, μελετήθηκαν για πρώτη φορά τα μοναδικά πεπτίδια ενός ιού ως προς το πρωτέωμα του Ανθρώπου. Η προσέγγιση των μοναδικών πεπτιδίων εφαρμόστηκε χαρτογραφώντας όχι τα CrUPs του ιού αυτού καθ' εαυτού, αλλά τα CrUPs του ιού που είναι μοναδικά έναντι ολόκληρου του πρωτεώματος του ξενιστή, δηλαδή εκείνα τα CrUPs του ιού τα οποία δεν υπάρχουν στο πρωτέωμα του ξενιστή. Με δεδομένο ότι ο ξενιστής του ιού SARS-CoV-2 είναι ο *Homo Sapiens* (Human), αναλύσαμε το SARS-CoV-2 πρωτέωμα για την ταυτοποίηση των CrUPs αυτού έναντι όλου του ανθρώπινου πρωτεώματος, με τη νέα αυτή κατηγορία πεπτιδίων να ονομάζεται πλέον C/H-CrUPs (Covid vs Human-Core Unique Peptides). Διαπιστώθηκε ότι τα C/H-CrUPs έδωσαν σημαντικές πληροφορίες για την τη διαλεύκανση του μηχανισμού των αλληλεπιδράσεων του ιού SARS-CoV-2 και ξενιστή Ανθρώπου. Όπως γίνεται αντιληπτό η χρήση τέτοιων πεπτιδίων μπορεί να εφαρμοστεί και σε άλλους ιούς (ξενιστές) με σκοπό την αποσαφήνιση του τρόπου με τον οποίο αλληλοεπιδρούν με τον άνθρωπο.

Οι νέες οντότητες πεπτιδίων που αναπτύχθηκαν για πρώτη φορά και περιλήφθηκαν στο Uniquome του ανθρώπου, ανέδειξαν μία ιδιαίτερη ομάδα βιολογικών μορίων (CrUPs και CmUPs) με μοναδικά χαρακτηριστικά. Σε πρώτο επίπεδο το Uniquome είναι ιδιαίτερα χρήσιμο στην πρωτεωμική καθόσον με τη χρήση του μπορεί να αυξηθεί αισθητά η ακρίβεια και τα ποσοστά της ταυτοποίησης των πρωτεϊνών με την μέθοδο της φασματομετρίας μάζας. Αυτό οδηγεί στην ταυτοποίηση βιοδεικτών με μεγαλύτερη ευαισθησία και με πιθανή διαγνωστική αξία. Περαιτέρω σε δεύτερο επίπεδο, το Uniquome μπορεί να φανεί πολύτιμο στην κατανόηση βιολογικών φαινομένων τόσο σε φυσιολογικό επίπεδο όσο και σε επίπεδο ασθενειών. Επεκτείνοντας τα παραπάνω το Uniquome έχει άμεση εφαρμογή και χρησιμότητα στην κατανόηση και στην πρόβλεψη της δράσης των ιών, όπως αποδείχθηκε με την χρήση των C/H-CrUPs για την κατανόηση της δράσης του Sars-Cov-2. Η προσέγγιση αυτή δίνει την δυνατότητα της δημιουργίας

νέων θεραπευτικών προσεγγίσεων στην αντιμετώπιση των ιών. Τα ευρήματα της παρούσας διατριβής σχετικά με τα ανοσοπεπτίδια καθώς και με τα αντιγονικά καρκινικά πεπτίδια υποδεικνύουν ότι το Uniquome, μέσω της χρήσης των μοναδικών πεπτιδίων, μπορεί να διαδραματίσει καθοριστικό ρόλο στην δημιουργία νέων ανοσοπεπτιδίων καθώς και αντιγονικών καρκινικών πεπτιδίων πιο εξειδικευμένων και πιο αποτελεσματικών από τα έως σήμερα διαθέσιμα. Τέτοια θεραπευτικά πεπτίδια μπορούν να εφαρμοστούν για την αποτελεσματική αντιμετώπιση παθήσεων μέσω της ανοσοθεραπείας.

Βιβλιογραφία

[1]. Βασικές Αρχές Μοριακής Βιολογίας (Burton E. Tropp) Α Ελληνική Έκδοση.

[2]. Kirschning A. On the Evolutionary History of the Twenty Encoded Amino Acids. Chemistry. 2022 Oct 4;28(55):e202201419. doi: 10.1002/chem.202201419. Epub 2022 Jul 28. PMID: 35726786; PMCID: PMC9796705.

[3]. Katchalski-Katzir, E., Kasher, R., Fridkin, M. (2006). Amino Acids: Physicochemical Properties. In: Encyclopedic Reference of Genomics and Proteomics in Molecular Medicine. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-29623-9_2260

[4]. Song, Lingshuang and Yang, Lin and Meng, Jie and Yang, Sichun, Thermodynamics of Hydrophobic Amino Acids in Solution: A Combined Experimental–Computational Study, The Journal of Physical Chemistry Letters, 8(2), 2017 doi: 10.1021/acs.jpclett.6b02673

[5]. Widyarani, Sari YW, Ratnaningsih E, Sanders JP, Bruins ME. Production of hydrophobic amino acids from biobased resources: wheat gluten and rubber seed proteins. Appl Microbiol Biotechnol. 2016 Sep;100(18):7909-20. doi: 10.1007/s00253-016-7441-8. Epub 2016 Apr 27. PMID: 27118013; PMCID: PMC4989023.

[6]. Mant CT, Kovacs JM, Kim HM, Pollock DD, Hodges RS. Intrinsic amino acid side-chain hydrophilicity/hydrophobicity coefficients determined by reversed-phase high-performance liquid chromatography of model peptides: comparison with other hydrophilicity/hydrophobicity scales. Biopolymers. 2009;92(6):573-95. doi: 10.1002/bip.21316. PMID: 19795449; PMCID: PMC2792893.

[7]. Martínez-Bachs, B.; Rimola, A. Prebiotic Peptide Bond Formation Through Amino Acid Phosphorylation. Insights from Quantum Chemical Simulations. *Life* 2019, *9*, 75. https://doi.org/10.3390/life9030075

[8]. Fahmeed Sheehan, Deborah Sementa, Ankit Jain, Mohit Kumar, Mona Tayarani-Najjaran, Daniela Kroiss, and Rein V. Ulijn, Peptide-Based Supramolecular Systems Chemistry, *Chemical Reviews* 2021 *121* (22), 13869-13914, doi: 10.1021/acs.chemrev.1c00089

[9]. Richard M. Twyman, (2020), Αρχές Πρωτεωμικής, (Πρόλογος - Γενική Επιμέλεια - Μετάφραση: Κωνσταντίνος Ε. Βοργιάς, Γεώργιος Θ. Τσάγκαρης) Broken Hill Publishers Ltd

[10]. Weil P. Protein Synthesis & the Genetic Code. In: Rodwell VW, Bender DA, Botham KM, Kennelly PJ, Weil P. eds. *Harper's Illustrated Biochemistry, 30e*. McGraw Hill; 2016. Accessed March 08, 2023. https://accessmedicine.mhmedical.com/content.aspx?bookid=1366&sectionid=73245262

[11]. Milton H. Saier, Understanding the Genetic Code, Journal of Bacteriology Vol. 201, No. 15, 10 July 2019, doi: https://doi.org/10.1128/JB.00091-19

[12]. Tseng, Yan Yuan, Li, Wen-Hsiung, Classification of protein functional surfaces using structural characteristics, Proceedings of the National Academy of Sciences, 109 (4) 2012, 1170-11, doi:10.1073/pnas.1119684109

[13]. Sanvictores T, Farci F. Biochemistry, Primary Protein Structure. 2022 Oct 31. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2022 Jan–. PMID: 33232013.

[14]. Mónika Bokor and Ágnes Tantos, Secondary Structures of Proteins: A Comparison of Models and Experimental Results, *J. Proteome Res.* 2021, 20, 3, 1802–1808, https://doi.org/10.1021/acs.jproteome.0c00986

[15]. Rehman I, Kerndt CC, Botelho S. Biochemistry, Tertiary Protein Structure. 2022 Sep 12. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2022 Jan–. PMID: 29262204.

[16]. Yu X, Wang C, Li Y. Classification of protein quaternary structure by functional domain composition. BMC Bioinformatics. 2006 Apr 4;7:187. doi: 10.1186/1471-2105-7-187. PMID: 16584572; PMCID: PMC1450311.

[17]. Korkmaz S, Duarte JM, Prlić A, Goksuluk D, Zararsiz G, Saracbasi O, Burley SK, Rose PW. Investigation of protein quaternary structure via stoichiometry and symmetry information. PLoS One. 2018 Jun 4;13(6):e0197176. doi: 10.1371/journal.pone.0197176. Erratum in: PLoS One. 2018 Jul 27;13(7):e0201403. PMID: 29864163; PMCID: PMC5986128.

[18]. Wang K, Huang C, Nice E. Recent advances in proteomics: towards the human proteome. Biomed Chromatogr. 2014 Jun;28(6):848-57. doi: 10.1002/bmc.3157. PMID: 24861753.

[19]. Lovell JT, Grimwood J. The first complete human genome. Nature. 2022 Jun;606(7914):468-469. doi: 10.1038/d41586-022-01368-w. PMID: 35606432.

[20]. Altelaar AF, Munoz J, Heck AJ. Next-generation proteomics: towards an integrative view of proteome dynamics. Nat Rev Genet. 2013 Jan;14(1):35-48. doi: 10.1038/nrg3356. Epub 2012 Dec 4. PMID: 23207911.

[21]. Zhu Z, Lu JJ, Liu S. Protein separation by capillary gel electrophoresis: a review. Anal Chim Acta. 2012 Jan 4; 709:21-31. doi: 10.1016/j.aca.2011.10.022. Epub 2011 Oct 19. PMID: 22122927; PMCID: PMC3227876.

[22]. Gao M, Qi D, Zhang P, Deng C, Zhang X. Development of multidimensional liquid chromatography and application in proteomic analysis. Expert Rev Proteomics. 2010 Oct;7(5):665-78. doi: 10.1586/epr.10.49. PMID: 20973640.

[23]. Janina Kneipp, Lisa M. Miller, Marion Joncic, Martin Kittel, Peter Lasch, Michael Beekes, Dieter Naumann,In situ identification of protein structural changes in prion-infected tissue, Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease, Volume 1639, Issue 3, 2003, Pages 152-158, ISSN 0925-4439, https://doi.org/10.1016/j.bbadis.2003.08.005.

[24]. Changhui Yan, Drena Dobbs, Vasant Honavar, A two-stage classifier for identification of protein–protein interface residues, *Bioinformatics*, Volume 20, Issue suppl_1, 4 August 2004, Pages i371–i378, https://doi.org/10.1093/bioinformatics/bth920

[25]. Domon B, Aebersold R. Mass spectrometry and protein analysis. Science. 2006 Apr 14;312(5771):212-7. doi: 10.1126/science.1124619. PMID: 16614208.

[26]. Zhe Yang, Zhengyi Ren, Yongjun Cheng, Wenjun Sun, Zhenghua Xi, Wenjie Jia, Gang Li, Yongjun Wang, Meiru Guo, Detian Li, Review and prospect on portable mass spectrometer for recent applications, Vacuum, Volume 199, 2022, 110889, ISSN 0042-207X, https://doi.org/10.1016/j.vacuum.2022.110889.

[27]. Kösling P, Rüger CP, Schade J, Fort KL, Ehlert S, Irsig R, Kozhinov AN, Nagornov KO, Makarov A, Rigler M, Tsybin YO, Walte A, Zimmermann R. Vacuum Laser Photoionization inside the C-trap of an Orbitrap Mass Spectrometer: Resonance-Enhanced Multiphoton Ionization High-Resolution Mass Spectrometry. Anal Chem. 2021 Jul 13;93(27):9418-9427. doi: 10.1021/acs.analchem.1c01018. Epub 2021 Jun 25. PMID: 34170684.

[28]. Valentine SJ, Liu X, Plasencia MD, Hilderbrand AE, Kurulugama RT, Koeniger SL, Clemmer DE. Developing liquid chromatography ion mobility mass spectometry techniques. Expert Rev Proteomics. 2005 Aug;2(4):553-65. doi: 10.1586/14789450.2.4.553. PMID: 16097888.

[29]. Rauser S, Höfler H, Walch A. In-situ-Proteomanalyse von Geweben: Mittels bildgebender Massenspektrometrie (MALDI Imaging) [MALDI imaging mass spectrometry for direct tissue analysis]. Pathologe. 2009 Dec;30 Suppl 2:140-5. German. doi: 10.1007/s00292-009-1185-5. PMID: 19756619.

[30]. Laughlin S, Wilson WD. May the Best Molecule Win: Competition ESI Mass Spectrometry. Int J Mol Sci. 2015 Oct 15;16(10):24506-31. doi: 10.3390/ijms161024506. PMID: 26501262; PMCID: PMC4632762.

[31]. Zaikin VG, Borisov RS. Options of the Main Derivatization Approaches for Analytical ESI and MALDI Mass Spectrometry. Crit Rev Anal Chem. 2022;52(6):1287-1342. doi: 10.1080/10408347.2021.1873100. Epub 2021 Feb 8. PMID: 33557614.

[32]. Matsumoto H, Haniu H, Komori N. Determination of Protein Molecular Weights on SDS-PAGE. Methods Mol Biol. 2019; 1855:101-105. doi: 10.1007/978-1-4939-8793-1_10. PMID: 30426411.

[33]. Perez-Riverol Y, Wang R, Hermjakob H, Müller M, Vesada V, Vizcaíno JA. Open-source libraries and frameworks for mass spectrometry-based proteomics: a developer's perspective. Biochim Biophys Acta. 2014 Jan;1844 (1 Pt A): 63-76. doi: 10.1016/j.bbapap.2013.02.032. Epub 2013 Mar 1. PMID: 23467006; PMCID: PMC3898926.

[34]. Brosch M, Swamy S, Hubbard T, Choudhary J. Comparison of Mascot and X!Tandem performance for low and high accuracy mass spectrometry and the development of an adjusted Mascot threshold. Mol Cell Proteomics. 2008 May;7(5):962-70. doi: 10.1074/mcp.M700293-MCP200. Epub 2008 Jan 23. PMID: 18216375; PMCID: PMC2656932.

[35]. Sacks GL, Derry LA, Brenna JT. Elemental speciation by parallel elemental and molecular mass spectrometry and peak profile matching. Anal Chem. 2006 Dec 15;78(24):8445-55. doi: 10.1021/ac0612170. PMID: 17165838.

[36]. Kawashima Y, Watanabe E, Umeyama T, Nakajima D, Hattori M, Honda K, Ohara O. Optimization of Data-Independent Acquisition Mass Spectrometry for Deep and Highly Sensitive Proteomic Analysis. Int J Mol Sci. 2019 Nov 26;20(23):5932. doi: 10.3390/ijms20235932. PMID: 31779068; PMCID: PMC6928715.

[37]. Therese Dau, Giulia Bartolomucci, and Juri Rappsilber, Proteomics Using Protease Alternatives to Trypsin Benefits from Sequential Digestion with Trypsin, *Anal. Chem.* 2020, 92, 14, 9523–9527 July 6, 2020, https://doi.org/10.1021/acs.analchem.0c00478

[38]. Duc T. Tran and Valerie J. Cavett and Vuong Q. Dang and Héctor L. Torres and Brian M. Paegel, Evolution of a mass spectrometry-grade protease with PTM-directed specificity, Proceedings of the National Academy of Sciences, 113(51), 2016, 14686-14691, doi: 10.1073/pnas.1609925113

[39]. Chen C, Hou J, Tanner JJ, Cheng J. Bioinformatics Methods for Mass Spectrometry-Based Proteomics Data Analysis. Int J Mol Sci. 2020 Apr 20;21(8):2873. doi: 10.3390/ijms21082873. PMID: 32326049; PMCID: PMC7216093.

[40]. Aggarwal S, Yadav AK. False Discovery Rate Estimation in Proteomics. Methods Mol Biol. 2016;1362:119-28. doi: 10.1007/978-1-4939-3106-4_7. PMID: 26519173.

[41]. Savitski MM, Nielsen ML, Zubarev RA. New data base-independent, sequence tag-based scoring of peptide MS/MS data validates Mowse scores, recovers below threshold data, singles out modified peptides, and assesses the quality of MS/MS techniques. Mol Cell Proteomics. 2005 Aug;4(8):1180-8.
doi: 10.1074/mcp.T500009-MCP200. Epub 2005 May 22. PMID: 15911534.

[42]. Check Hayden, E. Genome researchers raise alarm over big data. *Nature* (2015). https://doi.org/10.1038/nature.2015.17912

[43]. Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, et al. (2015) Big Data: Astronomical or Genomical? PLoS Biol 13(7): e1002195. https://doi.org/10.1371/journal.pbio.1002195

[44]. UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res. 2023 Jan 6;51(D1): D523-D531. doi: 10.1093/nar/gkac1052. PMID: 36408920; PMCID: PMC9825514.

[45]. Pundir S, Martin MJ, O'Donovan C; UniProt Consortium. UniProt Tools. Curr Protoc Bioinformatics. 2016 Mar 24;53:1.29.1-1.29.15. doi: 10.1002/0471250953.bi0129s53. PMID: 27010333; PMCID: PMC4941944.

[46]. Wang Y, Wang Q, Huang H, Huang W, Chen Y, McGarvey PB, Wu CH, Arighi CN; UniProt Consortium. A crowdsourcing open platform for literature curation in UniProt. PLoS Biol. 2021 Dec 6;19(12):e3001464. doi: 10.1371/journal.pbio.3001464. PMID: 34871295; PMCID: PMC8675915.

[47]. Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res. 2000 Jan 1;28(1):45-8. doi: 10.1093/nar/28.1.45. PMID: 10592178; PMCID: PMC102476.

[48]. Amos Bairoch, Rolf Apweiler, The SWISS-PROT protein sequence data bank and its supplement TrEMBL, *Nucleic Acids Research*, Volume 25, Issue 1, 1 January 1997, Pages 31–36, https://doi.org/10.1093/nar/25.1.31

[49]. Barker WC, Garavelli JS, Huang H, McGarvey PB, Orcutt BC, Srinivasarao GY, Xiao C, Yeh LS, Ledley RS, Janda JF, Pfeiffer F, Mewes HW, Tsugita A, Wu C. The protein information resource (PIR). Nucleic Acids Res. 2000 Jan 1;28(1):41-4. doi: 10.1093/nar/28.1.41. PMID: 10592177; PMCID: PMC102418.

[50]. Villalba GC, Matte U. Fantastic databases and where to find them: Web applications for researchers in a rush. Genet Mol Biol. 2021 Apr 2;44(2):e20200203. doi: 10.1590/1678-4685-GMB-2020-0203. PMID: 33821874; PMCID: PMC8022358.

[51]. Jo Guldi, The Algorithm: Mapping Long-Term Trends and Short-Term Change at Multiple Scales of Time, *The American Historical Review*, Volume 127, Issue 2, June 2022, Pages 895–911, https://doi.org/10.1093/ahr/rhac160

[52]. Nasar, Audrey A. (2016) "The history of Algorithmic complexity," The Mathematics Enthusiast: Vol. 13: No. 3, Article 4. DOI: https://doi.org/10.54870/1551-3440.1375

[53]. Zlatopolski, D. (2021). FROM THE HISTORY OF THE BINARY NUMBER SYSTEM. UNKNOWN AUTHOR'S BROCHURE. Informatics in school. 60-62. 10.32517/2221-1993-2021-20-3-60-62.

[54]. Perera, Piumi & Tennakoon, Geethya & Ahangama, Supunmali & Panditharathna, Rangana & Chathuranga, Buddhika. (2021). A Systematic Review of Introductory

Programming Languages for Novice Learners. IEEE Access. PP. 1-1. 10.1109/ACCESS.2021.3089560.

[55]. K P, Naveen Reddy &. Y, Geyavalli &. D, Sujani & Rajesh, More. (2018). Comparison of Programming Languages: Review. 9. 113-122.

[56]. Agrawal, Parag & Gupta, Amit & Mathur, Priya. (2021). CPU Scheduling in Operating System: A Review. 10.1007/978-981-15-9689-6_31.

[57]. Stephen Blair Chappel *Parallel Programming with Intel Parallel Studio X,* 2016

[58]. Matsuura, A. & Mattson, Tim. (2022). Introducing the Quantum Research Kernels: Lessons from Classical Parallel Computing. 10.48550/arXiv.2211.00844.

[59]. Amoretti, Michele. (2020). Review of Elements of Parallel Computing. ACM SIGACT News. 51. 10-13. 10.1145/3427361.3427365.

[60]. Al-Shafei, Ahmed & Zareipour, Hamidreza & Cao, Yankai. (2022). High-Performance and Parallel Computing Techniques Review: Applications, Challenges and Potentials to Support Net-Zero Transition of Future Grids. Energies. 15. 8668 doi:10.3390/en15228668.

[61]. Jacksi, Karwan & Najat, Zryan & Zeebaree, Subhi & Hussein, Karzan. (2018). Distributed Cloud Computing and Distributed Parallel Computing: A Review. 10.1109/ICOASE.2018.8548937.

[62]. R. L. R. Maata, R. Cordova, B. Sudramurthy and A. Halibas, "Design and Implementation of Client-Server Based Application Using Socket Programming in a Distributed Computing Environment," *2017 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, Coimbatore, India, 2017, pp. 1-4, doi: 10.1109/ICCIC.2017.8524573.

[63]. Y. Ma, C. Lu, B. Sinopoli and S. Zeng, "Exploring Edge Computing for Multitier Industrial Control," in *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 39, no. 11, pp. 3506-3518, Nov. 2020, doi: 10.1109/TCAD.2020.3012648.

[64]. Liu, L. (2017). Network computing principle and application analysis based on distributed peer-to-peer. Acta Technica CSAV (Ceskoslovensk Akademie Ved). 62. 415-424.

[65]. P., Kar, A.K. Big Data Analytics: A Review on Theoretical Contributions and Tools Used in Literature. *Glob J Flex Syst Manag* 18, 203–229 (2017). https://doi.org/10.1007

[66]. Gupta, Priyanka & Sawant, Vinaya. (2020). A Map Reduce Based Parallel Algorithm. Journal of University of Shanghai for Science and Technology. 22. 10.51201/jusst12486.

[67]. Anastasia Alexandridou, George Th. Tsangaris, Konstantinos Vougas, Konstantina Nikita, George Spyrou, Peptide Finder: mapping measured molecular masses to

peptides and proteins, *Bioinformatics*, Volume 24, Issue 19, October 2008, Pages 2267–2269, https://doi.org/10.1093/bioinformatics/btn413

[68]. Anastasia Alexandridou, George Th. Tsangaris, Konstantinos Vougas, Konstantina Nikita, George Spyrou, UniMaP: finding unique mass and peptide signatures in the human proteome, *Bioinformatics*, Volume 25, Issue 22, November 2009, Pages 3035–3037, https://doi.org/10.1093/bioinformatics/btp516

[69]. Alexandridou, Anastasia and Dovrolis, Nikolas and Tsangaris, George T and Konstantina S Nikita and George M Spyrou, PepServe: a web server for peptide analysis, clustering and visualization, Nuclear Acids Research 05/2011

[70]. Evangelos Kontopodis 2015, Ανάλυση των πεπτιδίων μοναδικής αλληλουχίας αμινοξέων (core unique peptides) στο ανθρώπινο πρωτέωμα (Μεταπτυχιακή Εργασία).

[71]. Vasileios Pierros 2018, Υπολογιστική μέθοδος για την κατασκευή και αξιολόγηση του UNIQUOME ενός οργανισμού (Μεταπτυχιακή Εργασία).

[72]. Yu, Dong & Lee, Dae-Hee & Kim, Seong & Lee, Choong & Song, Ju & Kong, Eun & Kim, Jihyun. (2012). Algorithm for predicting functionally equivalent proteins from BLAST and HMMER searches. Journal of microbiology and biotechnology. 22. 1054-8.

[73]. linares-López, F., Berthet, Q., Blondel, M. *et al.* Deep embedding and alignment of protein sequences. *Nat Method* 20,104–111 (2023). https://doi.org/10.1038/s41592-022-01700-2

[74]. Pieter Verschaffelt, Tim Van Den Bossche, Lennart Martens, Peter Dawyndt, and Bart Mesuere, Unipept Desktop: a faster, more powerful metaproteomics results analysis tool, Journal of Proteome Research, 2021, doi.org/10.1021/acs.jproteome.0c00855

[75]. Bart Mesuere, Felix Van der Jeugt, Bart Devreese, Peter Vandamme, and Peter Dawyndt The unique peptidome: Taxon-specific tryptic peptides as biomarkers for targeted metaproteomics, Proteomics, 2016, 16 (17), pp 2313–2318, doi.org/10.1002/pmic.201600023

[76]. Martini, S., Nielsen, M., Peters, B. et al. The Immune Epitope Database and Analysis Resource Program 2003–2018: reflections and outlook. Immunogenetics 72, 57–76 (2020). https://doi.org/10.1007/s00251-019-01137-6

[77]. Gori, A., Longhi, R., Peri, C. *et al.* Peptides for immunological purposes: design, strategies and applications. *Amino Acids* 45, 257–268 (2013). https://doi.org/10.1007/s00726-013-1526-9

[78]. León-Letelier, R.A.; Katayama, H.; Hanash, S. Mining the Immunopeptidome for Antigenic Peptides in Cancer. Cancers 2022, 14, 4968. https://doi.org/10.3390/cancers14204968

[79]. Agarwala, Richa & Barrett, Tanya & Beck, Jeff & Benson, Dennis & Bollin, Colleen & Bolton, Evan & Bourexis, Devon & Brister, J. & Bryant, Stephen & Canese, Kathi & Cavanaugh, Mark & Charowhas, Chad & Clark, Kami & Dondoshansky, Ilya & Feolo, Michael & Fitzpatrick, Lawrence & Funk, Kathryn & Geer, Lewis & Gorelenkov, Viatcheslav & Coordinators, NCBI. (2018). Database resources of the National Center for Biotechnology Information. Nucleic Acids Research. 46. D8-D13. 10.1093/nar/gkx1095.

[80]. Sofi, Mohammad Yaseen & Shafi, Afshana & Masoodi, Khalid. (2022). NCBI BLAST. 10.1016/B978-0-323-91128-3.00021-5.

[81]. Adam M. Johansen and Ludger Evers, Monte Carlo Methods

[82]. Hanada, Masanori & Matsuura, So. (2022). What is the Monte Carlo Method? A Simulation with Random Numbers. 10.1007/978-981-19-2715-7_2.

[83]. Anas El-Aneed, Aljandro Cohen & Joseph Banoub (2009) Mass Spectrometry, Review of the Basics: Electrospray, MALDI, and Commonly Used Mass Analyzers, Applied Spectroscopy Reviews, 44:3, 210-230, DOI: 10.1080/05704920902717872

[84]. Thiede B, Höhenwarter W, Krah A, Mattow J, Schmid M, Schmidt F, Jungblut PR. Peptide mass fingerprinting. Methods. 2005 Mar;35(3):237-47. doi: 10.1016/j.ymeth.2004.08.015. Epub 2005 Jan 12. PMID: 15722220.

[85]. Arumugaperumal, A., Velayudhan Krishna, D., Alaguponniah, S. *et al.* PeptCreatR: A Web App for Unique Peptides in Human. *Int J Pept Res Ther* **28**, 64 (2022). https://doi.org/10.1007/s10989-022-10375-4

[86]. Fernandez PC, Frank SR, Wang L, Schroeder M, Liu S, Greene J, Cocito A, Amati B. Genomic targets of the human c-Myc protein. Genes Dev. 2003 May 1;17(9):1115-29. doi: 10.1101/gad.1067003. Epub 2003 Apr 14. PMID: 12695333; PMCID: PMC196049.

[87]. Albert J Kooistra, Stefan Mordalski, Gáspár Pándy-Szekeres, Mauricio Esguerra, Alibek Mamyrbekov, Christian Munk, György M Keserű, David E Gloriam, GPCRdb in 2021: integrating GPCR sequence, structure and function, *Nucleic Acids Research*, Volume 49, Issue D1, 8 January 2021, Pages D335–D343, https://doi.org/10.1093/nar/gkaa1080

[88]. Neil D. Rawlings, Alex Bateman, Origins of peptidases, Biochimie Volume 166, 2019, Pages 4-18, ISSN 0300-9084, https://doi.org/10.1016/j.biochi.2019.07.026.

[89]. Kuznetsov A, Voronina A, Govorun V, Arapidi G. Critical Review of Existing MHC I Immunopeptidome Isolation Methods. *Molecules*. 2020; 25(22):5409. https://doi.org/10.3390/molecules25225409

[90]. John Colicelli, Human RAS Superfamily Proteins and Related GTPases, Sci STKE. 2004 Sep 7; 2004(250): RE13, Published online 2004 Sep 7. doi: 10.1126/stke.2502004re13

[91]. Norman RL, Singh R, Langridge JI, Ng LL, Jones DJL. The measurement of KRAS G12 mutants using multiplexed selected reaction monitoring and ion mobility mass spectrometry. Rapid Commun Mass Spectrom. 2020 Sep;34 Suppl 4(Suppl 4):e8657. doi: 10.1002/rcm.8657. Epub 2020 Feb 14. PMID: 31800120; PMCID: PMC7539944.

[92]. Friedman JR, Nunnari J. Mitochondrial form and function. Nature. 2014 Jan 16;505(7483):335-43. doi: 10.1038/nature12985. PMID: 24429632; PMCID: PMC4075653.

[93]. Antonarakis, S., Lyle, R., Dermitzakis, E. *et al.* Chromosome 21 and Down syndrome: from genomics to pathophysiology. *Nat Rev Genet* 5, 725–738 (2004). https://doi.org/10.1038/nrg1448

[94]. Gravholt, C.H., Viuff, M.H., Brun, S. *et al.* Turner syndrome: mechanisms and management. *Nat Rev Endocrinol* 15, 601–614 (2019). https://doi.org/10.1038/s41574-019-0224-4

[95]. Kristian A. Groth, Anne Skakkebæk, Christian Høst, Claus Højbjerg Gravholt, Anders Bojesen, Klinefelter Syndrome—A Clinical Update, *The Journal of Clinical Endocrinology & Metabolism*, Volume 98, Issue 1, 1 January 2013, Pages 20–30, https://doi.org/10.1210/jc.2012-2382

[96]. Laura Re, Jutta M. Birkhoff,The 47,XYY syndrome, 50 years of certainties and doubts: A systematic review, Aggression and Violent Behavior, Volume 22 2015, Pages 9-17,ISSN 1359-1789, https://doi.org/10.1016/j.avb.2015.02.003.

[97]. Donna Maria E. Cortezzo, MD Leandra K. Tolusso, MS, CGC , Perinatal Outcomes of Fetuses and Infants Diagnosed with Trisomy 13 or Trisomy 18, The journal of Pediatrics, Volume: 247, P116-123.e5, August 2022, doi:https://doi.org/10.1016/j.jpeds.2022.04.010

[98]. Scott McGinnis, Thomas L. Madden, BLAST: at the core of a powerful and diverse set of sequence analysis tools, *Nucleic Acids Research*, Volume 32, Issue suppl_2, 1 July 2004, Pages W20–W25, https://doi.org/10.1093/nar/gkh435

[99]. Shixin Ye, Jean Lehmann, Genetic code degeneracy is established by the decoding center of the ribosome, *Nucleic Acids Research*, Volume 50, Issue 7, 22 April 2022, Pages 4113–4126, https://doi.org/10.1093/nar/gkac171

[100]. Zubarev Roman A. and Makarov Alexander, Orbitrap Mass Spectrometry, Analytical Chemistry, 2013, 85, 11, 5288–5296, https://doi.org/10.1021/ac4001223

[101]. Mótyán, J.A.; Tóth, F.; Tőzsér, J. Research Applications of Proteolytic Enzymes in Molecular Biology. *Biomolecules* 2013, 923-942. https://doi.org/10.3390/biom3040923

[102]. Wai-Kok Choong, Ching-Tai Chen, Jen-Hung Wang, and Ting-Yi Sung, In Silico Human Proteome Digestion Map with Proteolytic Peptide Analysis and Graphical Visualizations, Journal of Proteome Research 2019 18 (12), 4124-4132 DOI: 10.1021/acs.jproteome.9b00350

[103]. Saveliev, S., Bratz, M., Zubarev, R. *et al.* Trypsin/Lys-C protease mix for enhanced protein mass spectrometry analysis. *Nat Methods* 10, i–ii (2013). https://doi.org/10.1038/nmeth.f.371

[104]. Anagnostopoulos AK, Tsiliki G, Spyrou G, Tsangaris GT. Bioinformatics approaches in the discovery and understanding of reproduction-related biomarkers. Expert Rev Proteomics. 2011 Apr;8(2):187-95. doi: 10.1586/epr.11.12. PMID: 21501012.

[105]. Tanishq Chamoli,Alisha Khera,Akanksha Sharma,Anshul Gupta,Sonam Garg,Kanishk Mamgain,Aayushi Bansal,Shriya Verma,Ankit Gupta,Hema K. Alajangi,Gurpal Singh,Ravi P. Barnwal, Peptide Utility (PU) search server: A new tool for peptide sequence search from multiple databases, Heliyon December 10 2022, https://doi.org/10.1016/j.heliyon.2022.e12283

[106]. Kalmykova, S.D., Arapidi, G.P., Urban, A.S. et al. In Silico Analysis of Peptide Potential Biological Functions. Russ J Bioorg Chem 44, 367–385 (2018). https://doi.org/10.1134/S106816201804009X

[107]. Trier N, Hansen P, Houen G. Peptides, Antibodies, Peptide Antibodies and More. Int J Mol Sci. 2019 Dec 13;20(24):6289. doi: 10.3390/ijms20246289. PMID: 31847088; PMCID: PMC6941022

[108]. Pierros V, Kontopodis E, Stravopodis DJ, Tsangaris GT. Unique peptide signatures of SARS-CoV-2 virus against human proteome reveal variants' immune escape and infectiveness. Heliyon. 2022 Apr;8(4): e09222. doi: 10.1016/j.heliyon.2022.e09222. Epub 2022 Apr 4. PMID: 35399374; PMCID: PMC8979629.

[109]. Kontopodis E, Pierros V, Stravopodis DJ, Tsangaris GT. Prediction of SARS-CoV-2 Omicron Variant Immunogenicity, Immune Escape and Pathogenicity, through the Analysis of Spike Protein-Specific Core Unique Peptides. Vaccines (Basel). 2022 Feb 24;10(3):357. doi: 10.3390/vaccines10030357. PMID: 35334990; PMCID: PMC8955659.

[110]. Hatmal MM, Alshaer W, Al-Hatamleh MAI, Hatmal M, Smadi O, Taha MO, Oweida AJ, Boer JC, Mohamud R, Plebanski M. Comprehensive Structural and Molecular Comparison of Spike Proteins of SARS-CoV-2, SARS-CoV and MERS-CoV, and Their Interactions with ACE2. Cells. 2020 Dec 8;9(12):2638. doi: 10.3390/cells9122638. PMID: 33302501; PMCID: PMC7763676.

[111]. Chen Y, Zhang YN, Yan R, Wang G, Zhang Y, Zhang ZR, Li Y, Ou J, Chu W, Liang Z, Wang Y, Chen YL, Chen G, Wang Q, Zhou Q, Zhang B, Wang C. ACE2-targeting monoclonal antibody as potent and broad-spectrum coronavirus blocker. Signal Transduct Target Ther. 2021 Aug 25;6(1):315. doi: 10.1038/s41392-021-00740-y. PMID: 34433803; PMCID: PMC8385704.

Παράρτημα

Πρωτότυπες Ερευνητικές Δημοσιεύσεις
Διδακτορικής Διατριβής

- **Evangelos Kontopodis, Vasileios Pierros, Dimitrios J. Stravopodis and George T.Tsangaris**
  Prediction of SARS-CoV-2 Omicron Variant Immunogenicity, Immune Escape and Pathogenicity, through the Analysis of Spike Protein-Specific Core Unique Peptides.
  Vaccines 2022, 10, 357. https://doi.org/10.3390/vaccines10030357

- **Vasileios Pierros, Evangelos Kontopodis, Dimitrios J. Stravopodis and George T.Tsangaris**
  Unique peptide signatures of SARS-CoV-2 virus against human proteome reveal variants' immune escape and infectiveness.
  Heliyon 2022, Volume 8, Issue 4, E09222, April 01, 2022.
  https://doi.org/10.1016/j.heliyon.2022.e09222

- **Evangelos Kontopodis, Vasileios Pierros, Dimitrios J. Stravopodis and George T.Tsangaris**
  Prediction of SARS-CoV-2 Omicron Variant Immunogenicity, Immune Escape and Pathogenicity, through Analysis of Spike Protein-specific Core Unique Peptides
  medRxiv 2021, December. https://doi.org/10.1101/2021.12.26.21268418

- **Evangelos Kontopodis, Vasileios Pierros, Dimitrios J. Stravopodis and George T.Tsangaris**
  Unique Peptide Signatures Of SARS-CoV-2 Against Human Proteome Reveal Variants' Immune Escape And Infectiveness
  BioRxiv 2021, October. Doi: 10.1101/2021.10.03.462911

Prediction of SARS-CoV-2 Omicron Variant Immunogenicity, Immune Escape and Pathogenicity, through the Analysis of Spike Protein-Specific Core Unique Peptides

Evangelos Kontopodis 1,2, Vasileios Pierros 1, Dimitrios J. Stravopodis 2 and George T. Tsangaris 1, *

1 Proteomics Research Unit, Biomedical Research Foundation of the Academy of Athens, 11527 Athens, Greece; kontopodisv@hotmail.gr (E.K.); pierrosv@gmail.com (V.P.)
2 Section of Cell Biology and Biophysics, Department of Biology, School of Science, National and Kapodistrian University of Athens, 15701 Athens, Greece; dstravop@biol.uoa.gr
* Correspondence: gthtsangaris@bioacademy.gr; Tel.: +30-210-659-7075

Abstract: The recently discovered Omicron variant of the SARS-CoV-2 coronavirus has raised a new, global, awareness. In this study, we identified the Core Unique Peptides (CrUPs) that reside. exclusively in the Omicron variant of Spike protein and are absent from the human proteome, creating a new dataset of peptides named as SARS-CoV-2 CrUPs against the human proteome (C/H-CrUPs), and we analyzed their locations in comparison to the Alpha and Delta variants. In Omicron, 115 C/HCrUPs were generated and 119 C/H-CrUPs were lost, almost four times as many compared to the other two variants. At the Receptor Binding Motif (RBM), 8 mutations were detected, resulting in the construction of 28 novel C/H-CrUPs. Most importantly, in the Omicron variant, new C/H-CrUPs carrying two or three mutant amino acids were produced, as a consequence of the accumulation of multiple mutations in the RBM. These C/H-CrUPs could not be recognized in any other viral Spike variant. Our findings indicated that the virus binding to the ACE2 receptor is facilitated by the herein identified C/H-CrUPs in contact point mutations and Spike cleavage sites, while the immunoregulatory NF9 peptide is not detectably affected. Thus, the Omicron variant could escape immune-system attack, while the strong viral binding to the ACE2 receptor leads to the highly efficient fusion of the virus to the target cell. However, the intact NF9 peptide suggests that Omicron exhibits reduced pathogenicity compared to the Delta variant.

1. Introduction

The SARS-CoV-2 virus has a high mutagenesis frequency, hitherto producing 63 different variants with 39 considered as the most predominant forms, including Delta, the dominant variant of the 4th pandemic wave [1]. Recently, a new variant, Omicron (B.1.1.529), was identified in South Africa. Omicron is characterized by 30 amino acid changes, three small deletions, and one small insertion in Spike protein, as compared to the original virus, with 15 of them residing in the Receptor Binding Domain (RBD) from 319 to 541 amino acid residues [2]. In our previous studies, we have defined as Unique Peptides (UPs) the peptides whose amino acid sequence appears only in one protein across a given proteome. We also introduced the term of Core Unique Peptides (CrUPs), which are the peptides with a minimum amino acid sequence length that appear only in one protein across a given proteome, thus having a unique signature for a particular protein identification [3]. Therefore, each peptide of any size that contains a CrUP is considered a UP. Peptides of bigger sizes than CrUPs being constructed by continuous CrUPs are considered as Composite Core Unique Peptides (CmUPs). Hitherto, our results regarding the analysis of CrUPs in different species and organisms strongly suggest that CrUPs constitute a concrete group of peptides within a given proteome, with specialized properties and functions Thereby, we have introduced the new term “Uniquome”, which is defined as the total set of UPs belonging to a given proteome and serving as its unique molecular signature. Hence, to map the UP landscape of a proteome under examination, we have herein developed a novel and advanced bioinformatics tool, including big data analysis, and we have applied this tool for the analysis of Uniquome typifying all model organisms. In Homo sapiens, the analysis of the 20,430 reviewed proteins resulted in the identification of 7,263,888 CrUPs which construct the human Uniquome. Most importantly, to elucidate SARS-CoV-2 virus–host organism interactions, we have further designed a novel bioinformatics platform to analyze the Core Unique Peptides (CrUPs) of the SARS-CoV-2 virus against the human proteome (C/H-CrUPs) [1]. C/HCrUPs represent a completely new set of peptides, which are the shortest in length peptides in a viral proteome that do not exist in the human proteome [3]. Based on their properties, the viral C/H-CrUPs could advance our knowledge regarding virus–host interactions, immune system response(s), and infectiveness and pathogenicity of the virus. Moreover, most importantly, they can be used as antigenic and diagnostic peptides, and likely druggable targets for successful therapeutic treatments. In the present study, we have identified, cataloged, and analyzed Omicron-specific C/H-CrUPs in order to illuminate the mechanisms controlling infectivity, immune escape, and pathogenicity of the new variant.

2. Materials and Methods

2.1 Methods

In our previous, recent studies, we developed a bioinformatics tool that can extract the Core Unique Peptides (CrUPs) from a given proteome, thus creating its Uniquome (Figure S1) [1,3]. We have expanded this tool by introducing a new feature that can extract the CrUPs of each individual protein of a given proteome (target) versus the proteins of a reference proteome. This new feature, like the initial implementation, will split each protein in the target proteome to all possible peptides of length minimum (4 amino acids) to length maximum (100 amino acids), and search them against the reference proteome. Each search will exclude all peptides that contain a smaller peptide already identified as CrUP (Figure S2).

For the present study, we have engaged this new feature of our tool. We created a "custom" proteome consisting of sequences from all variants of the SARS-CoV-2 Spike proteins and used it as the target versus the human proteome. The tool produces as output the C/H-CrUPs per protein of the target proteome, thus revealing the CrUPs for each Spike variant versus the human proteome.

Once we obtained the desired data, we ran a meta-analysis to identify how many C/H-CrUPs remained the same or were added or lost on each variant versus the wild- type Spike protein. For this analysis, initially we took the identified C/H-CrUPs of the wild-type sequence and checked their presence against the respective C/H-CrUPs of the other variants. We only cared for the amino acid sequence and not the position this could be found within the protein. If the sequence was found, then we considered the peptide to be the same, otherwise we considered it to be lost on the examined variant. Next, we analyzed the identified C/H-CrUPs of each variant versus the wild-type sequence. If the peptide was detected only on the variant's C/H-CrUPs, then we considered it as added. This meta-analysis also provided us with the position of each C/H-CrUP within the Spike protein, which we used to determine the area (e.g., RBD, RBM and S-cleavage site, as obtained by the Stanford COVID-19 Database) they resided in.

2.2 Databases

All proteomes and proteins were obtained from Uniprot. SARS-CoV-2 wild-type and variant sequences, and mutations were obtained from the Stanford COVID-19 Database (https://covdb.stanford.edu/page/mutation-viewer/, accessed on 23 December 2021).

3 Results and Discussion

3.1 Mapping the C/H-CrUPs Landscape of Spike Protein of the SARS-CoV-2 Omicron Variant

SARS-CoV-2 virus seems to be highly mutated, so far producing more than 60 distinct variants. Hitherto, the highest pathogenic form is the Delta variant (B.1.617.2), with 10 different sub-variants. Recently, a novel variant called Omicron has been identified. It is characterized by 30 amino acid changes, three small deletions, and one small insertion in the Spike protein area, as compared to the wild-type viral respective sequence (Figure S3) [2]. Out of these genetic changes, 15 reside in the Receptor Binding Domain (RBD) from amino acid position 318 to 541, and two are located around the S-cleavage site(s) (Figure S3).

Advanced bioinformatics analysis of the Omicron variant Spike protein showed that it contains 983 C/H-CrUPs, a number that is comparable to the one of wild-type Spike proteins (987 C/H-CrUPs) and to the mean SD value of Spike protein-specific C/H-CrUPs (983 2 C/H-CrUPs) (Table 1). Omicron variant Spike protein contains 34 mutations in total, which is the highest number of identified mutations among all virus variants.

Table 1. SARS-CoV-2 Spike protein C/H-CrUPs across variants, as compared to the wild-type virus respective sequence.

		Spike Protein					
Variant	**C/H-CrUPs**	**Same C/H-CrUPs**	**% of Same C/H-CrUPs**	**New C/H-CrUPs**	**% of New C/H-CrUPs**	**Lost C/H-CrUPs**	**% of Lost C/H-CrUPs**
Wild-type virus	987						
Alpha (B.1.1.7) + (Q1-Q4)	982	931	94.8	51	5.2	56	5.7
Alpha (B.1.1.7 + E484K)	983	928	94.4	55	5.6	59	6.0
Alpha (B.1.1.7 + L452R)	981	936	95.4	45	4.6	51	5.2
Alpha (B.1.1.7 + S494P)	981	936	95.4	45	4.6	51	5.2
Beta (B.1.351)	981	954	97.2	27	2.8	33	3.3
Beta (B.1.351 + E516Q)	981	949	96.7	32	3.3	38	3.8
Beta (B.1.351 + L18F) (B.1.351.2-3)	979	948	96.8	31	3.2	39	3.9
Beta (B.1.351 + P384L)	980	949	96.8	31	3.2	38	3.9
Gamma (P.1) (P.1.1 - P.1.2)	985	930	94.4	55	5.6	57	5.8
Gamma (P1 + P681H)	985	930	94.4	55	5.6	57	5.8
Delta (B.1.617.2)	984	948	96.3	36	3.7	39	4.0
Delta (B.1.617.2 + E484Q)	984	945	96.0	39	3.4	42	4.3
Delta (B.1.617.2 + K417N)	984	944	95.9	40	4.1	43	4.4
Delta (B.1.617.2 + Q613H)	984	947	96.2	37	3.8	40	4.1
Delta (AY.1)	984	944	95.9	40	4.7	43	4.1
Delta (AY.2)	985	939	95.3	46	4.8	48	4.9
Delta (AY.3 - AY.8) + (AY.12)	983	951	96.7	32	3.3	36	3.7
Delta (AY.9)	983	951	96.7	32	3.3	36	3.6
Delta (AY.10)	983	951	96.7	32	3.3	36	3.6

Delta (AY.11)	983	951	96.7	32	3.3	36	3.6
Eta (B.1.525)	990	956	96.5	34	3.4	31	3.1
Iota (B.1.526)	984	960	97.5	24	2.4	27	2.7
Kappa (B.1.617.1)	985	964	97.8	21	2.1	23	2.3
Lambda (C.37)	982	949	96.6	33	3.4	38	3.9
Mu (B.1.621)	983	953	96.9	30	3.1	34	3.4
Omicron (B.1.1.529)	983	868	88.3	115	11.7	119	12.1

New C/H-CrUPs is the number of new constructed peptides of each variant compared to C/H-CrUPs of wild-type virus; **% of new C/H-CrUPs** is the % of new constructed peptides compared to the total C/H-CrUPs number of each variant; **Lost C/H-CrUPs** is the number of peptides lost in each variant compared to C/H-CrUPs of wild-type virus; **% of lost C/H-CrUPS** is the % of lost peptides compared to the total C/H-CrUPs number of each variant.

These mutations seem to have a dramatic effect on the Spike protein C/H-CrUPs map. Compared to the wild-type Spike sequence, we found that 115 (new) C/H-CrUPs were created and 119 C/H-CrUPs were lost, almost twice as many when compared to the Alpha variant (51 and 56 C/H-CrUPs, respectively), and almost four times as many, compared to the other variants (Table 1). The distribution of these new C/H-CrUPs shows that the majority carry 6 amino acids in length (Figure S4).

3.2 Omicron-Specific C/H-CrUPs That belong to the Receptor Binding Domain

SARS-CoV-2 belongs to the β coronavirus group, which uses the plasma membrane receptor of Angiotensin-Converting Enzyme 2 (ACE2) to recognize and bind to the target cell [4]. The viral Spike protein attaches to ACE2 receptor by a Receptor Binding Domain (RBD) defined from amino acid position F318 to F541 [4,5]. The amino acid residues from W436 to Q506 inside RBD shape the Receptor Binding Motif (RBM), which carries 11 contact positions with ACE2 [5]. The RBD region has received great attention, as it seems to be a major target of antibodies against the virus and other therapeutic interventions [6–8].

In the RBD region, the Omicron variant carries 15 mutations, 10 of which are identified in the RBM area (Figure 1A). This results in the identification of the highest number of newly constructed C/H-CrUPs in the RBD/RBM region, as compared to all other previous virus variants examined (Table S1). Table 2 describes all the new, herein identified, C/H-CrUPs of Omicron variant in Spike's RBD region, in comparison to the Alpha and Delta variants, which represent two of the most predominant variants of the virus in human populations. Hence, it was proven that, in contrast to Alpha and Delta variants, at the end of Omicron variant RBM area from 440 to 508 amino acid position, 8 novel mutations were identified, resulting in the production of 28 new C/H-CrUPs. The most important

finding is that in Omicron variant, for the first time, new C/H-CrUPs including two or three mutant amino acids were generated, with the peptides "*QAGN*K*P*", "*N*K*PCN*","*LK*SYS*F*" and "*K*SYS*FR**" being characteristic examples, as a result of the accumulation of multiple mutations in the positions 440, 446, 477, 478 and 493–505. These novel C/H- CrUPs that contain several mutated amino acids could not be found in any other virus variants previously. Taking into consideration recent data about virus infectivity, the multimutated, new, C/H-CrUP collection seems to radically change the structure and the epitope regions of end positions of the RBM area in the Omicron variant, causing a serious compromise of its antigenic capacity and facilitating the immune escape of the virus [9].

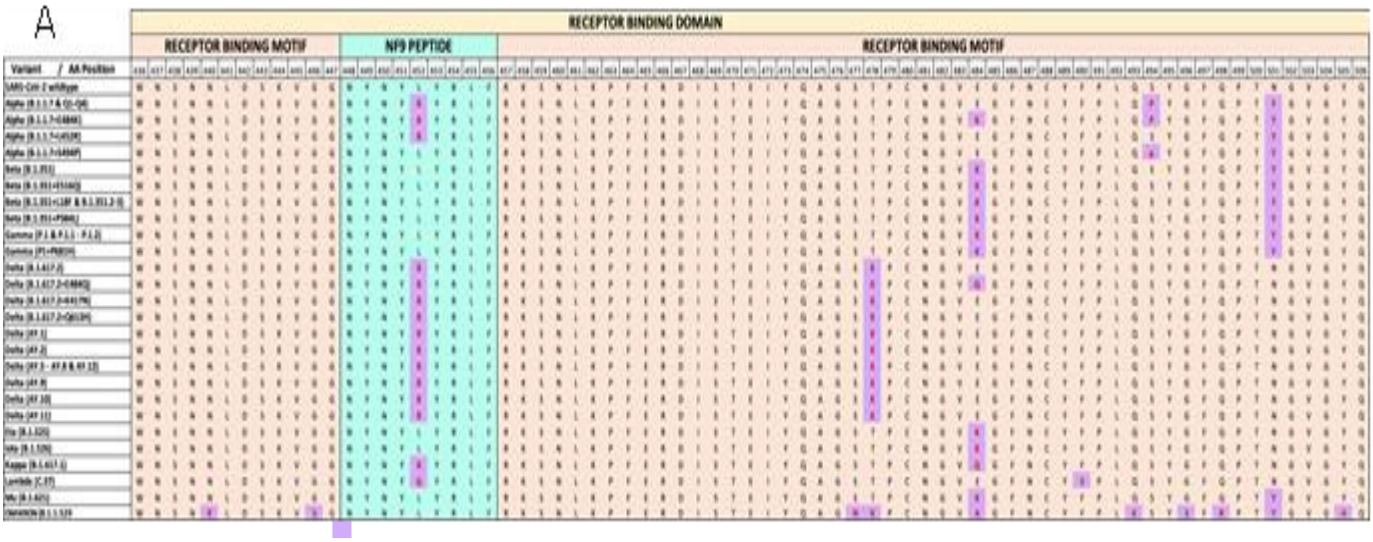


Figure 1. Mutations in different virus variants. (**A**) The mutations of the Receptor-Binding Motif (RBM) included in the Receptor-Binding Domain (RBD) are presented. (**B**) The mutations around the Spike cleavage sites are presented. Purple blocks mark the point mutation sites in the variants. Green colors indicate the Universal Peptides of the Spike proteins from Figure S2. Yellow colors mark the Receptor-Binding Domain of Spike protein interaction with ACE2. Pink colors mark the Receptor-Binding Motif. Cyan colors indicate the NF9 peptide, while light blue colors mark the Bridge between S1 and S2 domains. Red-colored arrows indicate the cleavage sites. With different colors in the upper side of the alignment, the different domains of the Spike protein are presented.

Table 2. C/H-CrUPs constructed aroud the mutations in RBD of Alpha, Delta and Omicron SARS-CoV-2 variants.

Alpha Variant					Delta Variant				
C/H-CrUP	**Position**	**Mutation**	**New C/H-CrUPs**	**Position**	**C/H-CrUP**	**Position**	**Mutation**	**New C/H-CrUPs**	**Position**
GNYNYL	447		GNYNYR	447	PGQTGKI	412			
NYNYLY	448				GQTGKIA	413		GQTGNI	413
			YNYRY	449			K417N	QTGNIA	414
NYLYRL	450	L452R	NYRYRL	450	TGKIAD	415		TGNIAD	415
YLYRLF	451		YRYRLF	451	GKIADY	416		GNIADY	416
LYRLFR	452								
					GNYNYL	447		GNYNYR	447
CNGVEG	480		CNGVKG	480	NYNYLY	448			

Omicron variant									
C/H-CrUP	**Position**	**Mutation**	**New C/H-CrUPs**	**Position**	**C/H-CrUP**	**Position**	**Mutation**	**New C/H-CrUPs**	**Position**
NLCPFG	334		NLCPFD	334	IYQAGS	472			
LCPFGE	335		LCPFDE	335	YQAGST	473		YQAGN	473
PFGEVF	337	G339D	PFDEV	337			S477N	QAGNKP	474
							T478K		
FGEVFN	338		FDEVFN	338	AGSTPC	475			
GEVFNA	339		DEVFNA	339	STPCN	477		NKPCN	477
								KPCNG	478
VLYNSA	367		VLYNLAP	367					
LYNSAS	368				CNGVEG	480		CNGVAG	480
SFSTFK	373				CYFPLQ	488		CYFPLK	488
			FFTFK	374	YFPLQS	489		YFPLKS	489
STFKC	375		FTFKCY	375	FPLQSY	490		FPLKSY	490
					PLQSYG	491	Q493K		
PGQTGKI	412						G496S	LKSYSF	492
GQTGKIA	413		GQTGNI	413	QSYGF	493	Q498R N501Y	KSYSFR	493
		K417N	QTGNIA	414	SYGFQP	494	Y505H	SYSFRP	494
TGKIAD	415		TGNIAD	415	YGFQPT	495		YSFRPT	495
GKIADY	416		GNIADY	416	GFQPTN	496			
					FQPTNG	497		FRPTY	497
WNSNN	436		WNSNKL	436	QPTNGV	498		RPTYGV	498
SNNLDS	438		SNKLDS	438	PTNGVG	499			
NNLDSK	439		NKLDSKV	439	TNGVGY	500		TYGVGH	500
			KLDSKVS	440	NGVGYQ	501			
LDSKVG	441	N440K G446S			GVGYQP	502		GVGHQ	502
DSKVGG	442		DSKVSG	442	VGYQPY	503		VGHQPY	503
KVGGNY	444		KVSGNY	444	GYQPYR	504			
VGGNYN	445		VSGNYN	445	YQPYRV	505		HQPYR	505
GGNYNY	446								

The original and newly constructed C/H-CrUPs around the native and mutant sites of RBD region of SARS-CoV-2 Spike protein in Alpha, Delta and Omicron variants are presented. With the red colors, the mutant amino acids in wild-type C/H-CrUPs and in the newly constructed peptides are marked.

Remarkably, RBM area contains 11 out of the 12 contact points of viral Spike protein with the ACE2 cellular receptor. Among them, 7 contact points remained intact, while 4 mutations in positions Q493K, Q498R, N501Y and Y505H were identified, resulting in the construction of 17 new C/H-CrUPs (Table 3). N501Y mutation was found to be a major determinant of increased viral transmission, due to the improved binding affinity of Spike

protein to ACE2 cellular receptor [10]. These findings indicate that virus binding to ACE2 receptor is notably affected by C/H-CrUP-specific mutations that can likely strengthen Spike-ACE2 protein–protein interaction(s).

Table 3. C/H-CrUPs around SARS-CoV-2 RBD contact positions.

WILD-TYPE							OMICRON VARIANT					
Contact Positions	C/H-CrUPs						Mutations	Newly Constructed C/H-CrUPs				
N439	AWNSN	WNSNN	SNNLDS	NNLDSK								
Y449	KVGGNY	VGGNYN	GGNYNY	GNYNYL	NYNYLY							
Y453	NYNYLY	NYLYRL	YLYRLF	LYRLFR	YRLFRK							
F486	NGVEGF	GVEGFN	GFNCY	FNCYF								
N487	GVEGFN	GFNCY	FNCYF									
Y489	GFNCY	FNCYF	CYFPLQ	YFPLQS								
Q493	CYFPLQ	YFPLQS	FPLQSY	PLQSYG	QSYGF		Q493K	CYFPLK	YFPLKS	FPLKSY	LKSYSF	KSYSFR
Q498	SYGFQP	YGFQPT	GFQPTN	FQPTNG	QPTNGV		Q498R	KSYSFR	SYSFRP	YSFRPT	FRPTY	RPTYGV
T500	YGFQPT	GFQPTN	FQPTNG	QPTNGV	PTNGVG	TNGVGY						
N501	GFQPTN	FQPTNG	QPTNGV	PTNGVG	TNGVGY	NGVGYQ	N501Y	FRPTY	RPTYGV	TYGVGH		
Y505	TNGVGY	NGVGYQ	GVGYQP	VGYQPY	GYQPYR	YQPYRV	Y505H	TYGVGH	GVGHQ	VGHQPY	HQPYR	

The original and newly constructed C/H-CrUPs residing around the native and contact positions of the SARS-CoV-2 Spike protein RBD region. The C/H-CrUPs of wild-type and Omicron variant are presented. With red colors, the mutant amino acids in wild-type C/H-CrUPs and in the newly constructed peptides are marked.

Interestingly, an important amino acid sequence in the RBM area is the "*NYNYLYRLF*" peptide (from 448 to 456 position). This Tyrosine (Y)-enriched peptide carries two contact sites (Y449 and Y453), and it is known as the NF9 peptide [11]. It seems to affect antigen recognition, by being an immunodominant HLA*24:02-restricted epitope identified by CD8+ T cells. Of note, NF9 presents immune stimulation activity, and increases cytokine production derived from CD8+ T cells, such as IFN-γ, TNF-α and IL-2 [12]. In contrast to Delta, in the Omicron variant the NF9 amino acid content is not changed by any mutation detected, thus suggesting that the NF9 peptide could induce early immune system activation and efficient cytokine production, leading to a faster immune response, and thus reducing SARS-CoV-2 virus pathogenicity.

3.3 C/H-CrUPs Altered Architecture around the Spike-Cleavage Site(s) of the Omicron Variant

The molecular mechanism of Spike protein's proteolytic activation has been shown to play a crucial role in the selection of host species, virus–cell fusion, and the viral infection of human lung cells [13–15]. Spike protein [SPIKE_SARS2 (P0DTC2)] contains three cleavage sites (known as S-cleavage sites) crucial for the virus fusion to the host cell: the

R685↓S and R815↓S positions that serve as direct targets of the Furin protease, and the T696↓M position that can be recognized by the TMPRSS2 protease [16–18].

In these cleavage sites, the Omicron variant carries only the critical mutation P681H, which also appears in the Alpha variant (Figure 1B). Strikingly, in contrast to the Delta variant, which contains the P681R mutation, the P681H mutation constructs several new C/H-CrUPs in the Alpha and Omicron variants, thus indicating their dispensable contribution to virus fusion to the host cell (Table 4).

Table 4. C/H-CrUPs arround the Spike protein cleavage sites.

Cleavage Site	Mutation	Variant	Position	New C/H-CrUPs
R^{685}↓S	P681R	Delta	680	SRRRAR↓S
	P681H	Alpha Omicron	677	QTNSH
			678	TNSHR
			680	SHRRAR
T^{696}↓M	A701V	Beta	None	
R^{815}↓S	None	None	None	

C/H-CrUPs around the mutant positions of Spike protein cleavage sites are presented. Symbol "↓" indicates the protein cleavage positions.

4 Conclusions

Core Unique Peptides constitute a distinct and important group of peptides within a proteome. The identification of CrUPs in an organism (e.g., virus, microbe, or mutant protein) against a distinct proteome of another organism is a completely novel approach, which could prove useful for the understanding of the action of microorganisms, the association of novel pharmacological targets with therapies, and the design of novel vaccines. It could be employed in many different kinds of diseases, such as cancer, athropozoans diseases, the design of vaccines for pathogenic viruses, and the identification of new antigenic epitopes capable for the development of new diagnostic or therapeutic antibodies. Therefore, we applied this dynamic and novel strategy, for the first time, in the identification of CrUPs derived from SARS-CoV-2 against the human proteome [1]. In that study, we analyzed all the CrUP peptides of all SARS-CoV-2 variants against the proteome of the host organism, which in our case was Human sapiens. Remarkably, this approach clearly revealed the immune escaping capacity, the contagious power and the high pathogenicity of Delta variant, in contrast to other variants. Notably, these findings have been confirmed by epidemiological data concerning the course of the disease.

In the present study, we engaged this approach to the analysis of the SARS-CoV-2 Omicron variant. The analysis of C/H-CrUP landscapes in the heavily mutated SARS-CoV- 2 Omicron variant Spike protein unveiled that the Omicron variant, by the generation of novel multi-mutated C/H-CrUPs, could escape the immune system defense mechanisms, while these C/H-CrUP-specific mutations could facilitate more efficient virus binding to the ACE2 cellular receptor, and a more productive fusion of the virus to the host cell. Most importantly, in contrast to the Delta variant, the intact NF9 peptide in the Omicron variant, which has a known immunostimulatory effect, suggests that Omicron exhibits reduced pathogenicity as compared to Delta.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/vaccines10030357/s1. Figure S1: Uniquome creation algorithm; Figure S2: Extracting CrUPs of Target vs Reference Proteome; Figure S3: Alignment of the SARS- CoV-2 Spike protein (SPIKE_SARS2, P0DTC2) of the 26 variants, together with the wild-type Spike Protein (SPIKE_SARS2, P0DTC2); Figure S4: Length distribution of Omicron variant Spike protein C/H-CrUPs; Table S1: New C/H-CrUPs located in the RBD and RBM regions of the Spike protein across virus variants.

Author Contributions: Conceptualization, V.P., E.K. and G.T.T.; methodology, V.P. and E.K.; investiga- tion, V.P., E.K. and G.T.T.; visualization, E.K., D.J.S. and G.T.T.; supervision, G.T.T.; writing—original draft, D.J.S. and G.T.T.; writing—review and editing, D.J.S. and G.T.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data of the present article are available in the main text or in the supplementary materials.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

UPs: Unique Peptides; CrUPs: Core Unique Peptides; C/H-CrUPs: SARS-CoV-2 Core Unique Peptides against Human Proteome; RBD: Receptor Binding Domain; RBM: Receptor Binding Motif.

References

1. Pierros, V.; Kontopodis, E.; Stravopodis, D.J.; Tsangaris, G.T. Unique Peptide Signatures of SARS-CoV-2 Against Human Proteome Reveal Variants' Immune Escape and Infectiveness. *bioRxiv* 2021. [CrossRef]

2. Wang, L.; Cheng, G. Sequence analysis of the emerging SARS-CoV-2 variant Omicron in South Africa. *J. Med. Virol.* 2022, *94*, 11728–11733. [CrossRef] [PubMed]

3. Kontopodis, E.; Pierros, V.; Anagnostopoulos, A.; Stravopodis, D.J.; Papassideri, I.; Vorgias, C.; Tsangaris, G.T. Data Processing Approach for the Construction and Evaluation of an Organism's UNIQUOME with Comparative Analysis for the Human, Rat and Mouse Uniquomes. In Proceedings of the XIII. Annual Congress of the European Proteomics Association: From Genes via Proteins and their Interactions to Functions, Potsdam, Germany, 24–28 March 2019.

4. Shang, J.; Ye, G.; Shi, K.; Wan, Y.; Luo, C.; Aihara, H.; Geng, Q.; Auerbach, A.; Li, F. Structural basis of receptor recognition by SARS-CoV-2. *Nature* 2020, *581*, 221–224. [CrossRef] [PubMed]

5. Hatmal, M.M.; Alshaer, W.; Al-Hatamleh, M.A.I.; Hatmal, M.; Smadi, O.; Taha, M.O.; Oweida, A.J.; Boer, J.C.; Mohamud, R.; Plebanski, M. Comprehensive Structural and Molecular Comparison of Spike Proteins of SARS-CoV-2, SARS-CoV and MERS-CoV, and Their Interactions with ACE2. *Cells* 2020, *9*, 2638. [CrossRef] [PubMed]

6. Chen, Y.; Zhang, Y.N.; Yan, R.; Wang, G.; Zhang, Y.; Zhang, Z.-R.; Li, Y.; Ou, J.; Chu, W.; Liang, Z.; et al. ACE2-targeting monoclonal antibody as potent and broad-spectrum coronavirus blocker. *Signal Transduct. Target. Ther.* 2021, *6*, 315. [CrossRef] [PubMed]

7. Zahradnik, J.; Marciano, S.; Shemesh, M.; Zoler, E.; Harari, D.; Chiaravalli, J.; Meyer, B.; Rudich, Y.; Li, C.; Marton, I.; et al. SARS-CoV-2 variant prediction and antiviral drug design are enabled by RBD in vitro evolution. *Nat. Microbiol.* 2021, *6*, 1188–1198. [CrossRef] [PubMed]

8. Hastie, K.M.; Li, H.; Bedinger, D.; Schendel, S.L.; Dennison, S.M.; Li, K.; Rayaprolu, V.; Yu, X.; Mann, C.; Zandonatti, M.; et al. Defining variant-resistant epitopes targeted by SARS-CoV-2 antibodies: A global consortium study. *Science* 2021, *374*, 472–478. [CrossRef] [PubMed]

9. Chen, J.; Wang, R.; Gilby, N.B.; Wei, G.-W. Omicron Variant (B.1.1.529): Infectivity, Vaccine Breakthrough, and Antibody Resistance. *J. Chem. Inf. Model.* 2022, *62*, 412–422. [CrossRef] [PubMed]

10. Liu, Y.; Liu, J.; Plante, K.S.; Plante, J.A.; Xie, X.; Zhang, X.; Ku, Z.; An, Z.; Scharton, D.; Schindewolf, C.; et al. The N501Y spike substitution enhances SARS-CoV-2 infection and transmission. *Nature* 2022, *602*, 294–299. [CrossRef] [PubMed]

11. Motozono, C.; Toyoda, M.; Zahradnik, J.; Saito, A.; Nasser, H.; Tan, S.T.; Ngare, I.; Kimura, I.; Uriu, K.; Kosugi, Y.; et al. SARS-CoV-2 spike L452R variant evades cellular immunity and increases infectivity. Cell Host Microbe 2021, 29, 1124–1136.e11. [CrossRef] [PubMed]

12. Kared, H.; Redd, A.D.; Bloch, E.M.; Bonny, T.S.; Sumatoh, H.; Kairi, F.; Carbajo, D.; Abel, B.; Newell, E.W.; Bettinotti, M.P.; et al. SARS-CoV-2-specific CD8+ T cell responses in convalescent COVID-19 individuals. J. Clin. Investig. 2021, 131, 1124–1136.e11. [CrossRef] [PubMed]

13. Peacock, T.P.; Goldhill, D.H.; Zhou, J.; Baillon, L.; Frise, R.; Swann, O.C.; Kugathasan, R.; Penn, R.; Brown, J.C.; Sanchez-David, R.Y.; et al. The furin cleavage site in the SARS-CoV-2 spike protein is required for transmission in ferrets. Nat. Microbiol. 2021, 6, 899–909. [CrossRef] [PubMed]

14. Whittaker, G.R. SARS-CoV-2 Spike and its Adaptable Furin Cleavage Site. Lancet Microbe 2021, 2, e488–e489. [CrossRef]

15. Shang, J.; Wan, Y.; Luo, C.; Ye, G.; Geng, Q.; Auerbach, A.; Li, F. Cell entry mechanisms of SARS-CoV-2. Proc. Natl. Acad. Sci. USA 2020, 117, 11727–11734. [CrossRef] [PubMed]

16. Hoffmann, M.; Kleine-Weber, H.; Schroeder, S.; Krüger, N.; Herrler, Y.; Erichsen, S.; Schiergens, T.S.; Herrler, G.; Wu, N.-H.; Nitsche, A.; et al. SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. Cell 2020, 181, 271–280.e8. [CrossRef] [PubMed]

17. Hoffmann, M.; Kleine-Weber, H.; Pohlmann, S. A Multibasic Cleavage Site in the Spike Protein of SARS-CoV-2 Is Essential for Infection of Human Lung Cells. Mol. Cell 2020, 78, 779–784.e5. [CrossRef] [PubMed]

18. Takeda, M. Proteolytic activation of SARS-CoV-2 spike protein. Microbiol. Immunol. 2022, 66, 15–23. [CrossRef] [PubMed]

Unique peptide signatures of SARS-CoV-2 virus against human proteome reveal variants' immune escape and infectiveness.

Vasileios Pierros [a,1], Evangelos Kontopodis [a,b,1], Dimitrios J. Stravopodis [b], George Th. Tsangaris [a,*]

[a] *Proteomics Research Unit, Biomedical Research Foundation of the Academy of Athens, 11527, Athens, Greece*

[b] *Section of Cell Biology and Biophysics, Department of Biology, School of Science, National and Kapodistrian University of Athens, 15701, Athens, Greece*

* *Corresponding author.*

[1] *These authors contributed equally to this work.*

Abstract: SARS-CoV-2 pandemic has necessitated the identification of sequence areas in the viral proteome that are capable to serve as antigenic sites and treatment targets. In the present study, we have applied a novel approach for mechanistically illuminating the virus-host organism interactions, by analyzing the Unique Peptides (UPs) of the virus featured by a minimum amino acid sequence length being defined as Core Unique Peptides (CrUPs), not of the virus *per se*, but against the entire proteome of the host organism. This approach resulted in the identification of CrUPs of the virus itself, which could not be recognized in the host organism proteome. Thereby, we analyzed the SARS-CoV-2 proteome for identification of CrUPs against the human proteome, which have been defined as C/ H-CrUPs. We herein reveal that SARS-CoV-2 include 7.503 C/H-CrUPs, with the SPIKE_SARS2 being detected as the protein with the highest density of C/H-CrUPs. Extensive analysis has indicated that the critical P681R mutation produces new C/H-CrUPs around the R685 cleavage site, while the L452R mutation causes loss of antigenicity of the NF9 peptide and strong(er) binding of the virus to its ACE2 receptor protein. Simultaneous formation of these mutations in detrimental variants like Delta leads to the immune escape of the virus, its massive entrance into the host cell, a notable increase in virus formation, and its massive release and thus elevated infectivity of human target cells.

1. Introduction

Covid-19 pandemic has emerged the urgent necessity of the identification of sequence sites of the SARS-CoV-2 viral proteome that can serve as appropriate treatment targets and antigenic positions suitable for production of therapeutic vaccines.

As we have recently described, a Unique Peptide (UP) is defined as the peptide carrying an amino acid sequence that appears only in one of all proteins in a particular proteome. To this direction, our team has also introduced, for the first time, the concept of Core Unique Peptide (CrUP), which represents the peptide bearing a minimum length of amino acid sequence that resides solely in one of all proteins in a profiled proteome, thereby rendering it a unique signature for identification and differential recognition of a given protein (Alexandridou et al., 2009; Kontopodis et al., 2019). Hence, to thoroughly map the UP-specific landscape of a proteome of interest, we have developed a novel bioinformatics tool that is based on advanced algorithms being dedicated to big-data analysis. Its engagement to deep and accurate processing of the 20.430 reviewed Homo sapiens (human) proteins led to the recognition and identification of more than $7X10^6$ CrUPs, which represent the backbone of human Uniquome that is mainly described as the voluminous collection of UPs shaping the human proteome (Kontopodis et al., 2022 and Kontopodis et al. manuscript in preparation).

Most importantly, to further illuminate the mechanisms controlling virus-host interactions, we have recently developed a novel, dynamic and advanced bioinformatics platform to thoroughly analyze and compare virus-derived CrUPs against host-organism proteome(s). This unique collection contains peptides that notably differ from the virus-specific CrUPs themselves, with each one of them being described as the pep- tide carrying an amino acid sequence of minimum length that is accommodated exclusively in one out of all proteins throughout the viral proteome. This virus against host CrUPs bear two cardinal properties: first, they are unique in virus proteome and, second, they do not exist in host-organism proteome. Therefore, the virus against host proteome derived CrUPs can advance our knowledge and understanding of virus- host interactions, and virus infectiveness and pathogenicity dynamics. Furthermore, they can be used as diagnostic and antigenic peptides, and likely therapeutic targets, as well. Altogether, these CrUPs seem to represent a completely new entity of peptides capable to significantly improve our view and comprehension regarding the structuring, functioning and mapping of virus and human Uniquomes, and their proteo-

mic “cross-talks”, towards immune escape and infectiveness (Kontopodis et al., 2022).

Since human cells can host the SARS-CoV-2 virus, we have herein engaged our novel bioinformatics platform not only for the profiling of CrUPs in the SARS-CoV-2 proteome per se, but, most importantly, for their identification against the human proteome (C/H-CrUPs). Remark- ably, C/H-CrUPs can likely serve as targets for the immune response upon infection, and antigenic sites with major pharmaceutical and diagnostic potential, for the successful clinical management of Covid-19 pandemic.

Table 1. Viral CrUPs against Human proteome (C/H-CrUPs).

VIRUS	Proteins (number)	Total number of AA	Total C/H-CrUPs (number)	C/H-CrUPs appeared 1 time (number)	C/H-CrUPs appeared 2 times (number)	C/H-CrUPs appeared 3 times (number)	C/H-CrUPs Density
SARS-CoV-2	16	14.401	7.503	4.213	3.289	1	75%
SARS-CoV	15	14.396	7.534	4.236	3.298	0	75%
MERS	10	14.216	7.413	4.077	3.336	0	76%

Viral proteomes of the β coronavirus group SARS-CoV-2, SARS-CoV and MERS-CoV were analyzed for core unique peptides (CrUPs) against the human proteome. The identified CrUPs of each virus against the human proteome are presented (C/H-CrUPs). C/H-CrUPs were further analyzed for the times by which they appear in each viral proteome. C/H-CrUP density is defined as the percentage of total amino acids contained in C/H-CrUPs of each virus to the total number of the virus amino acids.

2. Results and discussion

2.1. SARS-CoV-2 core unique peptides against human proteome

The SARS-CoV-2 proteome is structurally quite simple. In the UNI- PROT database (version 7/2021), 16 reviewed and 75.714 unreviewed proteins have been included (Jungreis et al., 2021). For the present study, only the 16 reviewed proteins are examined, since the unreviewed proteome components contain (among others) duplicate registrations, and unverified sequences and protein fragments, which could lead to unreliable data regarding the uniqueness of a protein sequence. To recognize all the CrUPs being embraced in the SARS-CoV-2 proteome against the human proteome, we in silico constructed a new, artificial, “hybrid-proteome” that contained all the reviewed human proteins (20.430 proteins), plus the one protein derived from the SARS-CoV-2 viral proteome (20.431 proteins). Thus, 16 “hybrid proteomes” including the 16 SARS-CoV-2 proteins were constructed. Hence, these “hybrid proteomes” were bioinformatically searched one by one for the identification of SARS-CoV-2-specific CrUPs in human protein sequence environments (C/H-CrUPs).

Strikingly, 7.503 C/H-CrUPs were detected, with 4.213 of them being presented one time in the SARS-CoV-2 proteome, 3.289 being observed two times in the viral proteome and only one peptide (“VNNATN”) with a 6 amino acid length being recognized three times (Table 1). Data processing and analysis unveiled that C/H-CrUPs retain a length range from 4 to 9 amino acids, while longer peptides could not be identified in the SARS-

CoV-2 virus proteome. Length distribution showed that the majority of C/H-CrUPs have a 6 amino acid length, whereas only one with 4 amino acids and only two with 9 amino acids C/H-CrUPs were observed (Figure 1).

The distribution of C/H-CrUPs across SARS-CoV-2 proteins demonstrated that the Replicase Polyprotein 1ab (R1AB_SARS2), which is the longest viral protein consisted of 7.096 amino acids, produces almost half of the identified C/H-CrUPs (5.334; 49,3%) (Table 2). On the other hand, the Putative ORF3b protein (ORF3B_SARS2), with a length of 22 amino acids, produces only 15 C/H-CrUPs that show a protein density of 68%. Notably, Spike glycoprotein (SPIKE_SARS2) is presented with the highest C/H-CrUPs density (78%), thus indicating its intriguing feature to carry the highest number of C/H-CrUPs (987), in terms of their physical length, as opposed to the ORF3c protein (ORF3C_SARS2), which is characterized by a respective density of only 56% (Table 2). A typical example for the construction of C/H-CrUPs is the peptide “PDEDEEEGD”. This peptide is a 9 amino acid in length C/H–CrUP that belongs to Replicase polyprotein 1a (R1A_SARS2), starting at position 927 and ending at position 935 (Figure 2). Around this peptide, 8 C/H-CrUPs were recognized with a 5–7 amino acid length range.

A

Viruses CrUPs length against Human proteome

CrUP length (AA number)	VIRUS		
	SARS-CoV 2	SARS-CoV	MERS-CoV
4	1	2	3
5	952	947	1.039
6	5.590	5.595	5.429
7	931	966	923
8	27	23	18
9	2	1	1

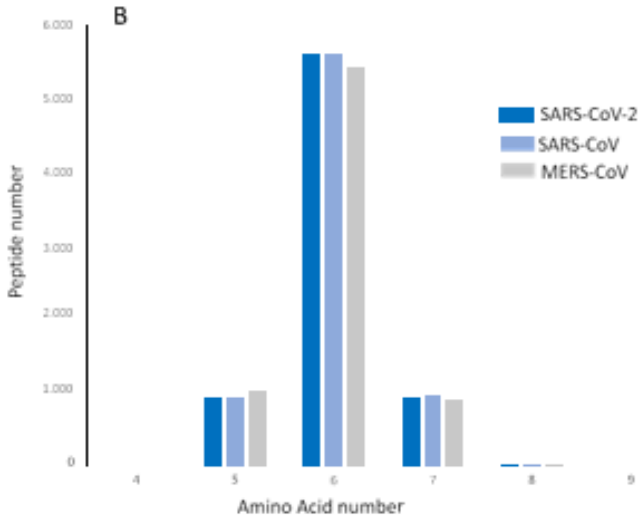


Figure 1. Amino acid length distribution of virus Core Unique Peptides (CrUPs) against human proteome. A) Set of CrUPs derived from SARS-CoV-2, SARS-CoV and MERS-CoV viruses against the human proteome. The CrUPs were identified, listed and grouped according to their amino acid length. B) Graphical presentation of CrUPs amino acid length across β coronavirus group.

Table 2. Virus detailed analysis.

SARS-CoV-2					
Entry ID	Entry Name Length (AA number) C/H-CrUPs Density	Protein C/H-CrUPs	Name (number)		
P0DTD1	R1AB_SARS2	Replicase polyprotein 1ab	7096	5334	75%
P0DTC1	R1A_SARS2	Replicase polyprotein 1a	4405	3294	75%
P0DTC2	SPIKE_SARS2	Spike glycoprotein	1273	987	78%
P0DTC9	NCAP_SARS2	Nucleoprotein	419	308	74%
P0DTC3	AP3A_SARS2	ORF3a protein	275	210	76%
P0DTC5	VME1_SARS2	Membrane protein	222	171	77%
P0DTC7	NS7A_SARS2	ORF7a protein	121	90	74%
P0DTC8	NS8_SARS2	ORF8 protein	121	82	68%
P0DTD2	ORF9B_SARS2	ORF9b protein	97	69	71%
P0DTD3	ORF9C_SARS2	Putative ORF9c protein	73	50	68%
P0DTC4	VEMP_SARS2	Envelope small membrane protein	75	48	64%
P0DTC6	NS6_SARS2	ORF6 protein	61	44	72%
P0DTG0	ORF3D_SARS2	Putative ORF3d protein	57	40	70%
P0DTD8	NS7B_SARS2	ORF7b protein	43	29	67%
P0DTG1	ORF3C_SARS2	ORF3c protein	41	23	56%
P0DTF1	ORF3B_SARS2	Putative ORF3b protein	22	15	68%
SARS-CoV					
Entry ID	Entry Name	Protein Name	Length (AA number)	S/H-CrUPs (number)	S/H-CrUPs Density
P0C6X7	R1AB_SARS	Replicase polyprotein 1ab	7.073	5.346	76%
P0C6U8	R1A_SARS	Replicase polyprotein 1a	4.382	3.301	75%
P59594	SPIKE_SARS	Spike glycoprotein	1.275	970	76%
P59595	NCAP_SARS	Nucleoprotein	422	319	76%
P59632	AP3A_SARS	ORF3a protein	274	208	76%
P59596	VME1_SARS	Membrane protein	221	162	73%
P59633	NS3B_SARS	ORF3b protein	154	113	73%
P59635	NS7A_SARS	ORF7a protein	122	93	76%
P59636	ORF9B_SARS	ORF9b protein	98	71	72%
Q80H93	NS8B_SARS	ORF8b protein	84	59	70%
P59637	VEMP_SARS	Envelope small membrane protein	75	47	63%
Q7TLC7	Y14_SARS	Uncharacterized protein 14	70	45	64%
P59634	NS6_SARS	ORF6 protein	63	44	70%
Q7TFA1	NS7B_SARS	Protein non-structural 7b	44	27	61%
Q7TFA0	NS8A_SARS	ORF8a protein	39	27	69%
MERS					
Entry ID	Entry Name	Protein Name	Length (AA number)	M/H-CrUPs (number)	M/H-CrUPs Density
K9N7C7	R1AB_MERS1	Replicase polyprotein 1ab	7.078	5.364	76%
K9N638	R1A_MERS1	Replicase polyprotein 1a	4.391	3.338	76%
K9N5Q8	SPIKE_MERS1	Spike glycoprotein	1.353	1.024	76%
K9N4V7	NCAP_MERS1	Nucleoprotein	411	301	73%
K9N643	ORF4B_MERS	Non-structural protein ORF4b	246	185	75%
K9N7D2	ORF5_MERS1	Non-structural protein ORF5	224	169	75%
K9N7A1	VME1_MERS1	Membrane protein	219	158	72%
K9N4V0	ORF4A_MERS1	Non-structural protein ORF4a	109	77	71%
K9N796	ORF3_MERS1	Non-structural protein ORF3	103	74	72%
K9N5R3	VEMP_MERS1	Envelope small membrane protein	82	59	72%

Analysis of the SARS-CoV-2, SARS-CoV and MERS-CoV virus is presented. All viruses' proteins have been in silico analyzed and each protein is shown by its Entry-ID, Entry Name and Protein Name according to the UNIPTOT database. The amino acid length of each protein and the number along with density of CrUPs per protein against the human proteome are shown. Density is defined as the percentage of total amino acids contained in CrUPs of each protein to the total number of the protein's amino acids.

2.2. Comparative analysis of SARS-CoV-2, SARS-CoV and MERS-CoV core unique peptides against human proteome

In order to illuminate the mechanisms orchestrating the differen- tial pathologies of SARS-CoV-2 compared to other coronavirus family members, we, next, applied the same strategy to other two similar viruses, the Severe Acute Respiratory Syndrome CoronaVirus (SARS- CoV) and the Middle East Respiratory Syndrome-related CoronaVirus (MERS-CoV). Among human viruses, SARS-CoV-2 (C) together with SARS-CoV (S) and MERS-CoV (M) constitute the β coronavirus group, and they use the same cellular receptor, the Angiotensin-Converting Enzyme 2 (ACE2), with SARS-CoV-2 sharing approXimately 80 and 70% amino acid sequence identity with SARS-CoV and MERS-CoV, respectively (Saputri et al., 2020; Walls et al., 2020). SARS-CoV viral proteome includes 15 reviewed proteins, while MERS-CoV contains 10 reviewed proteins in the UNIPROT database. Our find- ings confirm the strong similarities among these three coronaviruses at the level of CrUP structure and architecture against human pro- teome. Interestingly, a more comprehensive analysis of CrUPs per protein has revealed significant differences between them. The den- sity of M/H-CrUPs per protein ranges between 71-76% (5% range), the density of S/H-CrUPs per protein varies between 61-76% (15% range) and the density of C/H-CrUPs per protein fluctuates between 56-78% (22% range) (Table 2), thus indicating the comparatively more heterogenous CrUPs density in the SARS-CoV-2 coronaviral proteome.

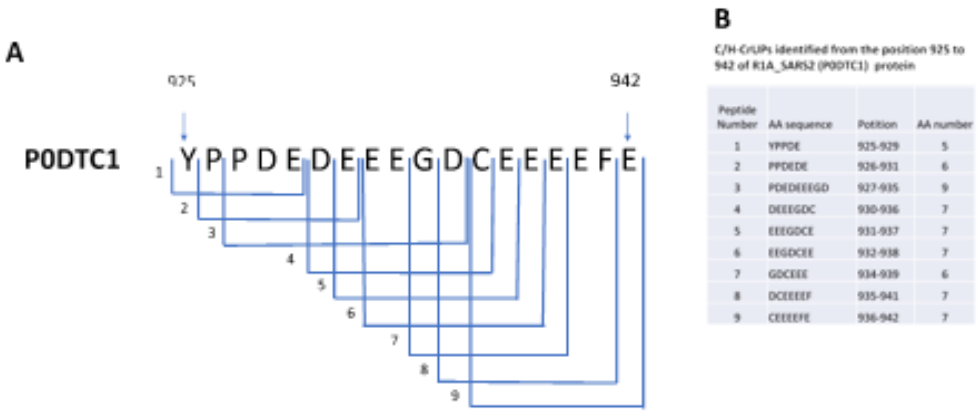


Figure 2. Identification of C/H-CrUPs around amino acid position 925-942 of the SARS-CoV-2 protein R1A_SARS2 (P0DTC1). In between these amino acid positions one of the two C/H-CrUPs with a 9 amino acid length is included (927–935). A) Schematic representation of the C/H-CrUPs included in that peptide (925-942), B) Table of C/H-CrUPs

2.3. Comparative analysis of viruses spike protein

Among all SARS-CoV-2 proteins, the SPIKE_SARS2 (P0DTC2) one (Spike) has received the greatest attention as a key element for virus attachment to the host cell, and as such it has become a principal target for therapeutic vaccine development (Papa et al., 2021; Xia 2021). To mechanistically couple protein's molecular features with virus pathology at the level of C/H-CrUPs, we comparatively analyzed the Spike proteins of the three coronaviruses, and, next, we projected the findings onto SPIKE_SARS2 mutation map. Spike glycoprotein presents a length of 1.273 amino acids in SARS-CoV-2, 1.275 amino acids in SARS-CoV and 1.373 amino acids in MERS-CoV (Agrawal et al., 2021). Their densities in CrUPs against the human proteome are measured as 78%, 76% and 76%, respectively, exhibiting the highest CrUP density values among all proteins for each virus herein studied (Table 2). Amino acid sequence alignment of SPIKE_SARS2 (P0DTC2), SPIKE_SARS (P59594) and R9UQ53_MERS (R9UQ53) proved that these three viral Spike proteins share a group of 12 regions, herein defined as Universal Peptides (UnPs) (Figure 3 and Table 3). The majority of coronaviral UnPs are clustered in the S2 domain of each Spike protein, with a critical one of them (UnPs) containing the Furin cleavage site 3 (R815↓S).

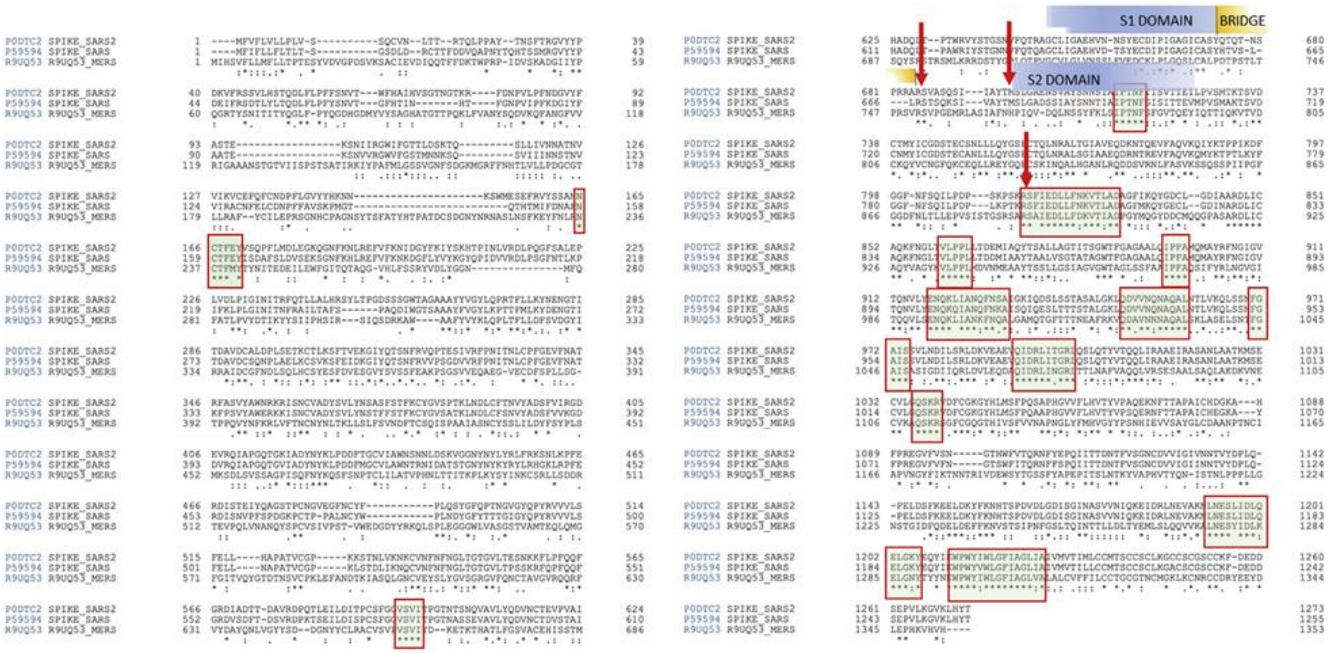


Figure 3. Alignment of the SARS-CoV-2, SARS-CoV and MERS-CoV Spike proteins. The amino acid sequence of sSpike proteins P0DTC2, P59594 and R9UQ53 derived from the SARS-CoV-2, SARS-CoV and MERS-CoV viruses, respectively, were obtained for Uniprot database and subsequently aligned, according to an online available bioinformatic tool in that database. Green blocks with red outline mark the identical peptidic sequences between the alignment sequences. The identical peptide sequences are considered as Universal Peptides (UnPs). Red arrows indicate the cleavage sites of the SARS-CoV-2 Spike protein.

Table 3. Spike-derived Universal Peptides (UnPs) and their residing CrUPs against the human proteome.

SITE				SEQUENCE	C/H-CrUP	DOMAIN	
165-168	170			NCTF*Y	CTFEY	S1 Domain	N-Terminal domain
595-598				VSVI	VSVITP	S1 Domain	
714-718				IPTNF	IPTNFT	S1 Domain	
815-816	818-823	825-827	829-830	RS*IEDLLF*KVT*AD	RSFIED	S2 Domain	S3 Cleavage site (Furin)
					SFIEDL		
					FIEDLL		
					IEDLLF		
					EDLLFN		
					DLLFNK		
					LLFNKV		
					LFNKVT		
860-864				VLPPL	VLPPLLT	S2 Domain	
896-899				IPFA	IPFAMQ	S2 Domain	Internal fusion peptide
918-921	923-925	927-928	930	ENQK*IAN*FN*A	ENQKLI	S2 Domain	
					QKLIAN		
					KLIANQ		
					LIANQF		
					IANQFN		
					ANQFNS		
					NQFNSA		
949-950	952-953	955-959		QD*VN*NAQAL	DVVNQN	S2 Domain	Heptad Repeats 1
					VVNQNA		
					NQNAQA		
					NAQALN		
970-974				FGAIS	FGAISSV	S2 Domain	Heptad Repeats 1
992-997	999-1001			QIDRLI*GRL	**QIDRLI**	S2 Domain	
					IDRLIT		
					DRLITG		
					RLITGR		
1036-1039				QSKR	QSKRVD	S2 Domain	
1193-1204	1206			LNESLIDLQELG*Y	**NESLID**	S2 Domain	Heptad Repeats 2
					SLIDLQ		
					LIDLQE		
					IDLQELG		
					DLQELGK		
					QELGKY		
1211-1216	1217-1224		1226	KWPWY*WLGFIAGL*A	**WPWY**	S2 Domain	Trans membrane domain
					PWYIWL		

Collection of the Universal peptides of SARS-CoV-2, SARS-CoV and MERS-CoV spike proteins according to Figure 4B alignement. The position in each protein sequence and the peptide sequence are shown. "" symbol indicates positions with different amino acids residues among the examined proteins. CrUPs being contained in Universal Peptides (UnPs) are recorded. Notably, they are followed by the domain of Spike protein which they belong in. Yellow blocks indicate complete sequence CrUPs that appear in the Universal Peptides (UnPs) in all Spike proteins alignment.*

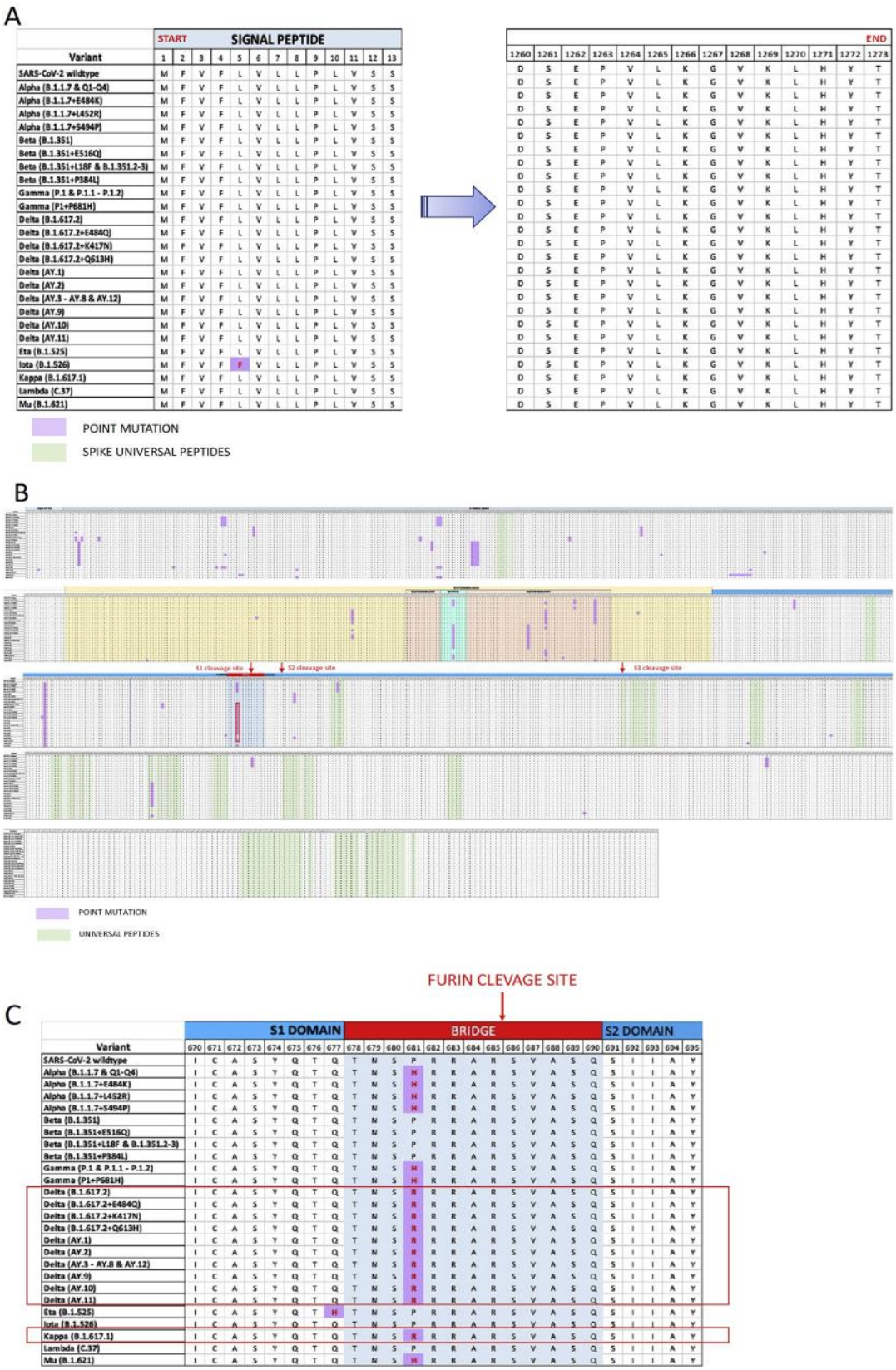


Figure 4. Alignment of the SARS-CoV-2 Spike protein (SPIKE_SARS2, P0DTC2) of the 25 sup-variants belonging to the major 9 virus variants, together with the native (wild-type) Spike Protein. A) N-terminal and C-terminal areas of the native (wild-type) Spike protein, and the 25 sup-variants are presented. B) Complete Spike protein sequence alignment. Purple blocks mark the point mutation sites in variants; green color indicates the Unique Peptides (UnPs) of the Spike proteins

from Figure 3; yellow color denotes the Receptor-Binding Domain (RBD) of Spike protein to ACE2; pink color indicates the Receptor-Binding Motif (RBM); cyan color marks the NF9 peptide; light blue color indicates the bridge between S1 and S2 domains; red arrows denote the cleavage sites. Different domains of the Spike protein are marked with different colors in the upper side of the alignment. C) The Spike protein alignment around the bridge domain (light blue color) between the S1 and S2 domains is presented. Red arrow denotes the Furin cleavage site $R^{685}\downarrow S$. Purple blocks mark the point mutations around this position, while red outline indicates the Delta and Kappa variants carrying the critical mutation P681R.

2.4. Analysis of SARS-CoV-2 variants spike protein

Most importantly, SARS-CoV-2 Spike protein has presented a significant mutational diversity (Sanches et al., 2021; Tzou et al., 2020). Hitherto, 9 main variants with adaptive mutations and high spread to human populations, named from Alpha to Lambda, respectively, have been thoroughly mapped and characterized. These 9 variants are divided in 39 sub-variants, while other 32 sporadic variants have also been described (Tzou et al., 2020). To investigate the association of mutational profiling with C/H–CrUP landscaping of SARS-CoV-2 Spike protein, the 39 sub-variants together with the wild-type Spike protein (SPIKE_SARS2, P0DTC2) were suitably aligned (Figure 4). This multiple alignment il- lustrates all the herein identified Universal Peptides (UnPs) (Table 3) and all the mutations previously announced per isolated variant (Figure 4B). Notably, it seems that almost all the hitherto characterized mutations are identified in regions being located outside the UnPs group. Their majority are clustered in the S1 domain of Spike protein, with two critical mutations being detected in the S1–S2 bridge region, at the amino acid residue 681 that resides in proXimity to the first cleavage position by Furin protease, in between the 685th and 686th amino acid residue (Figure 4C) (Davidson et al., 2020; Coutard et al., 2020).

Remarkably, all the examined mutations herein prove to create new CrUPs against the human proteome compared to the wild-type Spike protein, thus indicating that the mutant virus strains need novel clinical treatments. This is an important finding, since these new C/H-CrUPs do not exist in the human proteome, but are observed exclusively in the mutant virus proteomes, thereby justifying the great attention Alpha, Delta, Kappa, Lambda and Mu variants have recently received at the worldwide level (Tzou et al., 2020). Table 4 lists all the novel C/H-CrUPs being created by the hitherto reported mutations in coronavirus variants. These variants include 25 mutations, which produce 44 new CrUPs against the human proteome. It may be these novel C/H-CrUPs that give rise to formation of new Intrinsically Disordered Regions (IDRs) and Small Linear Motifs

(SLiMs) in the SARS-CoV-2 Spike protein mutant versions (van der Lee et al., 2014; Hraber et al., 2020).

The molecular mechanism of Spike protein's proteolytic activation has been shown to play a crucial role in the selection of host species, virus binding to the ACE2 receptor, virus-cell fusion, and viral infection of human lung cells (Peacock et al., 2021; Whittaker 2021; Shang et al., 2020a). Spike protein contains three cleavage sites: the R685↓S and the R815↓S positions that serve as direct targets of Furin protease, and the T696↓M position that can be recognized by TMPRSS2 protease (Hoffmann et al., 2020a, 2020b; Takeda, 2021). Analysis of the wild-type C/H-CrUPs and the new formed, mutation-induced, C/H-CrUPs in Spike protein unveiled that the mutation-driven, novel, peptides are created exclu- sively around the critical R685↓S cleavage site by the two pathogenic mutations P681H and P681R (Table 5).

2.5. Analysis of C/H-CrUPs around the R685↓S cleavage site

Notably, among these four new peptides (Table 5), the only one that embraces Furin's cleavage site is the “SRRRAR↓S” C/H–CrUP, which is solely generated by the P681R mutation carried by the Delta and Kappa coronavirus variants, while at the same time the “PRRARSV” peptide maintains its uniqueness even after the replacement of Proline (P) with Arginine (R) and its transformation to “RRRARSV” (Figure 5A,B).

The Furin cleavage site R685↓S has been characterized as a 20 amino acid sequence motif that corresponds to the amino acid sequence A672- S691 of the Spike protein (Figure 4A,B) (Wu and Zhao, 2020). The 8 amino acid sequence peptide “SPRRAR↓SV” (S680–V687) serves as the core region of the motif, while two flanking solvent-accessible regions of 8 amino acids (A672-N679) and 4 amino acids (A688-S691) long, respectively, are recognized (Takeda, 2021; Wu and Zhao, 2020).

Pro-protein Convertase (PC) Furin and/or Furin-like PCs act as sequence-specific proteases and can cleave the Spike protein in a position recognizing the unique, and positively charged by the Arginine, motif “R- X-X-R↓S” (Wu and Zhao, 2020). Since Furin and/or Furin like PCs are secreted from host cells and bacteria in the airway epithelium, while other PCs, such as PC5/6A and PACE4, exhibit widespread tissue distribution, it is likely that their activities may be critically implicated in the SARS-CoV-2-induced damage and pathology of multiple infected organs. It seems that Furin’s cleavage site essentially contributes to the infection process and disease progression, and offers a

powerful target for immunogenetic, antigenic and therapeutic in terventions, as strongly supported by the recently developed new anti- body against Furin's cleavage site (Braun and Sauter, 2019; Zahradník et al., 2021; Wu et al., 2020).

Most importantly, the SARS-CoV-2 Delta variant that carries the critical mutation P681R seems to be more infectious and pathogenic than the wild-type virus form, while the importance of this mutation has very recently begun to be recognized (Wu et al., 2020). Replacement of Proline (P) with Arginine (R) at position 681 causes the loss of amino acid sequence uniqueness that characterizes the wild-type “PRRARSV” C/H–CrUP and likely increases the possibility of Furin's cleavage site (core region) to be significantly stabilizing its conformation, thus facilitating a more efficient Spike protein cleavage process by the Furin pro- tease (Whittaker, 2021; Callaway, 2021).

To the same direction, novel SLiMs, such as “SRRR”, “RRR”, “RRRAR” and “RRRARS”, can be produced by the mutant C/H-CrUPs, which may act as specific targets of other than Furin PCs, thereby enabling the stronger (and quicker) binding of the mutant virus to its host ACE2 re- ceptor, which likely leads to a comparatively more generalized infection and massive mutant virus production (Table 6) (Shorthouse and Hall, 2021; Davey et al., 2015). This finding seems to be evidenced by the remarkable increase of the total number of motifs created by the P681R mutation identified within the human proteome (Table 6). Of note, the mutant C/H–CrUP-derived new SLiMs, in the SARS-CoV-2 Delta variant, could render Spike protein antigenically weak or defective, fostering it to lose its capacity to serve as antibody target and thus promoting the virus immune escape (Davey et al., 2015; Almehdi et al., 2021).

2.6. Analysis of C/H-CrUPs around the ACE2 receptor site

An important issue for viral infectivity and pathogenesis is the receptor recognition and binding of the virus to the host cell surface. SARS- CoV-2 belongs to the β coronavirus group and, like SARS-CoV, uses the same cellular receptor, the Angiotensin-Converting Enzyme 2 (ACE2) (Walls et al., 2020; Wang et al., 2020). The SARS-CoV-2 Spike protein attaches to ACE2 receptor by a Receptor-Binding Domain (RBD) defined in the Spike protein from positions F318 up to F541 (Shang et al., 2020b). Nowadays, this region has received great attention, as it seems to be the target of antibodies against the virus and other therapeutic interventions (Chen et al., 2021; Zahradník et al., 2021; Hastie et al., 2021). Additional studies have shown that from the amino acid residue W436 up to the Q506 one the RBD contains the Receptor-Binding Motif (RBM), which carries 12 contact positions with ACE2 (Hatmal et al., 2020). Mutation analysis revealed that in 10 positions of the RBD region 13 mutations were described (Figure 4 and Table

7). In RBM, 10 mutations in 6 sequence positions were reported for different virus variants (Table 7), while from the 10 contact positions only the P501Y in Alpha, Beta, Gamma, and Mu variants was found to be mutated.

Table 4. New C/H-CrUPs of SARS-CoV-2 Spike protein in Alpha, Delta, Kappa and Lambda variants.

Mutations position	Mutation	Variant	New C/H-CrUPs first AA position	New C/H-CrUPs
19	T19R	Delta_P0DTC2	-	-
70	V70F	Delta_P0DTC2	69	HFSGTN
			70	FSGTNG
75 - 76	G75V&T76I	Lambda_P0DTC2	71	SGTNVI
			75	VIKRFD
222	A222V	Delta_P0DTC2	218	QGFSVL
258	W258L	Delta_P0DTC2	-	-
417	K417N	Delta_P0DTC2	413	GQTGNI
			414	QTGNIA
452	L452R	Delta_P0DTC2	449	YNYRY
		Kappa_P0DTC2		
		Alpha_P0DTC2		
	L452Q	Lambda_P0DTC2	448	NYNYQ
			449	YNYQY
478	T478K	Delta_P0DTC2	474	QAGSKP
			478	KPCNG
484	E484Q	Kappa_P0DTC2	481	NGVQG
			483	VQGFN
			484	QGFNC
	E484K	Alpha_P0DTC2	484	KGFNC
490	F490S	Lambda_P0DTC2	487	NCYSP
494	S494P	Alpha_P0DTC2	-	-
501	N501Y	Alpha_P0DTC2	498	QPTY
			499	PTYG
			500	TYGV
			501	YGVG
570	A570D	Alpha_P0DTC2	568	DIDDTT
614	D614G	Delta_P0DTC2	609	AVLYQG
		Kappa_P0DTC2		
		Alpha_P0DTC2		
		Lambda_P0DTC2		
		Delta_P0DTC2	610	VLYQGV
		Kappa_P0DTC2		
		Alpha_P0DTC2		
		Lambda_P0DTC2		
681	P681R	Delta_P0DTC2	680	SRRRARS
		Kappa_P0DTC2		

	P681H	Alpha_P0DTC2	677	QTNSH
			678	TNSHR
			680	SHRRAR
716	T716I	Alpha_P0DTC2	714	IP|NF
859	T859N	Lambda_P0DTC2	855	FNGLNV
			857	GLNVLP
950	D950N	Delta_P0DTC2	946	GKLQN
			947	KLQNVV
			948	LQNVVN
			949	QNVVNQ
982	S982A	Alpha_P0DTC2	978	NDILAR
1071	Q1071H	Kappa_P0DTC2	1067	YVPAH
			1069	PAHEKN
			1071	HEKNF
1118	D1118H	Alpha_P0DTC2	1113	QIITTH
			1115	ITTHN
			1116	TTHNT
			1117	THNTF
			1118	HNTFV

The new C/H-CrUPs of SARS-CoV-2 spike protein (SPIKE_SARS2, P0DTC2) across the variants Alpha, Delta, Kappa and Lambda are presented. In the first column, the position of each mutation in the Spike protein sequence is shown. In the second column the mutation is recorded. In the third column, the SARS-CoV-2 main variant which each mutation is appeared in, is recorded. In the fourth column, the position of the first amino acid residues of the new C/H–CrUP created by each mutation is shown. In the last column, the new created C/H-CrUPs by each mutation is recorded. Each mutant amino acid residue in the new C/H-CrUPs is denoted by red color. Mutations that not create new C/H-CrUPs are indicated by the symbol '-'. Some mutations produce multiple new C/H-CrUPs, while 4 new C/H-CrUPs are created in more than one variant.

2.7. C/H-CrUPs around the NF9 peptide

The most important region in RBM is the peptide "NYNYLYRLF" (from 448 to 456 position). This Tyrosine (Y)-enriched peptide contains two contact site (Y449 and Y453) and it is known as the NF9 peptide (Motozono et al., 2021). It seems to affect antigen recognition, by being an immunodominant HLA*24:02-restricted epitope identified by the CD8þ T-cells. Furthermore, NF9 stimulation also increases cytokine production by the CD8þ T-cells, such as IFN-γ, TNF-α and IL-2 (Kared et al., 2021). Analysis of C/H-CrUPs that are being associated with the NF9 peptide showed that it contains 3 UPs (Figure 5D,E, and Table 7). Mutation analysis indicated that in the NF9 peptide the mutation L452R is carried by the variants Alpha, Delta, Lamda and Kappa, while the mutation L452Q appears in the variant Lambda. Further analysis un- veiled that these mutations

are observed in the amino acid that resides at position 5, exactly in the middle of the peptide, creating 3 and 4 new C/H CrUPs, respectively (Table 8). These mutations have a dramatic effect in the uniqueness of the NF9 peptide(s). Namely, the 6 amino acid length C/H-CrUPs “NYNYLY” lose their uniqueness against the human proteome, while only by the mutation L452Q a new CrUP with 5 amino acid length is surprisingly created (Figure 5D,E). The loss of uniqueness of this peptide, which notably is located at the beginning of NF9 peptide, seems to be crucial, as it leads to the loss of antigenic capacity of the NF9 peptide, thus evading the HLA-A24-restricted immunity and inducing the immune escape of the virus. Interestingly, related studies have shown that the L452R mutation (and subsequently the new created C/H-CrUPs herein characterized) increases the infectiveness of SARS-CoV-2, by strengthening the electrostatic interactions of this region on Spike pro- tein with the ACE2 virus receptor (Motozono et al., 2021). Hitherto, epidemiological data indicated that the dominant variant of SARS-CoV-2 is the Delta variant (Micochova et al., 2021). Under the light of the findings, variant's enhanced pathogenicity seems to be the outcome of the simultaneous presence (accumulation) of two critical mutations, the L452R and P681R ones, in Delta variant. The mutation L452R, through the loss of NF9 peptide uniqueness, causes virus immune escape and strong(er) binding of the virus to its cognate receptor, while at the same time the mutation P681R facilitates the Spike protein cleavage process by different proteases, inducing a generalized infection and a massive virus release. Therefore, the Delta variant gains a significant advantage of escape from the immune system per se, as well as from the vaccination-induced immunity, together with an increased infectiveness as a result of virus entrance into the host cell, and an in- crease of virus formation and its massive release.

3. Conclusion

Since mutations outside the Spike protein locus in SARS-CoV-2 coronavirus genome have not been yet completely mapped, in a systematic manner, our study importantly reveals novel and useful information of all the remaining, Spike protein-independent, C/H-CrUPs that seem to hold strong promise and open new therapeutic windows for the Covid-19 pandemic. Finally, the approach of virus-host UP-specific signature identification could prove a useful tool for the elucidation of virus infectiveness, prevention of virus immune escape, domination of pathogenic variants, and identification of new antigenic and pharmaco- logical targets.

Table 5. New C/H-CrUPs around the SARS-CoV-2 Spike protein cleavage sites.

Cleavage site	Mutation	Variant	New C/H-CrUPs first AA position	New C/H-CrUP
$R^{685}\downarrow S$	P681R	Delta & Kappa	680	SRRRAR↓S
	P681H	Alpha & Gamma	677	QTNSH
			678	TNSHR
			680	SHRRAR
$T^{696}\downarrow M$	A701V	Beta	None	
$R^{815}\downarrow S$	None		None	

The new C/H-CrUPs created by the mutations around the SARS-CoV-2 spike protein (SPIKE_SARS2, P0DTC2) are identified. Fist column: The cleavage site of SARS-CoV-2 Spike protein. Second column: The mutation identified around the cleavage site. Third column: The virus variants in which the mutation appears in. Fourth column: The position in the SARS-CoV-2 Spike protein sequence which the first amino acid of the C/H–CrUP appears in. Fifth column: The sequence of the new C/H–CrUP. "↓" symbol indicates the cleavage site within this peptide.

4. Materials and methods

4.1. Methods

A new bioinformatics tool that has been recently built on an advanced big-data algorithm was herein developed to extract CrUP collections from proteomes of interest and, thereby, create organism-specific Uniquomes. The user can specify the min and max peptide lengths that the tool will analyze. The tool will split each protein to all possible peptides of length min to length max, thus generating a very large set of peptides (for a

protein of length L with a window of size W, a set of “C ¼ L - W þ 1” will be generated). In the next step, all these peptides, starting from smallest

Since mutations outside the Spike protein locus in SARS-CoV-2 coronavirus genome have not been yet completely mapped, in a sys- tematic manner, our study importantly reveals novel and useful infor- mation of all the remaining, Spike protein-independent, C/H-CrUPs that seem to hold strong promise and open new therapeutic windows for the Covid-19 pandemic. Finally, the approach of virus-host UP-specific signature identification could prove a useful tool for the elucidation of virus infectiveness, prevention of virus immune escape, domination of pathogenic variants, and identification of new antigenic and pharmaco- logical targets.

4. Materials and methods

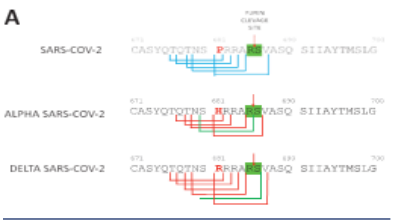
4.1. Methods

A new bioinformatics tool that has been recently built on an advanced big-data algorithm was herein developed to extract CrUP collections from proteomes of interest and, thereby, create organism-specific Uniquomes. The user can specify the min and max peptide lengths that the tool will analyze. The tool will split each protein to all possible peptides of length min to length max, thus generating a very large set of peptides (for a protein of length L with a window of size W, a set of “C ¼ L - W þ 1” will be generated). In the next step, all these peptides, starting from smallest and ending to largest, will be searched against the rest of the proteome to decide whether the peptide exists on another protein or not. Since the search is dedicated for the smallest possible peptide

(Core Unique Pep- tide: CrUP), the tool will first make sure that the peptide under examination does not already contain a smaller CrUP. This is ensured by examining if any of the already identified CrUPs of the protein is contained within the peptide under examination. All peptides that conform to these two rules are considered as CrUPs. Figure 6 describes the algorithm we have herein developed and used to recognize these novel CrUPs.

In Figure 7, a sliding window of 9 amino acids is applied on O00400 ACATN_HUMAN protein, generating the candidate peptides “VYVKNFGRR” and “YVKNFGRRK”. These peptides will be searched against the rest of the proteome, to determine their uniqueness once we have ensured that they do not already contain a smaller CrUP. The latter is determined by examining whether an already defined CrUP is included within the peptide.

To address the question of the present study, the aforementioned tool was expanded by developing a new feature, where the user can give a reference and a target proteome. This new feature allows the tool to search all the peptides of the target proteome against the reference proteome, thus creating a set of CrUPs of target versus reference proteome. To this direction, the tool (similar to the initial implementation) will split all proteins in the target proteome to all possible peptides of length min to length max. Now, instead of searching for the uniqueness of each peptide within the same proteome, it performs that search against the reference proteome. Like before, the peptide under examination must not contain any smaller peptides already identified as CrUPs. The algorithm we have employed to identify these CrUPs is described diagram- matically in Figure 6.



B C/H-CrUPs of wild-type and mutant SARS-CoV-2 around the $R^{685}\downarrow S$ cleavage site.

	PEPTIDES		
		MUTATION	
POSITION	SARS-CoV-2	P681R	P681H
676	TQTNSP	TQTNSR	
677	QTNSPR	QTNSRR	QTNSH
678	TNSPRR	TNSRRR	TNSHR
679	NSPRRA	NSRRRA	NSHRRA
680		SRRRAR↓S	SHRRAR
681	PRRAR↓SV	RRRAR↓SV	HRRAR↓SV

D NF9 wild-type and mutant C/H-CrUPS.

	PEPTIDES		
		MUTATION	
POSITION	SARS-CoV-2	L452R	L452Q
448	NYNYLY		NYNYQ
449		YNYRY	YNYQY
450	NYLYRL	NYRYRL	NYQYRL
451	YLYRLF	YRYRLF	YQYRLF

Figure 5. C/H-CrUPs residing around the $R^{685}\downarrow S$ cleavage site and belonging to the NF9 peptide of Spike protein (SPIKE_SARS2, P0DTC2). A) Amino acid sequences of Spike protein between position 671 and 700 in wild-type, Alpha and Delta variants of SARS-CoV-2 virus are shown. In each variant, the identified C/H-CrUPs are marked. Blue lines indicate C/H-CrUPs derived from wild-type protein around the $R^{685}\downarrow S$ cleavage site. Red lines denote C/H-CrUPs produced by the P681H and P681R mutations. Green lines indicate the new created mutant C/H-CrUPs that derive from the P681H and P681R mutations in Alpha and Delta variants, respectively. B) Set of C/H-CrUPs generated around the $R^{685}\downarrow S$ cleavage site of wild-type and mutant Spike protein forms. C) Amino acid sequences of the NF9 peptide between positions 448 and 456 in wild-type Spike protein, before and after creation of the L452R and L452Q mutations. Blue lines indicate C/H-CrUPs that belong to the NF9 peptide. Red lines denote C/H-CrUPs that are produced by the L452R and L452Q mutations. Green lines indicate the new generated mutant collection of C/H-CrUPs derived from the L452R and L452Q mutations. D) Set of C/H-CrUPs residing in the NF9 peptide in wild-type, and L452R and L452Q mutated protein forms.

Table 6. Small Linear Motifs (SLiMs) of wild-type C/H-CrUPs and C/H-CrUPs created by the critical mutation P681R being detected in human proteome.

Motif	Number of proteins in UNIPROT contain the motif	Motif found	Protein Entry ID	Protein Entry Name	Protein full Name
PRRARSV	0	-	-		
XRRARSV	1	ARRARSV	P37088	SCNNA_HUMAN	Amiloride-sensitive sodium channel subunit alpha
PXRARSV	0	-	-		
PRXARSV	1	PRPARSV	Q96PD5	PGRP2_HUMA	N-acetylmuramoyl-L-alanine amidase
PRRXRSV	1	PRRSRSV	Q9UQ35	SRRM2_HUMAN	Serine/arginine repetitive matrix protein 2
PRRAXSV	2	PRRASSV	Q04844	ACHE_HUMAN	Acetylcholine receptor subunit epsilon
		PRRALSV	Q5VZ46	K1614_HUMAN	Uncharacterized protein KIAA1614
PRRARXV	0	-	-		
PRRARSX	1	PRRARSS	Q92902	HPS1_HUMAN	Hermansky-Pudlak syndrome 1 protein
TOTAL TIMES	6				

Motif	Number of proteins in UNIPROT contain the motif	Motif found	Protein Entry ID	Protein Entry Name	Protein full Name
SRRRARS	0	-	-		
XRRRARS	6	RRRRARS	Q8WUQ7	CATIN_HUMAN	Cactin
		RRRRARS	P18825	ADA2C_HUMAN	Alpha-2C adrenergic receptor
		DRRRARS	Q96QZ7	MAGI1_HUMAN	Membrane-associated guanylate kinase, WW and PDZ domain-containing protein 1
		PRRRARS	C9J069	ALM1_HUMAN	Apical junction component 1 homolog
		WRRRARS	O00198	HRK_HUMAN	Activator of apoptosis harakiri
		PRRRARS	Q9NZV5	SELN_HUMAN	Selenoprotein N
SXRRARS	3	SDRRARS	Q8N2C7	UNC80_HUMAN	Protein unc-80 homolog
		SPRRARS	Q92902	HPS1_HUMAN	Hermansky-Pudlak syndrome 1 protein

		SCRRARS	Q5T4W7	ARTN_HUMAN	Artemin
SRXRARS	1	SRDRARS	Q92917	GPKOW_HUMAN	G-patch domain and KOW motifs-containing protein
SRRXARS	1	SRRQARS	Q9NSI2	F2007A_HUMAN	Protein FAM207A
SRRRXRS	10	SRRRPRS	Q70EL4	UBP43_HUMAN	Ubiquitin carboxyl-terminal hydrolase 43
		SRRRIRS	P05198	IF2A_HUMAN	Eukaryotic translation initiation factor 2 subunit 1
		SRRRRRS	P18583	SON_HUMAN	Protein SON
		SRRRSRS	Q8N2M8	CLASR_HUMAN	CLK4-associating serine/arginine rich protein
		SRRRSRS	Q15058	KIF_HUMAN	Kinesin-like protein KIF14
		SRRRRRS	Q5M9Q1	NKAPL_HUMAN	NKAP-like protein
		SRRRSRS	Q14498	RBM39_HUMAN	NA-binding protein 39
		SRRRSRS	Q96T37	RBM15_HUMAN	RNA-binding protein 15
		SRRRSRS	Q13247	SRSF6_HUMAN	Serine/arginine-rich splicing factor 6
		SRRRQRS	Q9UQ35	SRRM2_HUMAN	Serine/arginine repetitive matrix protein 2
SRRRAXS	6	SRRRAIS	Q00987	MDM2_HUMAN	3 ubiquitin-protein ligase Mdm2
		SRRRAQS	Q9NQU5	PAK_HUMAN	Serine/threonine-protein kinase PAK 6
		SRRRADS	Q53GL0	PKHO1_HUMAN	Pleckstrin homology domain-containing family O member 1
		SRRRAWS	Q9UKN7	MYO15_HUMAN	Unconventional myosin-XV
		SRRRAFS	Q9GZK7	011A1_HUMAN	Olfactory receptor 11A1
		SRRRAVS	Q9BYX2	TBD2A_HUMAN	TBC1 domain family member 2A
SRRRARX	3	SRRRARR	O95450	ATS2_HUMAN	A disintegrin and metalloproteinase with thrombospondin motifs 2
		SRRRARD	Q8N5L8	RP25L_HUMAN	Ribonuclease P protein subunit p25-like protein
		SRRRARV	Q9GZQ6	NPFF1_HUMAN	Neuropeptide FF receptor 1
TOTAL TIMES	**30**				

Motif	Number of proteins in UNIPROT contain the motif	Motif found	Protein Entry ID	Protein Entry Name	Protein full Name
RRRARSV	0	-	-		
XRRARSV	1	ARRARSV	P37088	SCNNA_HUMAN	Amiloride-sensitive sodium channel subunit alpha
RXRARSV	1	RPRARSV	Q86X29	LSR_HUMAN	Lipolysis-stimulated lipoprotein receptor
RRXARSV	2	RRDARSV	Q8WWN8	ARAP3_HUMAN	Arf-GAP with Rho-GAP domain, ANK repeat and PH domain-containing protein 3
		RRPARSV	Q9H427	KCNKF_HUMAN	Potassium channel subfamily K member 15
RRRXRSV	3	RRRSRSV	P18583	SON_HUMAN	Protein SON
		RRRKRSV	P49685	GPR15_HUMAN	G-protein coupled receptor 15
		RRRASSV	O14681	EI24_HUMAN	Etoposide-induced protein 2.4 homolog
RRRAXSV	3	RRRAQSV	Q7LDG7	GRP2_HUMAN	RAS guanyl-releasing protein 2
		RRRAPSV	P21333	FLNA_HUMAN	Filamin-A
		RRRARPV	Q7RTU4	BHA09_HUMAN	Class A basic helix-loop-helix protein 9
RRRARXV	4	RRRARQV	Q8N9Z2	CC71L_HUMAN	Coiled-coil domain-containing protein 71L
		RRRARAV	Q6NUJ1	SAPL1_HUMAN	Proactivator polypeptide-like 1
		RRRARVV	Q9GZQ6	NPFF1_HUMAN	Neuropeptide FF receptor 1
		RRRARSW	Q8WUQ7	CATIN_HUMAN	Cactin
RRRARSX	5	RRRARSS	P18825	ADA2C_HUMAN	Alpha-2C adrenergic receptor
		RRRARSP	Q96QZ7	MAGI1_HUMAN	Membrane-associated guanylate kinase, WW and PDZ domain-containing protein 1
		RRRARSK	C9J069	AJM1_HUMAN	Apical junction component 1 homolog
		RRRARSR	O00198	HRK_HUMAN	Activator of apoptosis harakiri
		RRRARSL	Q9NZV5	SELN_HUMAN	Selenoprotein N
TOTAL TIMES	**19**				

Motif	Number of proteins in UNIPROT contain the motif	Motif	Number of proteins in UNIPROT contain the motif	Motif	Number of proteins in UNIPROT contain the motif
XRRRARX	47	SRRXXRS	46	RXXRS	3774
XRRRAXS	44	SRRRXRX	53	**TOTAL TIMES**	**3774**
XRRRXRS	139	SRRRXXS	50		
XRRXARS	30	SRRRAXX	25		
XRXRARS	22	**TOTAL TIMES**	**174**		
XXRRARS	27				
SXRRARX	29				
SXRRAXS	29				
SXRRXRS	72				
SXRXARS	17				
SXXRARS	20				
SRXRARX	19				
SRXRAXS	35				
SRXRXRS	175				
SRXXARS	16				
SRRXARX	19				
SRRXAXS	24				
TOTAL TIMES	**735**				

The list of SLiMs of wild-type and mutant C/H-CrUPs produced by the critical mutation P681R in SPIKE_SARS2, and being detected in the human proteome, are presented. Green block indicates the C/H–CrUP in wild-type protein; blue block denotes the mutant C/H–CrUP peptide derived from the P681R mutation; yellow block descibed the newly created C/H–CrUP by the same mutation. X (in red color) is used for the position within the peptide to create the motif. In the third column, the detected motif is recorder, and is followed by the Protein Entry-ID and the protein name it is detected in. "Total" summarizes the time for which the motifs related to C/H–CrUP are recorded in the human proteome.

4.2. Motifs and SLiMs search

For Motif and SLiM identification, and search, the tool offers the user the ability to perform a motif search to identify putative SLiMs. User gives an N-length peptide, as well as the number of amino acids that can vary in the given peptide. Then, the tool creates all possible combinations of peptides that can be produced by considering in each combination exactly N-amino acid(s) as unknown. Once these combinations are produced, an exhaustive search using regular expressions is performed against the reference proteome, to locate all possible proteins containing such peptides. To better highlight the process, if the user provides the peptide “TQYILG” and N ¼ 2, the following combinations will be generated:

- ??YILG
- ?Q?ILG
- ?QY?LG
- ?QYI?G
- ?QYIL?
- T??ILG
- T?Y?LG
- T?YI?G
- T?YIL?
- TQ??LG
- TQ?I?G
- TQ?IL?
- TQY??G
- TQY?L?
- TQYI??

User will receive a list of all proteins containing peptides that match the criteria, including the motif against which the peptide was matched, and all the positions within the protein sequence where that peptide can be found. All proteomes were taken from the UNIPROT database.

Table 7. C/H-CrUPs of wild-type and mutant Receptor-Binding Domain (RBD) of SARS-Cov-2 Spike protein.

SARS-CoV-2 SPIKE PROTEIN RECEPTOR BINDING DOMAIN														
WILDTYPE								MUTANT						
Position	C/H-CrUP						Peptide number / peptide length	Mutation	VARIANT	NEW C/H-CrUP					Peptide number/ peptide length
346	VFNATR	FNATRF	NATRFA	ATRFAS	TRFASV	RFASVY	6/6AA	R346K	Mu	VFNATK	FNATKF	ATKFAS	TKFASV	KFASVY	5/6AA
384	YGVSPT	GVSPTK	VSPTKL	SPTKLN	PTKLND		5/6AA	P384L	Beta	YGVSLT	GVSLTK	SLTKLN	LTKLND		4/6AA
417	PGQTGKI	GQTGKIA	TGKIAD	GKIADY			2/7AA, 2/6AA	K417N	Beta, Delta	GQTGNI	QTGNIA	TGNIAD	GNIADY		4/6AA
								K417T	Gamma	PGQTGT	GQTGTI	QTGTIA	TGTIAD		4/6AA
452	GNYNYL	NYNYLY	NYLYRL	YLYRLF	LYRLFR		5/6AA	L452R	Alpha, Delta, Iota, Kappa	GNYNYR	YNYRY	NYRYRL	YRYRLF		1/5AA, 3/6AA
								L452Q	Lambda	NYNYQ	YNYQY	NYQYRL	YQYRLF	QYRLFR	2/5AA, 3/6AA
478	YQAGST	AGSTPC	STPCN				1/5AA, 2/6AA	T478K	Delta	YQAGSK	QAGSKP	AGSKPC	GSKPCN	KPCNG	1/5AA, 4/6AA
484	CNGVEG	NGVEGF	GVEGFN				3/6AA	E484K	Alpha, Beta, Gamma, Eta, Mu	CNGVKG	NGVKGF	GVKGFN	KGFNC		1/5AA, 3/6AA
								E484Q	Kappa	NGVQG	VQGFN	QGFNC			3/5AA
490	FNCYF	CYFPLQ	YFPLQS	FPLQSY			1/5AA, 3/6AA	F490S	Lambda	FNCYS	NCYSP	CYSPLQ	SPLQSY		2/5AA, 2/6AA
494	YFPLQS	FPLQSY	PLQSYG	QSYGF	SYGFQP		1/5AA, 4/6AA	S494P	Alpha	PLQP	LQPY	QPYG	PYGF		4/4AA
501	GFQPTN	FQPTNG	QPTNGV	PTNGVG	TNGVGY	NGVGYQ	6/6AA	N501Y	Alpha, Beta, Gamma, Mu	GFQPTY	QPTYG	YGVGY			2/5AA, 1/6AA
516	VVLSFE	VLSFELL	LSFELLH	FELLHA	ELLHAP		2/7AA, 3/6AA	E516Q	Beta	VVVLSFQ	VLSFQL	SFQLLH	FQLLHA	QLLHAP	4/6AA, 1/7AA

Novel C/H-CrUPs created by critical mutations in the Receptor-Binding (RBD) domain of SARS-CoV-2 wild-type and mutant Spike protein (SPIKE_SARS2, P0DTC2) amino acid sequence are identified. Peptide number/peptide length is the number of a given length C/H–CrUP around the position. By red color the amino acids in wild-type C/H-CrUPs, which will be modified, and the mutated amino acids in the new C/H-CrUPs are marked. Light blue color indicates the peptides which disappear from the wild- type viral proteome by the mutation, yellow color shows the completely new created C/H-CrUPs peptides by the mutation.

Table 8. NF9-specific C/H–CrUPS.

	PEPTIDES		
		MUTATION	
POSITION	SARS-CoV-2	L452R	L452Q
448	NYNYLY		NYNYQ
449		YNYRY	YNYQY
450	NYLYRL	NYRYRL	NYQYRL
451	YLYRLF	YRYRLF	YQYRLF

The C/H-CrUPs in wild-type and mutant NF9 peptide are listed. By red color the mutant amino acids are marked

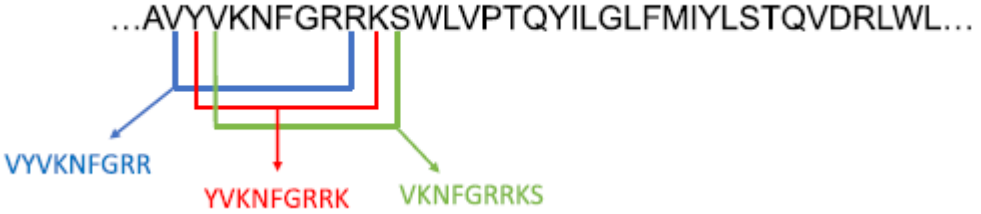


Figure 7. Presentation of the bioinformatic process developed for the identification of the CrUPs peptides, performed amino acid by amino acid residue.

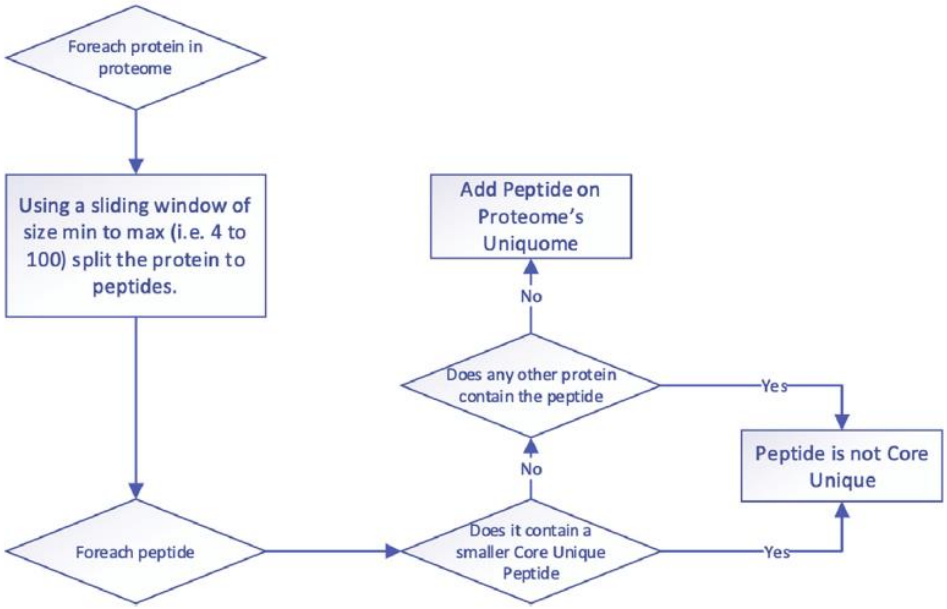


Figure 6. Schematic presentation of the algorithm herein developed for the identification of Core Unique Peptides (CrUPs).

4.3. Algorithm's application to the identification of virus CrUPs against human proteome

To recognize all the CrUPs being embraced in a virus proteome against the human proteome, we in silico constructed a new, artificial, "hybrid-proteome" that contained all the reviewed human proteins (20.430 proteins), plus the one protein derived from the viral proteome (20.431 proteins). Thereby, n "hybrid proteomes", including the n viral proteins, were constructed, with n representing the number of viral proteins. Hence, these "hybrid proteomes" were bioinformatically searched one by one for the identification of virus-specific CrUPs in human protein sequence environments.

4.4. Databases

All proteomes and proteins were obtained from UNIPROT [http s://www.uniprot.org]. SARS-CoV-2 wild-type and variant/mutated se- quences derived from Stanford COVID database [https://covdb.stanfo rd.edu/page/mutation-viewer/]. Motifs were taken from the Eukary- otic Linear Motif resource for Functional Sites in Proteins [http://el m.eu.org/index.html] and KEGG/GenomeNet/MOTIF2 [https://www.genome.jp/tools/motif/MOTIF2.html]. SLiM-containing pro- teins were taken from Davey lab SLiM servers (The Institute of Cancer Research {ICR}, UK) [http://slim.icr.ac.uk/slimsearch/] and [http://sli m.icr.ac.uk/index.php?page¼tools].

Declarations

Autor contribution statement

Vasileios Pierros: Conceived and designed the experiments; Performed the experiments; Contributed reagents, materials, analysis tools or data.

Evangelos Kontopodis: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data.

Dimitrios J. Stravopodis, George Th. Tsangaris: Conceived and designed the experiments; Analyzed and interpreted the data; Contrib- uted reagents, materials, analysis tools or data; Wrote the paper.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability statement

Data included in article/supp. material/referenced in article.

Declaration of interest’s statement

The authors declare no conflict of interest.

Additional information

No additional information is available for this paper

References

Agrawal, A., Varshney, R., Pathak, M., Patel, S.K., Rai, V., Sulabh, S., Gupta, R., Solanki, K.S., Varshney, R., Nimmanapalli, R., 2021. EXploration of antigenic determinants in spike glycoprotein of SARS-CoV2 and identification of five salient potential epitopes. Virusdisease 1-10.

Alexandridou, A., Tsangaris, G. Th., Vougas, K., Nikita, K., Spyrou, G., 2009. UniMaP: finding unique mass and peptide signatures in the human proteome. Bioinformatics 25, 3035–3037.

Almehdi, A.M., Khoder, G., Alchakee, A.S., Alsayyid, A.T., Sarg, N.H., Soliman, S.S.M., 2021. SARS-CoV-2 spike protein: pathogenesis, vaccines, and potential therapies. Infection 49, 855–876.

Braun, E., Sauter, D., 2019. Furin-mediated protein processing in infectious diseases and cancer. Clin. Transl. Immunol. 8, e1073.

Callaway, E., 2021. The mutation that helps Delta spread like wildfire. Nature 596, 472–473.

Chen, Y., Zhang, Y.N., Yan, R., Wang, G., Zhang, Y., Zhang, Z.R., Li, Y., Ou, J., Chu, W., Liang, Z., Wang, Y., Chen, Y.L., Chen, G., Wang, Q., Zhou, Q., Zhang, B., Wang, C., 2021. ACE2-targeting monoclonal antibody as potent and broad-spectrum coronavirus blocker. Signal Transduct. Targeted Ther. 6 (1), 315.

Coutard, B., Valle, C., de Lamballerie, X., Canard, B., Seidah, N.G., Decroly, E., 2020. The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade. Antivir. Res. 176, 104742.

Davey, N.E., Cyert, M.S., Moses, A.M., 2015. Short linear motifs - ex nihilo evolution of protein regulation. Cell Commun. Signal. 13, 43.

Davidson, A.D., Williamson, M.K., Lewis, S., Shoemark, D., Carroll, M.W., Heesom, K.J., Zambon, M., Ellis, J., Lewis, P.A., HiscoX, J.A., Matthews, D.A., 2020.

Characterization of the transcriptome and proteome of SARS-CoV-2 reveals a cell passage induced in-frame deletion of the furin-like cleavage site from the spike glycoprotein. Genome Med. 12, 68.

Hastie, K.M., Li, H., Bedinger, D., Schendel, S.L., Dennison, S.M., Li, K., Rayaprolu, V., Yu, X., Mann, C., Zandonatti, M., et al., 2021. Defining variant-resistant epitopes targeted by SARS-CoV-2 antibodies: a global consortium study. Science eabh2315.

Hatmal, M.M., Alshaer, W., Al-Hatamleh, M.A.I., Hatmal, M., Smadi, O., Taha, M.O., Oweida, A.J., Boer, J.C., Mohamud, R., Plebanski, M., 2020. Comprehensive structural and molecular comparison of spike proteins of SARS-CoV-2, SARS-CoV and MERS-CoV, and their interactions with ACE2. Cells 9, 2638.

Hoffmann, M., Kleine-Weber, H., Schroeder, S., Krüger, N., Herrler, T., Erichsen, S., Schiergens, T.S., Herrler, G., Wu, N.H., Nitsche, A., et al., 2020a. SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. Cell 181, 271–280.

Hoffmann, M., Kleine-Weber, H., Po€hlmann, S., 2020b. A multibasic cleavage site in the spike protein of SARS-CoV-2 is essential for infection of human lung cells. Mol. Cell 78, 779–784.

Hraber, P., O'Maille, P.E., Silberfarb, A., Davis-Anderson, K., Generous, N., McMahon, B.H., Fair, J.M., 2020. Resources to discover and use short linear motifs in viral proteins. Trends Biotechnol. 38, 113–127.

Jungreis, I., Sealfon, R., Kellis, M., 2021. SARS-CoV-2 gene content and COVID-19 mutation impact by comparing 44 Sarbecovirus genomes. Nat. Commun. 12, 2642.

Kared, H., Redd, A.D., Bloch, E.M., Bonny, T.S., Sumatoh, H., Kairi, F., Carbajo, D., Abel, B., Newell, E.W., Bettinotti, M.P., et al., 2021. SARS-CoV-2-specific CD8 T cell responses in convalescent COVID-19 individuals. J. Clin. Invest. 131, e145476.

Kontopodis, E., Pierros, V., Anagnostopoulos, A., Stravopodis, D., Papassideri, I., Vorgias, C., Tsangaris, G.T., 2019. Data processing approach for the construction and evaluation of an organism's UNIQUOME with comparative analysis for the Human, Rat and Mouse Uniquomes. Paper presented at XIII. In: Annual Congress of the European Proteomics Association: from Genes via Proteins and Their Interactions to Functions, Potsdam, Germany. March 24-28, P194.

Kontopodis, E., Pierros, V., Stravopodis, D., Tsangaris, G.T., 2022. Prediction of SARS-CoV-2 Omicron variant immunogenicity, immune escape and pathogenity, through the analysis of spike protein-specific core unique peptides. Vaccines 10, 357.

Mlcochova, P., Kemp, S., Dhar, M.S., Papa, G., Meng, B., Ferreira, I., Datir, R., Collier, D.A., Albecka, A., Singh, S., et al., 2021. SARS-CoV-2 B.1.617.2 Delta variant replication and immune evasion. Nature 599, 114–119.

Motozono, C., Toyoda, M., Zahradnik, J., Saito, A., Nasser, H., Tan, T.S., Ngare, I., Kimura, I., Uriu, K., Kosugi, Y., Yue, Y., Shimizu, R., Ito, J., Torii, S., Yonekawa, A., Shimono, N., Nagasaki,

Y., Minami, R., Toya, T., Sekiya, N., Sato, K., 2021. SARS- CoV-2 spike L452R variant evades cellular immunity and increases infectivity. Cell Host Microbe 29 (7), 1124–1136.e11.

M., Faustova, I., Loog, M., 2020. The sequence at Spike S1/S2 site enables cleavage by furin and phospho-regulation in SARS-CoV2 but not in SARS-CoV1 or MERS-CoV. Sci. Rep. 10, 16944.

Papa, G., Mallery, D.L., Albecka, A., Welch, L.G., Cattin-Ortol´a, J., Luptak, J., Paul, D., McMahon, H.T., Goodfellow, I.G., Carter, A., Munro, S., James, L.C., 2021. Furin cleavage of SARS-CoV-2 Spike promotes but is not essential for infection and cell-cell fusion. PLoS Pathog. 17, e1009246.

Peacock, T.P., Goldhill, D.H., Zhou, J., Baillon, L., Frise, R., Swann, O.C., Kugathasan, R., Penn, R., Brown, J.C., Sanchez-David, R.Y., et al., 2021. The furin cleavage site in the SARS-CoV-2 spike protein is required for transmission in ferrets. Nat. Microbiol. 6, 899–909.

Sanches, P., Charlie-Silva, I., Braz, H., Bittar, C., Freitas Calmon, M., Rahal, P., Cilli, E.M., 2021. Recent advances in SARS-CoV-2 Spike protein and RBD mutations comparison between new variants Alpha (B.1.1.7, United Kingdom), Beta (B.1.351, South Africa), Gamma (P.1, Brazil) and Delta (B.1.617.2, India). J. Virus Erad. 7, 100054.

Saputri, D.S., Li, S., van Eerden, F.J., Rozewicki, J., Xu, Z., Ismanto, H.S., Davila, A., Teraguchi, S., Katoh, K., Standley, D.M., 2020. Flexible, functional, and familiar: characteristics of SARS-CoV-2 spike protein evolution. Front. Microbiol. 11, 2112.

Shang, J., Wan, Y., Luo, C., Ye, G., Geng, Q., Auerbach, A., Li, F., 2020a. Cell entry mechanisms of SARS-CoV-2. Proc. Natl. Acad. Sci. U. S. A. 117, 11727–11734.

Shang, J., Ye, G., Shi, K., Wan, Y., Luo, C., Aihara, H., Geng, Q., Auerbach, A., Li, F., 2020b. Structural basis of receptor recognition by SARS-CoV-2. Nature 581, 221–224.

Shorthouse, D., Hall, B.A., 2021. SARS-CoV-2 variants are selecting for spike protein mutations that increase protein stability. J. Chem. Inf. Model. 61, 4152–4155.

Takeda, M., 2021. Proteolytic activation of SARS-CoV-2 spike protein. Microbiol. Immunol. 66.

Tzou, P.L., Tao, K., Nouhin, J., Rhee, S.-Y., Hu, B.D., Pai, S., Parkin, N., Shafer, R.W., 2020. Coronavirus antiviral research database (CoV-RDB): an online database designed to facilitate comparisons between candidate anti-coronavirus compounds. Viruses 12, 1006.

van der Lee, R., Buljan, M., Lang, B., Weatheritt, R.J., Daughdrill, G.W., Dunker, A.K., FuXreiter, M., Gough, J., Gsponer, J., Jones, D.T., et al., 2014. Classification of intrinsically disordered regions and proteins. Chem. Rev. 114, 6589–6631.

Walls, A.C., Park, Y.J., Tortorici, M.A., Wall, A., McGuire, A.T., Veesler, D., 2020. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. Cell 181, 281–292.

Wang, Q., Zhang, Y., Wu, L., Niu, S., Song, C., Zhang, Z., Lu, G., Qiao, C., Hu, Y., Yuen, K.Y., et al., 2020. Structural and functional basis of SARS-CoV-2 entry by using human ACE2. Cell 181, 894–904.e9.

Whittaker, G.R., 2021. SARS-CoV-2 spike and its adaptable furin cleavage site. Lancet Microbe 2, e488–e489.
Wu, Y., Zhao, S., 2020. Furin cleavage sites naturally occur in coronaviruses. Stem Cell Res. 50, 102115.

Wu, C., Zheng, M., Yang, Y., Gu, X., Yang, K., Li, M., Liu, Y., Zhang, Q., Zhang, P., Wang, Y., et al., 2020. Furin: a potential therapeutic target for COVID-19. iScience 23, 101642.

Xia, X., 2021. Domains and functions of spike protein in SARS-CoV-2 in the context of vaccine design. Viruses 13, 109.

Zahradník, J., Marciano, S., Shemesh, M., Zoler, E., Harari, D., Chiaravalli, J., Meyer, B., Rudich, Y., Li, C., Marton, I., Dym, O., et al., 2021. SARS-CoV-2 variant prediction and antiviral drug design are enabled by RBD in vitro evolution. Nat. Microbiol. 6, 1188–1198.

ΒΙΟΓΡΑΦΙΚΟ ΣΗΜΕΙΩΜΑ

ΠΡΟΣΩΠΙΚΑ ΣΤΟΙΧΕΙΑ

Ονοματεπώνυμο	Ευάγγελος Κοντοπόδης		
Ημ/νία γέννησης	03/08/1988	**Τόπος γέννησης**	Ηράκλειο Κρήτης
Διεύθυνση	Ανωγείων 77, Τ.Κ.: 71304, Ηράκλειο Κρήτης		
Τηλέφωνα επικοινωνίας	2810258806, 6947873337	**E-mail**	kontopodisv@hotmail.gr
Στρατιωτικές Υποχρεώσεις	Εκπληρωμένες		

ΕΚΠΑΙΔΕΥΣΗ

2016 έως σήμερα Διδακτορικό Δίπλωμα - Εθνικού και Καποδιστριακού Πανεπιστημίου Αθηνών στο Πρόγραμμα Μεταπτυχιακών Σπουδών του τμήματος Βιολογίας.

Γνωστικό αντικείμενο Διδακτορικού Διπλώματος: Αποκωδικοποίηση του Ανθρώπινου UNIQOME, Εξελικτικές, Μηχανιστικές και θεραπευτικές Προσεγγίσεις.

2012-2015 Μεταπτυχιακό Δίπλωμα Ειδίκευσης - Εθνικού και Καποδιστριακού Πανεπιστημίου Αθηνών, Τμήμα Πληροφορικής και Τηλεπικοινωνιών - στη Βιοπληροφορική (8,55/10)

Θέμα μεταπτυχιακής εργασίας: Ανάλυση των πεπτιδίων μοναδικής αλληλουχίας αμινοξέων (core unique peptides) στο ανθρώπινο πρωτέωμα

2006-2012 Πτυχίο Πανεπιστημίου Πατρών - Πολυτεχνική Σχολή, Τμήμα Μηχανικών Ηλεκτρονικών Υπολογιστών και Πληροφορικής (6,81/10)

Θέμα πτυχιακής εργασίας: «Η κατάσταση του e-banking στην Ελλάδα και η αξιοπιστία των συναλλασσόμενων»

2006 Απολυτήριο Λυκείου (4ο Ενιαίο Λύκειο Ηρακλείου) (17,82/20)

ΓΛΩΣΣΕΣ

Αγγλικά	First Certificate in English, University of Cambridge (ESOL) Edexcel, University of London, England, UK

ΕΡΓΑΣΙΑΚΗ ΕΜΠΕΙΡΙΑ

2/2022 έως 9/2022	Συμμετοχή με υποτροφία στο Πρόγραμμα «Foodomics-GR: Ενδελεχής Χαρακτηρισμός Τροφίμων», του τμήματος I.I.B.E.A.A.
02/2017 έως σήμερα	Διοικητικός Υποδιευθυντής και Υπεύθυνος I.T. (information technology) στην ιδιωτική κλινική Creta Inter Clinic (Ηράκλειο Κρήτης)
06/2016 έως 02/2017	Αναλυτής – Προγραμματιστής Ηλεκτρονικών Υπολογιστών στο Κέντρο Πληροφορικής Υποστήριξης Ελληνικού Στρατού (ΚΕΠΥΕΣ), για την εκπλήρωση της στρατιωτικής θητείας
10/2013 έως 05/2016	Υπεύθυνος I.T. (information technology) στην ιδιωτική κλινική Creta Inter Clinic (Ηράκλειο Κρήτης)

Δημοσιεύσεις

1. Dataset of milk whey proteins of three indigenous Greek sheep breeds.
Anagnostopoulos AK, Katsafadou AI, Pierros V, **Kontopodis E**, Fthenakis GC, Arsenos G, Karkabounas SCh, Tzora A, Skoufos I, Tsangaris GT.
Data Brief. 2016 Jun 29;8:877-80. doi: 10.1016/j.dib.2016.06.040. eCollection 2016 Sep.

2. Dataset of milk whey proteins of two indigenous greek goat breeds.
Anagnostopoulos AK, Katsafadou AI, Pierros V, **Kontopodis E**, Fthenakis GC, Arsenos G, Karkabounas SCh, Tzora A, Skoufos I, Tsangaris GT.
Data Brief. 2016 Jun 28;8:692-6. doi: 10.1016/j.dib.2016.06.038. eCollection 2016 Sep.

3. Milk of Greek sheep and goat breeds; characterization by means of proteomics.
Anagnostopoulos AK, Katsafadou AI, Pierros V, **Kontopodis E**, Fthenakis GC, Arsenos G, Karkabounas SC, Tzora A, Skoufos I, Tsangaris GT.
J Proteomics. 2016 Sep 16;147:76-84. doi: 10.1016/j.jprot.2016.04.008. Epub 2016 Apr 14.

4. Unique Peptide Signatures Of SARS-CoV-2 Against Human Proteome Reveal Variants' Immune Escape And Infectiveness
Evangelos Kontopodis, Vasileios Pierros, Dimitrios J. Stravopodis and George T.Tsangaris
BioRxiv 2021, October. DOI: 10.1101/2021.10.03.462911

5. Prediction of SARS-CoV-2 Omicron Variant Immunogenicity, Immune Escape and Pathogenicity, through Analysis of Spike Protein-specific Core Unique Peptides
Evangelos Kontopodis, Vasileios Pierros, Dimitrios J. Stravopodis and George T.Tsangaris
medRxiv 2021, December. https://doi.org/10.1101/2021.12.26.21268418

6. Prediction of SARS-CoV-2 Omicron Variant Immunogenicity, Immune Escape and Pathogenicity, through the Analysis of Spike Protein-Specific Core Unique Peptides.
Evangelos Kontopodis, Vasileios Pierros, Dimitrios J. Stravopodis and George T.Tsangaris
Vaccines 2022, 10, 357. https://doi.org/10.3390/vaccines10030357

7. Unique peptide signatures of SARS-CoV-2 virus against human proteome reveal variants' immune escape and infectiveness
Vasileios Pierros, **Evangelos Kontopodis,** Dimitrios J. Stravopodis and George T.Tsangaris
Heliyon 2022, Volume 8, Issue 4, E09222, April 01, 2022.
https://doi.org/10.1016/j.heliyon.2022.e09222

8. Unique Peptides of Cathelicidin-1 in the Early Detection of Mastitis—In Silico Analysis
Bourganou, M.V.; **Kontopodis, E.**; Tsangaris, G.T.; Pierros, V.; Vasileiou, N.G.C.; Mavrogianni, V.S.; Fthenakis, G.C.; Katsafadou, A.I.
Int. J. Mol. Sci. 2023, *24*, 10160. https://doi.org/10.3390/ijms241210160

Παρουσιάσεις σε Συνέδρια

1. Insights of the Human Proteome Based on Unique Peptide Analysis.
E. Kontopodis, V. Pierros, G. M. Spyrou and G. Th. Tsangaris
Oral presentation, EUPA Annual Meeting, 2015, 22-26 of June, Milano Italy. Abstract Book, Page 50

2. Insights of the Human Proteome Based on Unique Peptide Analysis.
E. Kontopodis, V. Pierros, A. K. Anagnostopoulos, G. Spyrou and G. Th. Tsangaris.
Oral presentation, 66ο Πανελλήνιο Συνέδριο ΕΕΒΜΒ, 2015, 11-13 Δεκεμβρίου, Αθήνα, Ελλάδα

3. Comparative Analysis of the Human, Rat and Mouse Uniquomes
E. Kontopodis, V. Pierros, A. K. Anagnostopoulos, D. J. Stravopodis, I. S. Papassideri, C. E. Vorgias and G. Th. Tsangaris
Oral presentation, 69ο Πανελλήνιο Συνέδριο ΕΕΒΜΒ, 2018, 23-25 Νοεμβρίου, Λάρισα, Ελλάδα.
Abstract Book, Page 12

4. Data processing approach for the construction and evaluation of an organisms UNIQUOME with comparative analysis for the Human, Rat and Mouse Uniquomes
E. Kontopodis, V. Pierros, A. K. Anagnostopoulos, D. J. Stravopodis, I. S. Papassideri, C. E. Vorgias and G. Th. Tsangaris
Poster presentation, EUPA Proteomic Forum, 2019, 24-28 of March, Potsdam Germany.

5. Data processing approach for the construction and evaluation of an organisms UNIQUOME with comparative analysis for the Human, Rat and Mouse Uniquomes
E. Kontopodis, V. Pierros, A. K. Anagnostopoulos, D. J. Stravopodis, I. S. Papassideri, C. E. Vorgias and G. Th. Tsangaris
Oral presentation, ItPa XIV Italian Proteomics Association Annual meeting, 2019, 25-27 of June, Catanzaro, Italy.
Abstract Book, Page 36

6. Constructing the EvolUniquome: an Evolutionary Process of Progressive Complexity from Popular Model-organism to Human-species Uniquome
E. Kontopodis, V. Pierros, A. K. Anagnostopoulos, I. S. Papassideri, C. E. Vorgias, D. J. Stravopodis and G. Th. Tsangaris
Poster presentation, 70ο Πανελλήνιο Συνέδριο ΕΕΒΜΒ, 2019, 29 Νοεμβρίου - 1 Δεκεμβρίου, Αθήνα, Ελλάδα
Abstract Book, Page 38

7. Uniquome: deciphering the Human Uniquome
E. Kontopodis, V. Pierros, I. S. Papassideri, C. E. Vorgias, D. J. Stravopodis and G. Th. Tsangaris
Oral presentation, ItPa XV Congress Roma, September 8-10, 2021, Italy
Abstract Book, Page38

8. Uniquome: deciphering the Human Uniquome
E. Kontopodis, V. Pierros, I. S. Papassideri, C. E. Vorgias, D. J. Stravopodis and G. Th. Tsangaris
Poster presentation, PROTEOMIC FORUM | EuPA 2022 | XIV Annual Congress European Proteomic Association Proteomics Association, 3–7 April 2022 Leipzig
Abstract Book, Page 202, P189