Κεφάλαιο 2

Βιολογικές Βάσεις Δεδομένων

Σύνοψη
Στο κεφάλαιο αυτό, θα γίνει η απαραίτητη εισαγωγή στις βιολογικές βάσεις δεδομένων έτσι ώστε ο αναγνώστης να μπορεί, στα επόμενα κεφάλαια, να ανατρέχει στις πηγές που χρησιμοποιούνται για την ανάλυση των αντίστοιχων κάθε φορά δεδομένων (αλληλουχίες, δομές, οικογένειες πρωτεϊνών, δεδομένα έκφρασης, πολυμορφισμοί κ.ο.κ.) Ανάλογα με το είδος της πληροφορίας που περιέχουν, θα παρουσιαστούν οι κύριες βάσεις κάθε κατηγορίας και θα τονιστούν τα βασικά χαρακτηριστικά τους. Ειδικό κομμάτι στο τέλος του κεφαλαίου, θα αφιερωθεί στις εξειδικευμένες βάσεις (κυριώς πρωτεϊνικών) δεδομένων, οι οποίες καταλαμβάνουν σημαντικό μερίδιο στην έρευνα των μικρών και μεσαίου μεγέθους ερευνητικών εργαστηρίων και αποτελούν σημαντικό εργαλείο στη βιοπληροφορική μελέτη των πρωτεϊνών.

Προαπαιτούμενη γνώση
Προαπαιτούμενο για το κεφάλαιο αυτό, είναι η εξοικείωση με τις βασικές γνώσεις και έννοιες της μοριακής βιολογίας (DNA, RNA, πρωτεϊνες κλπ).

2. Εισαγωγή

Οι βιολογικές βάσεις δεδομένων, αποτελούν βασικό κομμάτι της σύγχρονης βιοπληροφορικής, καθώς αποτελούν τη βασική πηγή δεδομένων από την οποία ένας ερευνητής αντλεί τα δεδομένα στα οποία θα βασίσει την ανάλυση του. Ακόμα και για αυτούς οι οποίοι παράγουν οι ίδιοι πρωτογενή δεδομένα, η ύπαρξη τους είναι σημαντική καθώς τις περισσότερες φορές είναι αναγκασμένοι να καταθέτουν τα δεδομένα τους σε αυτές, έτσι ώστε να είναι διαθέσιμα στην επιστημονική κοινότητα. Όταν δημιουργήθηκαν οι πρώτες βάσεις δεδομένων ο όγκος της πληροφορίας ήταν μικρός, με αποτέλεσμα η συντήρηση και η ανανέωση των βάσεων να απαιτεί μικρό κόστος τόσο σε υποδομές όσο και σε ανθρώπινο δυναμικό. Η πρόσβαση στις εγγραφές γινόταν μέσω επικοινωνίας με τους επιστημονικούς υπευθύνους της βάσης, οι οποίοι συνήθως έστελναν στον ενδιαφερόμενο όλη την βάση αποθηκευμένη σε δισκέτες ή μαγνητοταινία, με συμβατικό ταχυδρομείο.

Η τεχνολογική εξέλιξη όμως οδήγησε στην αύξηση του όγκου των πειραματικών εργασιών και της διεκπεραίωσής τους, που σε συνδυασμό με τον διαρκή προσδιορισμό γονιδιωμάτων διαφόρων οργανισμών, αύξησε σημαντικά τον όγκο της πληροφορίας σε όλα τα επίπεδα και ιδιαίτερα στο επίπεδο της αλληλουχίας. Στις μέρες μας οι βάσεις περιέχουν πολύ μεγάλο όγκο δεδομένων ενώ είναι απαραίτητο να ανανεώνονται καθημερινά. Η συντήρηση μιας βάσης απαιτεί μεγάλο αριθμό εξειδικευμένων επιστημόνων που θα ασχολούνται αποκλειστικά με την επισήμανση ενδεχόμενων λαθών καθώς και με το σχολιασμό (annotation) των νεοεισερχόμενων δεδομένων.

Δυο χαρακτηριστικά παραδείγματα βάσεων αποτελούν η UiprotKB/SWISS-PROT, η κύρια βάση πρωτεϊνικών αλληλουχιών που περιέχει 547.599 αλληλουχίες (Rel. 2015_02 – Φεβρουάριος 2015) και η EMBL Nucleotide Sequence Database που περιέχει νουκλεοτιδικές αλληλουχίες και έχει 510.014.239 εγγραφές (Rel. 122 - Νοέμβριος 2014). Κάθε ερευνητής μπορεί να έχει πρόσβαση στις βάσεις αυτές μέσω της χρήσης διαδικτύου. Αρκεί η επίσκεψη στην ιστοσελίδα της βάσης, η αναζήτηση των δεδομένων ενδιαφέροντος και στη συνέχεια η αποθήκευσή τους στον υπολογιστή. Παράλληλα έχουν δημιουργηθεί βάσεις στις οποίες η πληροφορία στο επίπεδο της αλληλουχίας και της δομής είναι ταξινομημένες ώστε η πληροφορία να είναι οργανωμένη για την εξαγωγή συμπερασμάτων ως προς την βιολογική τους σημασία.

Οι βιολογικές βάσεις δεδομένων, γενικά, μπορούν να διακριθούν σε 2 μεγάλες κατηγορίες, με επιμέρους κατηγοριοποιήσεις, όπως περιγράφονται παρακάτω:

1. Πρωτογενείς βάσεις δεδομένων, οι οποίες περιέχουν τα πρωτογενή πειραματικά δεδομένα που αναλύονται κυρίως σε:
    Α) Βάσεις δεδομένων νουκλεοτιδικών αλληλουχιών
    Β) Βάσεις δεδομένων αμινοξικών αλληλουχιών πρωτεϊνών
    Γ) Βάσεις δεδομένων τρισδιάστατων βιολογικών δομών
    Δ) Βάσεις δεδομένων γονιδιακής έκφρασης

Ε) Βάσεις δεδομένων γενετικής ποικιλομορφίας
ΣΤ) Βάσεις δεδομένων βιβλιογραφίας

2. Δευτερογενείς βάσεις δεδομένων, στις οποίες υπάρχουν κυρίως ταξινομήσεις των πρωτογενών δεδομένων, χρήσιμες για αναλυτικούς σκοπούς, και διακρίνονται περαιτέρω σε:
   Α) Βάσεις δεδομένων οικογενειών (κυρίως πρωτεϊνών)
   Β) Εξειδικευμένες βάσεις δεδομένων

2.1 Πρωτογενείς βάσεις δεδομένων

Οι πρωτογενείς βάσεις δεδομένων, είναι οι βάσεις που περιέχουν τα βιολογικά δεδομένα όπως αυτά προσδιορίζονται πειραματικά, και συνήθως περιέχουν επιπλέον ταξινόμηση και σχολιασμό. Γενικά, θα μπορούσαμε να τοποθετήσουμε σε αυτή την κατηγορία τις γενικές βάσεις δεδομένων αλληλουχιών, δομών, δεδομένων έκφρασης, γενετικής ποικιλομορφίας αλλά και για λόγους που θα γίνουν κατανοητοί αργότερα, και τις βάσεις δεδομένων βιβλιογραφίας.

2.1.1 Βάσεις δεδομένων νουκλεοτιδικών αλληλουχιών

Ο όγκος της πληροφορίας που περιέχεται στις βάσεις δεδομένων νουκλεοτιδικών αλληλουχιών καθώς και ο εκθετικός ρυθμός συσσώρευσης των δεδομένων που εμφανίζουν, τις έχουν καταστήσει ως τις μεγαλύτερες βάσεις της Βιολογίας. Η εξέλιξη της τεχνολογίας στην εύρεση της αλληλουχίας (sequencing) κυρίως του DNA αλλά και δευτερευόντως του RNA οδήγησε στον προσδιορισμό της αλληλουχίας ολόκληρων γονιδιωμάτων αρκετών οργανισμών (π.χ. ο άνθρωπος) και τη δημιουργία εξειδικευμένων βάσεων δεδομένων που περιέχουν τις αλληλουχίες για έναν και μόνο από αυτούς.

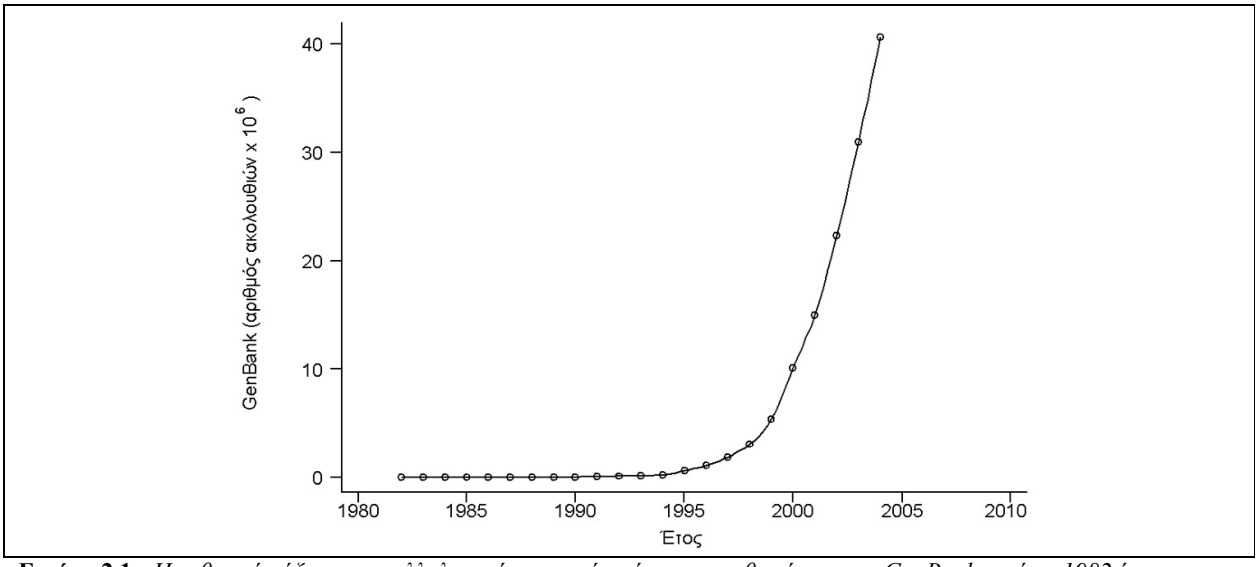
Οι τρεις μεγαλύτερες βάσεις δεδομένων νουκλεοτιδικών αλληλουχιών που είναι ελεύθερα διαθέσιμες στην ακαδημαϊκή κοινότητα είναι οι GENBANK (NCBI), DNA Data Bank of Japan (DDBJ) και EMBL Nucleotide Sequence Database (EBI). Οι τρεις αυτές βάσεις, βρίσκονται σε συνεργασία, δηλαδή ανταλλάσσουν σε καθημερινή βάση τις εγγραφές που κατατίθενται ανεξάρτητα σε καθεμία, έχοντας θέσει παράλληλα κοινούς κανόνες ταξινόμησης και σχολιασμού δεδομένων. Από αυτήν την συνεργασία έχει δημιουργηθεί η International Nucleotide Sequence Database Collaboration. Παρακάτω παρουσιάζονται τα βασικά χαρακτηριστικά των βάσεων δεδομένων που συμμετέχουν στην International Nucleotide Sequence Database Collaboration :

GENBANK: Η GENBANK (http://www.ncbi.nlm.nih.gov/Genbank/index.html) είναι μια βάση νουκλεοτιδικών αλληλουχιών (Benson et al., 2014), διατίθεται ελεύθερα στην επιστημονική κοινότητα και βρίσκεται και υπό την αιγίδα του Εθνικού Ινστιτούτου Υγείας των Η.Π.Α (National Institutes of Health). Τα δεδομένα της βάσης προέρχονται από υποβολές δεδομένων διαφόρων ερευνητικών ομάδων όπως αυτά προκύπτουν από πειραματικές διεργασίες. Η διαδικασία υποβολής γίνεται με την συμπλήρωση κατάλληλης φόρμας μέσω διαδικτύου. Τα δεδομένα που υποβάλλονται στην βάση επεξεργάζονται, σχολιάζονται (annotate) από τους υπεύθυνους της βάσης και στη συνέχεια δημοσιοποιούνται σε αυτήν. Σε συχνά χρονικά διαστήματα τα δεδομένα που έχουν καταχωρηθεί στη βάση επανεξετάζονται και διορθώνονται σε περίπτωση που έχουν προκύψει νέα δεδομένα. Ο αριθμός των νουκλεοτιδικών βάσεων που περιέχονται στην GENBANK διπλασιάζεται κάθε 14 μήνες με αποτέλεσμα η τελευταία έκδοση (Rel. 206, Φεβρουάριος 2015) να περιέχει 181.336.445 αλληλουχίες και 187.893.826.750 συνολικό αριθμό βάσεων.

EMBL-Bank: Η EMBL Nucleotide Sequence Database (http://www.ebi.ac.uk/embl/) αποτελεί τη μεγαλύτερη βάση νουκλεοτιδικών αλληλουχιών στην Ευρώπη, βρίσκεται υπό την αιγίδα του Ευρωπαϊκού Εργαστηρίου Μοριακής Βιολογίας (EMBL) ενώ εδράζεται και συντηρείται από το Ευρωπαϊκό Ινστιτούτο Βιοπληροφορικής (EBI) στο Cambridge, UK. Οι αλληλουχίες κατατίθενται στην EMBL-Bank μέσω διαδικτύου, ακολουθώντας μία απλή διαδικασία από ανεξάρτητα ερευνητικά εργαστήρια ή ομάδες που ασχολούνται με τον προσδιορισμό των γονιδιωμάτων διαφόρων οργανισμών. Αντίστοιχα με την GENBANK, οι νέες καταχωρήσεις αλληλουχιών επεξεργάζονται, σχολιάζονται από τους υπεύθυνους της βάσης και δημοσιοποιούνται. Παράλληλα διατίθενται διάφορα εργαλεία ανάλυσης των δεδομένων όπως το Fasta και το BLAST. Η παρούσα έκδοση της EMBL-Bank (Rel. 122 - Νοέμβριος 2014) περιέχει 510.014.239 εγγραφές. Ο συνολικός αριθμός νουκλεοτιδίων φτάνει τα 1.094.969.877.589.

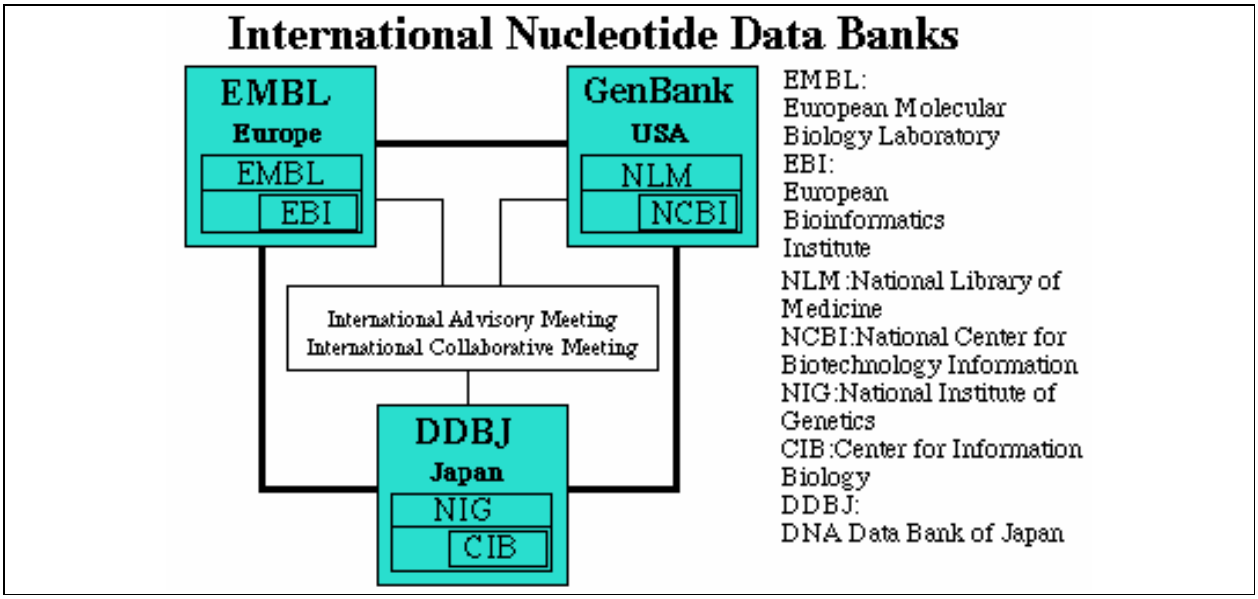
DDBJ: Η DNA Databank of Japan (DDBJ - http://www.ddbj.nig.ac.jp/) είναι η μοναδική διεθνώς αναγνωρισμένη βάση νουκλεοτιδικών αλληλουχιών στην Ιαπωνία. Ιδρύθηκε το 1986 στο Εθνικό Ινστιτούτο Γενετικής (NIG) και βρίσκεται υπό την αιγίδα του Υπουργείου Παιδείας, Επιστημών και Αθλητισμού της

Ιαπωνίας. Βασική πηγή δεδομένων της βάσης αποτελούν οι εργασίες των Ιαπώνων ερευνητών. Επιπλέον στην DDJB είναι διαθέσιμα διάφορα εργαλεία ανάλυσης νουκλεοτιδικών αλληλουχιών. Η παρούσα έκδοση της DDJB (Rel. 99, Δεκέμβριος 2014) περιέχει 178.825.615 εγγραφές και συνολικά 184.410.381.191 νουκλεοτιδικές βάσεις που περιέχονται στις αλληλουχίες.



Εικόνα 2.1: *Η εκθετική αύξηση των αλληλουχιών οι οποίες είναι κατατεθειμένες στην GenBank, από το 1982 έως το τέλος του 2004.*

Οι κυριότερες βάσεις δεδομένων με αλληλουχίες DNA στον διεθνή χώρο, η Genbank, στις ΗΠΑ, η DDBJ στην Ιαπωνία, και η EMBL Data Bank στην Ευρώπη, συνεργάζονται μέσω του International Nucleotide Sequence Collaboration, μιας οργάνωσης που οι ίδιοι δημιούργησαν, και έτσι μια αλληλουχία αφού καταχωρηθεί σε μια από αυτές μέσα από μια διαδικασία εγκρισης, καταχωρείται και στις άλλες. Πρακτική συνέπεια αυτού του γεγονότος, είναι ότι εκτός ελαχίστων εξαιρέσεων, οι 3 βάσεις περιέχουν τις ίδιες καταχωρήσεις, άρα δεν έχει και πολύ μεγάλη σημασία σε ποια από τις 3 βάσεις δεδομένων θα απευθυνθούμε για μια έρευνα.

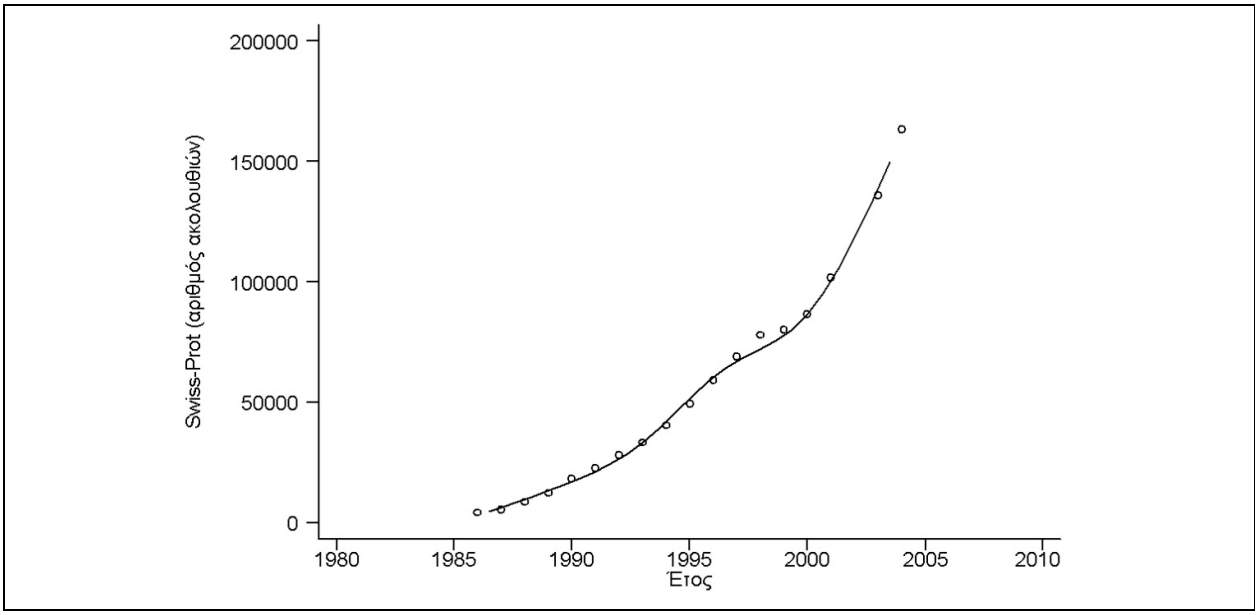


Εικόνα 2.2: *Διάγραμμα που απεικονίζει τη συνεργασία και τη ροή δεδομένων των 3 μεγάλων βάσεων νουκλεοτιδικών αλληλουχιών*

2.1.2 Βάσεις δεδομένων πρωτεϊνικών αλληλουχιών.

Οι βάσεις δεδομένων πρωτεϊνικών αλληλουχιών, αποτελούν το δεύτερο μεγαλύτερο σε όγκο τμήμα του συνόλου των βιολογικών βάσεων δεδομένων (μετά τις αλληλουχίες DNA), αλλά αποτελούν ίσως το σημαντικότερο τμήμα, καθώς οι αμινοξικές αλληλουχίες πρωτεϊνών παρουσιάζουν μεγάλη ποικιλομορφία τόσο στη δομή όσο και στη λειτουργία. Κατά συνέπεια, μεγάλο μέρος της σύγχρονης βιοπληροφορικής ανάλυσης, αναφέρεται σε αυτές και υπάρχει τεράστιος όγκος λειτουργικών δεδομένων που παράγονται συνεχώς πειραματικά, και τα οποία αποτελούν ή θα έπρεπε να αποτελούν μέρος της πληροφορίας που περιέχεται σε αυτές τις βάσεις.

Η **UniprotKB** (Uniprot Knowledgebase, http://www.uniprot.org/), αποτελεί την κύρια, σε παγκόσμιο επίπεδο βάση δεδομένων πρωτεϊνικών αλληλουχιών (UniProt, 2014). Απότελείται από δύο υποσύνολα, την Uniprot/SwissProt η οποία περιέχει τις καλά σχολιασμένες πρωτεϊνικές αλληλουχίες, και την Uniprot/TrEMBL η οποία περιέχει τις πρωτεϊνικές αλληλουχίες που έχουν προκύψει από αυτόματη (ηλεκτρονική) μετάφραση γονιδιωματικών αλληλουχιών. Η UniprotKB/SwissProt η οποία περιέχει 547.599 αλληλουχίες (Rel. 2015_02 – Φεβρουάριος 2015) οι οποίες έχουν περάσει από κάποιου είδους έλεγχο και συνοδεύονται από συμπληρωματικά σχόλια όπως, βιβλιογραφικές αναφορές, γενικά στοιχεία δευτεροταγούς δομής, σύνδεσμοι σε άλλες βάσεις δεδομένων σχετικές με κάθε εγγραφή καθώς και σημειώσεις για τη βιολογική λειτουργία (αν είναι γνωστές) καθώς και άλλες χρήσιμες πληροφορίες. Η Uniprot/TrEMBL περιέχει σήμερα (Rel. 2015_02 – Φεβρουάριος 2015) 92.124.243 αλληλουχίες η οποίες όμως δεν έχουν υποστεί ανθρώπινο σχολιασμό. Περιοδικά, οι σχολιαστές της UniprotKB εντοπίζοντας δεδομένα από τη βιβλιογραφία αλλά και με χρήση αυτοματοποιημένων εργαλείων, αλλάζουν το σχολιασμό των καταχωρήσεων και έτσι μια πρωτεϊνική αλληλουχία ενδέχεται να "περάσει" από την Uniprot/TrEMBL στην Uniprot/SwissProt. Το είδος, το εύρος και η μεγάλη ποικιλομορφία του σχολιασμού που μπορεί να υπάρχει σε επίπεδο πρωτεϊνικής αλληλουχίας είναι τεράστιο (σε ποιο κυτταρικό οργανίδιο υπάρχει, σε ποιον ιστό εκφράζεται, ποια είναι η δευτεροταγής δομή της, ποιος ο βιολογικός της ρόλος, ποια τα μονοπάτια στα οποία εμπλέκεται κ.ο.κ.), και κατά συνέπεια, ο όγκος της πληροφορίας στην Uniprot/SwissProt είναι τεράστιος, όπως επίσης και η πιθανότητα (παρόλες τις προσπάθειες), η πληροφορία αυτή να είναι λαθεμένη ή απλά ελλειπής. Περισσότερα, αναλύονται στην ειδική ενότητα παρακάτω που αφορά στις εξειδικευμένες βάσεις δεδομένων και στα προβλήματα σχολιασμού. Ένα τυπικό αρχείο Uniprot με τις επεξηγήσεις των πιο σημαντικών πεδίων, παρουσιάζεται στο παράρτημα.



Εικόνα 2.3: *Η εκθετική αύξηση των αμινοξικών αλληλουχιών πρωτεϊνών οι οποίες είναι κατατεθειμένες στην Swiss-Prot, από το 1986 έως το τέλος του 2004.*

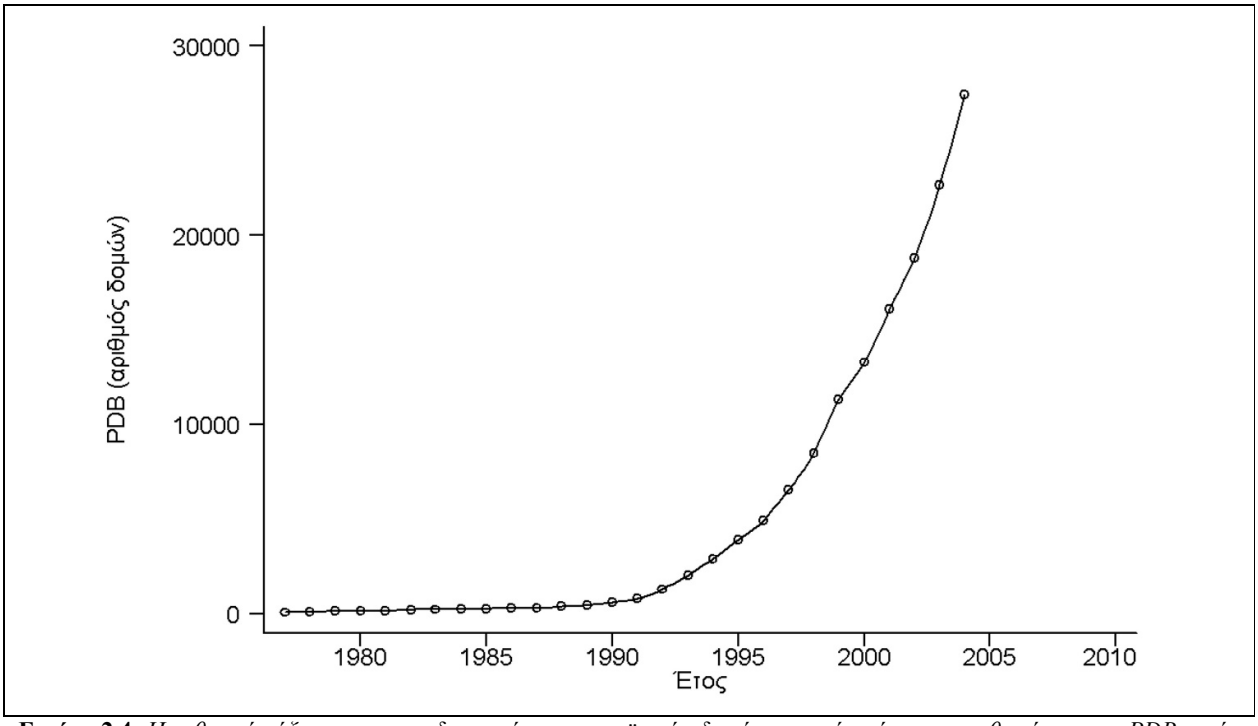
Ιστορικά, αξίζει να αναφερθεί ότι η Uniprot προέκυψε το 2002 από μια συνένωση των δύο μεγαλύτερων τότε βάσεων δεδομένων, της SwissProt και της PIR. Η SwissProt Ιδρύθηκε το 1986 στο Ελβετικό ινστιτούτο Βιοπληροφορικής (Swiss Institute of Bioinformatics) και λειτουργούσε σε συνεργασία με το Ευρωπαϊκό Ινστιτούτο Βιοπληροφορικής (European Bioinformatics Institute). Η **Protein Information Resource** (PIR - http://pir.georgetown.edu/) ήταν η αντίστοιχη Αμερικάνικη βάση δεδομένων. Η έδρα της ήταν στο Πανεπιστήμιο του Georgetown και αποτελούσε τμήμα του Εθνικού Ιδρύματος Βιοϊατρικής Έρευνας (NBRF) των Η.Π.Α. Η κυριότερη βάση που περιέχει είναι η PIR-International Protein Sequence Database (PSD), της οποίας τα δεδομένα προκύπτουν από την συνεργασία της PIR με το Munich Information Center for Protein Sequences (MIPS) και την Japanese International Protein Information Database (JIPID). Το 2002, η PIR σε μια κοινή προσπάθεια με το EBI (European Bioinformatics Institute) και το SIB (Swiss Institute of Bioinformatics) σχημάτισαν το UniProt consortium. Με αυτόν τον τρόπο οι αλληλουχίες της PIR-PSD αλλα και ο σχολιασμός τους ενσωματώθηκαν στην UniProt Knowledgebase. Προστέθηκαν διασυνδέσεις μεταξύ των καταχωρήσεων της UniProt και της PIR-PSD για να διευκολυνθεί ο εντοπισμός παλαιών καταχωρήσεων της PIR-PSD. Πρωτεΐνες που ήταν μοναδικές στην PIR-PSD όπως και οι αναφορές τους αλλά και τα πειραματικά δεδομένα που υπήρχαν στις σχετικές καταχωρήσεις μπορούν πλέον να βρεθούν στις αντίστοιχες καταχωρήσεις της UniProt.

2.1.3 Βάσεις δεδομένων τρισδιάστατων βιολογικών δομών.

Οι βάσεις αυτές περιέχουν δεδομένα που έχουν να κάνουν με την τρισδιάστατη δομή βιολογικών μακρομορίων. Οι τρισδιάστατες δομές αποτελούν το τελικό στάδιο μιας επίπονης διαδικασίας η οποία μετά τη χρήση μοριακών τεχνικών (κλωνοποίηση, απομόνωση, κρυστάλλωση κ.ο.κ.), οδηγεί τελικά στην υπολογιστική επίλυση της δομής μέσω της διαδικασίας της κρυσταλλογραφίας ακτίνων Χ, ή, σε πιο σπάνιες περιπτώσεις με φασματοσκοπία NMR. Το μεγαλύτερο ενδιαφέρον, βέβαια, έχουν οι δομές πρωτεϊνών, καθώς οι πρωτεΐνες είναι τα μακρομόρια των οποίων η μεγάλη ποικιλομορφία της δομής συνδέεται άμεσα με την βιολογική δράση. Η μοναδική βάση αυτού το είδους παγκοσμίως, είναι η PDB, η οποία και αναλύεται παρακάτω.

Protein Data Bank: Η Protein Data Bank (PDB, www.rcsb.org) είναι παγκοσμίως η μοναδική βάση στην οποία περιέχονται τρισδιάστατες δομές βιολογικών μακρομορίων (Kouranov et al., 2006). Ιδρύθηκε το 1971 στα εργαστήρια Brookhaven National Laboratories (BNL) των ΗΠΑ. Αρχικά αποτελούνταν από 7 δομές μακρομορίων οι οποίες προέκυψαν από κρυσταλλογραφικές μελέτες ενώ είχε μικρό ρυθμό αύξησης εγγραφών μέχρι τα τέλη της δεκαετίας του '70. Την δεκαετία του '80 παρατηρήθηκε σημαντική αύξηση του ρυθμού προσθήκης δεδομένων λόγω της τεχνολογικής εξέλιξης σε κάθε στάδιο του προσδιορισμού των δομών, ενώ πλέον η PDB περιέχει και δομές που έχουν προκύψει με φασματοσκοπία Πυρηνικού Μαγνητικού Συντονισμού (NMR). Σήμερα (Φεβρουάριος 2015) η PDB περιλαμβάνει 106.858 δομές βιομορίων. Οι εγγραφές της PDB περιλαμβάνουν εκτός από τις συντεταγμένες των ατόμων που απαρτίζουν τη δομή και επιπρόσθετα βοηθητικά στοιχεία όπως βιβλιογραφικές αναφορές, λεπτομέρειες για τον προσδιορισμό της δομής καθώς και άλλα στοιχεία που προκύπτουν από τη συγκεκριμένη δομή. Κάθε δομή πριν δημοσιευθεί στην βάση ελέγχεται για την ορθότητα της με τη χρήση ειδικού λογισμικού. Στη συνέχεια εφόσον περάσει τις δοκιμές με επιτυχία αποκτά ένα χαρακτηριστικό κωδικό και προστίθεται στη βάση.

Πρέπει να τονιστεί, ότι η καταχώρηση στην PDB είναι η τρισδιάστατη δομή, και όχι η πρωτεΐνη. Κατά συνέπεια, είναι δυνατόν να υπάρχει μια καταχώρηση της PDB η οποία να περιέχει περισσότερες από μία (ακόμα και μερικές δεκάδες) αμινοξικές αλληλουχίες πρωτεϊνών, όπως για παράδειγμα όταν αναφερόμαστε σε πολυενζυμικά σύμπλοκα τα οποία περιέχουν πολλές υπομονάδες. Επίσης, είναι δυνατόν να υπάρχουν περισσότερες από μία δομές μιας συγκεκριμένης πρωτεΐνης, καθώς είναι δυνατόν να έχουν γίνει διαφορετικά πειράματα είτε σε διαφορετικές συνθήκες, είτε παρουσία άλλων παραγόντων, είτε και απλά με άλλη τεχνική για να επιτευχθεί καλύτερη ευκρίνεια. Φυσικά, όπως είναι αναμενόμενο, μόνο ένα μικρό υποσύνολο των γνωστών πρωτεϊνών έχουν γνωστή τρισδιάστατη δομή, γιατί η διαδικασία επίλυσης της δομής είναι χρονοβόρα και δύσκολη. Αυτό φαίνεται ξεκάθαρα αν συγκρίνουμε τον αριθμό των καταχωρήσεων της Uniprot με αυτόν της PDB. Ειδικότερα δε, για κάποιες ειδικές κατηγορίες πρωτεϊνών όπως οι διαμεμβρανικές πρωτεΐνες, τα πράγματα είναι ακόμα πιο δύσκολα από πειραματικής πλευράς και οι τρισδιάστατες δομές τους, είναι ακόμα πιο σπάνιες. Ένα τυπικό αρχείο PDB με τις επεξηγήσεις των πιο σημαντικών πεδίων, παρουσιάζεται στο παράρτημα. Τέλος, αξίζει να αναφερθεί, ότι παρόμοια βάση (MMDB) συντηρείται και στις ΗΠΑ στα πλαίσια του NCBI, με συνεχή όμως επαφή και ενημέρωση από την PDB.



Εικόνα 2.4: *Η εκθετική αύξηση των προσδιορισμένων πρωτεϊνικών δομών οι οποίες είναι κατατεθειμένες στην PDB, από το 1977 έως το τέλος του 2004.*

2.1.4 Βάσεις δεδομένων γονιδιακής έκφρασης

Εκτός από τις βάσεις δεδομένων αλληλουχιών και δομών, σημαντική είναι τα τελευταία χρόνια και η ανάπτυξη των βάσεων δεδομένων γονιδιακής έκφρασης. Με την εξέλιξη της τεχνολογίας και τη δημιουργία νέων οικονομικότερων τσιπ μικροσυστοιχιών, αλλά και με την εμφάνιση των τεχνολογιών Next Generation Sequencing, τα πειράματα ανάλυσης γονιδιακής έκφρασης πραγματοποιούνται με μεγαλύτερο ρυθμό και έτσι υπάρχει ανάγκη αποθήκευσης και ανάλυσης όλων αυτών των δεδομένων. Τη λύση στο παραπάνω πρόβλημα έδωσαν οι βάσεις δεδομένων οι οποίες περιέχουν δεδομένα από χιλιάδες πειράματα μικροσυστοιχιών. Οι βάσεις δεδομένων αυτές επιτρέπουν την καταχώρηση αποτελεσμάτων από πειράματα μικροσυστοιχιων, ενώ κάποιες από αυτές προσφέρουν και επιπλέον εργαλεία ανάλυσης. Επίσης, παρέχουν πληροφορίες σχετικά με το είδος των δεδομένων, την πλατφόρμα μικροσυστοιχίων που χρησιμοποιήθηκε στο πείραμα, τα γονίδια τα οποία μελετώνται καθώς επίσης και πληροφορίες σχετικά με τα είδη των δειγμάτων τα οποία χρησιμοποιήθηκαν. Η βασική δομή αυτών των αρχείων, διαφέρει πολύ από αυτά που αναφέραμε μέχρι τώρα, καθώς έχουμε να κάνουμε με έναν πίνακα, στον οποίο αναγράφονται τιμές "έκφρασης" ενός γονιδίου για κάθε άτομο. Συνήθως τα πειράματα αυτά αφορούν λίγα άτομα, αλλά ανάλογα με την πλατφόρμα μπορούμε να έχουμε δεδομένα έκφρασης για μερικές εκατοντάδες έως μερικές δεκάδες χιλιάδες γονίδια.

Επειδή ο όγκος των δεδομένων γονιδιακής έκφρασης είναι μεγάλος και πολύπλοκος, για να καταχωρηθούν τα δεδομένα των μικροσυστοιχιών στις δημόσιες βάσεις δεδομένων θα πρέπει να ακολουθούν ένα συγκεκριμένο πρωτόκολλο με βάση το οποίο καταχωρείται η ελάχιστη πληροφορία που περιγράφει ένα πείραμα μικροσυστοιχιών (MIAME:Minimun Information About a Microarray Experiment). Τα τελευταία χρόνια, γίνεται μεγάλη προσπάθεια το πρωτόκολλο αυτό να "επιβάλλεται" στους συγγραφείς οι οποίοι πρόκειται να δημοσιεύσουν μια σχετική εργασία. Δηλαδή, πριν η εργασία γίνει αποδεκτή από το επιστημονικό περιοδικό, θα πρέπει οι συγγραφείς να έχουν καταθέσει τα δεδομένα τους σε μια σχετική βάση δεδομένων (κάτι παρόμοιο ισχύει από χρόνια για τις αλληλουχίες και τις δομές μακρομορίων). Οι πιο γνώστες και συχνά χρησιμοποιούμενες βάσεις δεδομένων μικροσυστοιχιών είναι:

GeneExpression Omnibus (GEO): Βάση δεδομένων του NCBI που παρέχει δεδομένα γονιδιακής έκφρασης, τόσο από μικροσυστοιχίες όσο και από αλληλούχιση (next generation sequenicng) (Barrett & Edgar, 2006) Είναι διαθέσιμη στην ιστοσελίδα http://www.ncbi.nlm.nih.gov/geo/, ενώ στην ίδια διεύθυνση

υπάρχουν διαθέσιμα και κάποια διαδικτυακά εργαλεία που επιτρέπουν απλές αναλύσεις των δεδομένων της βάσης. Τα δεδομένα υπάρχουν τόσο σε ακατέργαστη (raw) όσο και σε επεξεργασμένη μορφή (με κανονικοποιήσεις κ.ο.κ.). Η βάση περιέχει (τον Φεβρουάριο του 2015), δεδομένα από 14.031 διαφορετικές πλατφόρμες έκφρασης, προερχόμενα από 1.357.732 "δείγματα", δηλαδή άτομα (στα οποία όμως δεν περιέχονται μόνο άνθρωποι, μπορεί να υπάρχουν δεδομένα από ζώα, φυτά ή ακόμα και μικρο-οργανισμούς), ταξινομημένα 55.725 "σειρές" (series) και 3.848 "σύνολα δεδομένων" (datasets). Το ίδιο δείγμα μπορεί να περιέχεται σε διαφορετικές σειρές και η ίδια σειρά σε ένα ή περισσότερα σύνολα δεδομένων.

Array Express: Δημόσια βάση δεδομένων μικροσυστοιχιων η οποία διατηρείται στο Ευρωπαϊκό Ινστιτούτο Βιοπληροφορικής, EBI, διαθέσιμη στην ιστοσελίδα http://www.ebi.ac.uk/arrayexpress/ (Brazma et al., 2003). Είναι της ίδιας λογικής με την GEO, την οποία περιέχει ως υποσύνολο βάσει της συνεργασίας των ιδρυμάτων. Στην ιστοσελίδα υπάρχουν επίσης διαθέσιμα εργαλεία για ανάλυση, οδηγίες για προγραμματιστική πρόσβαση στις υπηρεσίες και tutorials. Τον Φεβρουάριο του 2015, η βάση περιέχει δεδομένα για 57.009 πειράματα (experiments, τα οποία αντιστοιχούν στα series της GEO) και 1.689.237 μετρήσεις (assays, τα οποία περιέχουν ένα ή περισσότερα δείγματα).

Stanford Microarray Database (SMD): Βάση δεδομένων που κατασκευάστηκε αρχικά για να καλύπτει τις ανάγκες διαμοιρασμού αρχείων των ερευνητών του Stanford, αλλά μετεξελίχθηκε σταδιακά σε ένα δημόσιο αποθετήριο δεδομένων για μικροσυστοιχίες, http://smd.stanford.edu/ (Demeter et al., 2007). Περιέχει μικρότερο αριθμό δεδομένων από τις υπόλοιπες βάσεις, καθώς αυτή τη στιγμή έχει δεδομένα για 84.051 πειράματα από 631 δημοσιεύσεις.

2.1.5 Βάσεις δεδομένων γενετικής ποικιλομορφίας

Οι βάσεις αυτές, αν και συνδέονται στενά με τις βάσεις δεδομένων αλληλουχιών DNA, δεν αποτελούν ευθέως παράγωγα τους, αλλά μάλλον ανεξάρτητες οντότητες. Τούτο είναι κατανοητό αν σκεφτούμε ότι σε μια δεδομένη θέση ενός γονιδιώματος ενός είδους (πχ του ανθρώπου), τα διαφορετικά άτομα είναι δυνατόν να έχουν διαφορετική γενετική πληροφορία (πχ Α αντί για Τ, κ.ο.κ.). Η βάση η οποία καταγράφει τους πολυμορφισμούς και τις συχνότητες τους στους διάφορους πληθυσμούς είναι η dbSNP, ενώ η βάση που καταγράφει πρωτογενώς τουλάχιστον τις αλληλοσυσχετίσεις των πολυμορφισμών αυτών, είναι η HapMap.

dbSNP: Η dbSNP είναι η δημόσια βάση για τους νουκλεοτιδικούς πολυμορφισμούς http://www.ncbi.nlm.nih.gov/snp (Sherry et al., 2001). Εκτός από νουκλεοτιδικούς πολυμορφισμούς (single nucleotide polymorphisms - SNPs), περιέχει και δεδομένα για πολυμορφικές θέσεις που αφορούν απαλοιφές ή εισαγωγές βάσεων (deletion insertion polymorphisms -DIPs), καθώς και για ένθετα μεταθετά στοιχεία και μικροδορυφορικές επαναλήψεις (short tandem repeats - STRs). Κάθε καταχώρηση στην dbSNP περιέχει πληροφορίες για το που βρίσκεται ο πολυμορφισμός (δηλαδή την περιβάλλουσα αλληλουχία), τη συχνότητα του πολυμορφισμού σε διάφορους πληθυσμούς, αλλά και για την πειραματική μέθοδο, τα πρωτόκολλα και τις συνθήκες με τις οποίες μετρήθηκε η ποικιλομορφία. Η dbSNP δέχεται επίσης υποβολές για καταχωρήσεις πολυμορφισμών από κάθε είδος, αλλά και από διαφορετικά σημεία του γονιδιώματος. Λεπτομερής περιγραφή της βάσης δεδομένων υπάρχει στο ελεύθερο διαδικτυακό βιβλίο του NCBI στη διεύθυνση http://www.ncbi.nlm.nih.gov/books/NBK3848/. Στην έκδοση 129 (2008) η βάση είχε πάνω από 14 εκατομύρια πολυμορφισμούς, αλλά προφανώς ο αριθμός αυτός αυξάνεται συνεχώς.

HapMap: Το International HapMap Project (http://hapmap.ncbi.nlm.nih.gov/) είναι το αποτέλεσμα μια διεθνούς συνεργασίας σε μια προσπάθεια να εντοπισθούν και να καταγραφούν οι γενετικές διαφορές αλλά και οι ομοιότητες των ανθρώπινων πληθυσμών (HapMap, 2003). Ο σκοπός του προγράμματος είναι να συγκρίνει τις γενετικές αλληλουχίες διαφορετικών ατόμων (από διαφορετικούς πληθυσμούς) και να εντοπίσει με αυτόν τον τρόπο χρωμοσωμικές περιοχές στις οποίες οι γενετικές παραλλαγές (συνήθως, νουκλεοτιδικοί πολυμορφισμοί), κληρονομούνται μαζί. Στην αρχική φάση του προγράμματος, έγινε χρήση γενετικών δεδομένων από 4 πληθυσμούς Αφρικανικής, Ασιατικής και Ευρωπαϊκής καταγωγής. Σε μεταγενέστερες εκδόσεις, προστέθηκαν και άλλοι πληθυσμοί, σε μια προσπάθεια να υπάρχει όσο το δυνατό μεγαλύτερη κάλυψη παγκοσμίως. Τα τελικά δεδομένα που είναι διαθέσιμα από τη βάση αυτή, είναι οι απλότυποι, δηλαδή οι συνδυασμοί πολυμορφισμών που συνκληρονομούνται, και ακριβέστερα οι συντελεστές ανισορροπίας σύνδεσης (Linkage Disequilibrium), των διαφόρων πολυμορφισμών του ιδίου χρωμοσώματος, μεταξύ τους. Με τη χρήση αυτής της πληροφορίας, είναι δυνατόν να σχεδιαστούν μεθόδοι και αλγόριθμοι στατιστικής γενετικής με τους οποίους θα επιχειρείται να απαντηθούν ερωτήματα σχετικά με τη γενετική προδιάθεση σε

ασθένειες και την ανταπόκριση σε φάρμακα. Επιπλέον, τέτοια δεδομένα είναι πολύ χρήσιμα στη μελέτη της γενετικής δομής των ανθρώπινων πληθυσμών.

2.1.6 Βάσεις δεδομένων βιβλιογραφίας

Παρόλο που οι βάσεις αυτές δεν είναι με την στενή έννοια «βιολογικές βάσεις δεδομένων», ιστορικά, αλλά και για λόγους που θα φανούν στην πορεία, είναι καλό να γίνεται αναφορά και σε αυτές. Οι βάσεις αυτές, έχουν σαν «καταχώρηση» τα στοιχεία μιας επιστημονικής δημοσίευσης (συγγραφέας, περιοδικό, περίληψη κ.ο.κ.). Η κυριότερη βάση του είδους, είναι η **PubMed** (http://www.ncbi.nlm.nih.gov/pubmed) η οποία στεγάζεται στο NCBI και περιλαμβάνει περισσότερα από 24 εκατομύρια καταχωρήσεις επιστημονικών άρθρων από τη βιοϊατρική βιβλιογραφία (έχοντας κάλυψη της MEDLINE, άλλων περιοδικών των επιστημών της ζωής αλλά και από κάποια online βιβλία). Οι αναφορές μπορεί να περιέχουν συνδέσμους στο πλήρες κείμενο των εργασιών, είτε μέσω της PubMed Central (το υποσύνολο με τις ελεύθερα διαθέσιμα δημοσιεύσεις πλήρους κειμένου), είτε απευθείας μέσω των ιστοσελίδων των εκδοτικών οίκων. Παρόλο που τα στοιχεία της PubMed είναι δημόσια διαθέσιμα, το να έχει πρόσβαση κανείς στο πλήρες κείμενο μιας εργασίας, εξαρτάται από την πολιτική του εκδοτικού οίκου. Στον ίδια ιστοσελίδα, υπάρχουν διαθέσιμα και tutorials για τη χρήση της υπηρεσίας (http://www.nlm.nih.gov/bsd/disted/pubmed.html).

Άλλες βάσεις δεδομένων, παρόμοιας φύσης, είναι το SCOPUS (http://www.scopus.com/) και το Web of Science (http://webofknowledge.com/). Οι βάσεις αυτές, παρέχουν περισσότερες πληροφορίες, με την κυριότερη να είναι οι βιβλιογραφικές αναφορές (citations) που έχει πάρει κάθε δημοσιευμένη εργασία. Αυτό επιτρέπει την αντίστροφη αναζήτηση (πχ εύρεση του ποια εργασία έχει αναφέρει μια δεδομένη εργασία), αλλά και την αξιολόγηση του συνολικού έργου (ενός συγγραφέα, ενός περιοδικού ή ενός ιδρύματος). Το βασικότερο μειονέκτημα αυτών των βάσεων είναι ότι διατίθενται από ιδιωτικούς οργανισμούς και απαιτούν συνδρομή του χρήστη είτε του ινστιτούτου του.

Η πρόσβαση στη βιβλιογραφία, εκτός του ότι είναι απαραίτητη εργασία στην καθημερινότητα ενός επιστήμονα, αποτελεί επιπλέον, ένα ιδιαίτερα αναπτυσσόμενο κομμάτι της επιστήμης της πληροφορικής (text mining), το οποίο έχει βρει ιδιαίτερες εφαρμογές στη βιοπληροφορική, καθώς η ύπαρξη ενός τεράστιου όγκου δεδομένων από κείμενα (περιλήψεις εργασιών κυρίως), έχει δώσει την αφορμή για μελέτες αυτών των κειμένων με σκοπό την ανακάλυψη συσχετίσεων και την εξαγωγή βιολογικών συμπερασμάτων (Ananiadou, Kell, & Tsujii, 2006; Scherf, Epple, & Werner, 2005).

2.2 Δευτερογενείς βάσεις δεδομένων

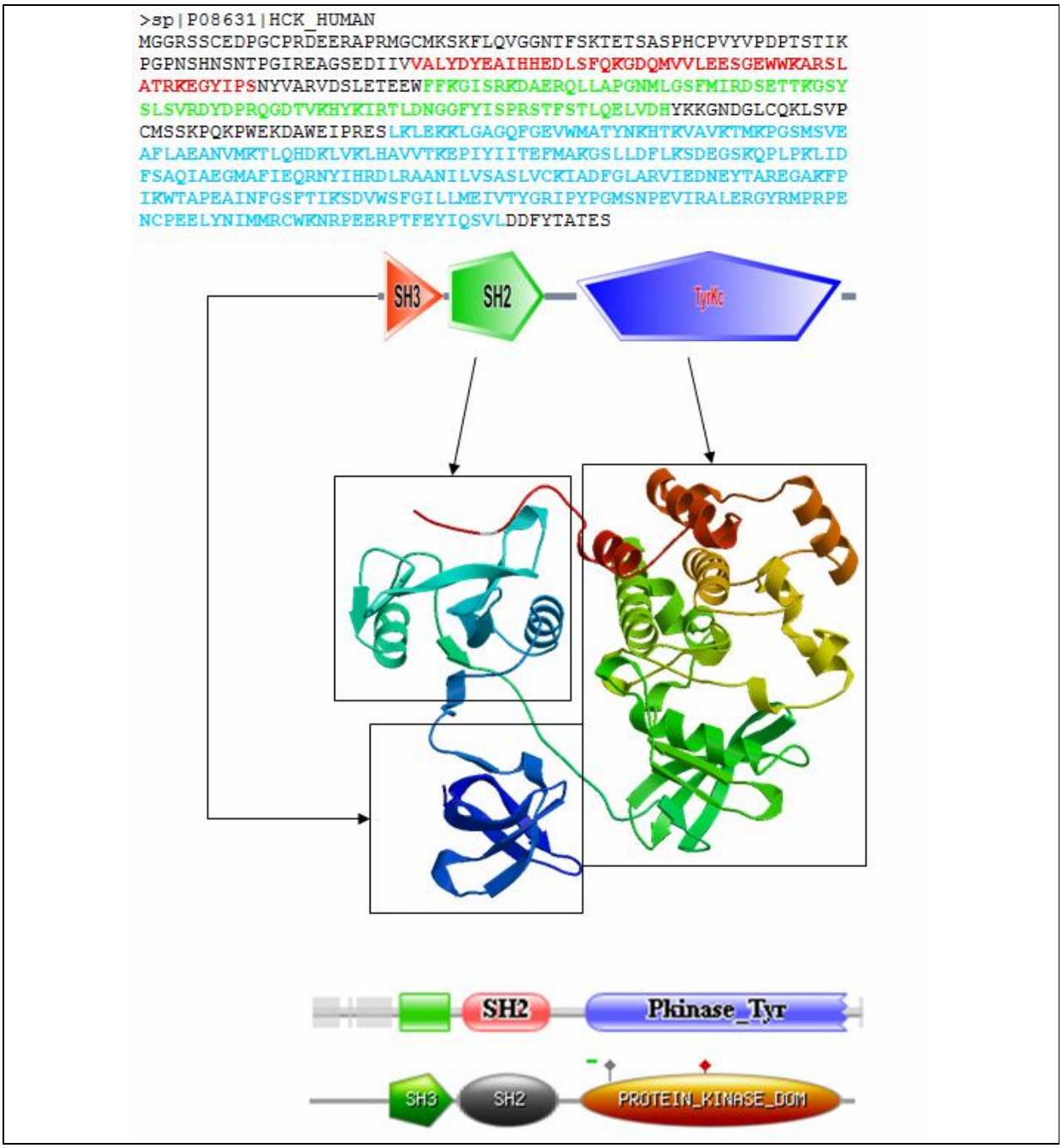
Σε αυτήν την μεγάλη αλλά και ετερογενή κατηγορία περιλαμβάνονται κυρίως βάσεις δεδομένων που περιέχουν διαφόρων ειδών ταξινομήσεις των πρωτογενών δεδομένων, χρήσιμες για αναλυτικούς σκοπούς, και διακρίνονται περαιτέρω σε βάσεις οικογενειών και σε εξειδικευμένες βάσεις δεδομένων.

2.2.1 Βάσεις δεδομένων οικογενειών

Όπως είναι γνωστό, οι πρωτεΐνες γενικά απότελούνται από μία ή περισσότερες διακριτές λειτουργικές περιοχές (domains), οι οποίες πολλές φορές είναι και δομικά αυτοτελείς. Οι περιοχές αυτές, θεωρείται ότι μπορούν να λειτουργήσουν αλλά και και να εξελιχθούν ανεξάρτητα από το υπόλοιπο τμήμα της πρωτεΐνης. Διαφορετικοί συνδυασμοί τέτοιων περιοχών οδηγούν σε μια μεγάλη ποικιλία των πρωτεϊνών στη φύση. Συνεπώς, η ανίχνευση τέτοιων περιοχών είναι σημαντική στην προσπάθεια λειτουργικής ταξινόμησης των πρωτεϊνών. Στο κεφάλαιο της στοίχισης αλληλουχιών θα μιλήσουμε αναλυτικά για το ρόλο που παίζει αυτό το φαινόμενο στην αναζήτηση ομοιότητας αλληλουχιών (τοπική στοίχιση), ενώ στο κεφάλαιο της γονιδιωματικής θα μιλήσουμε για το πως μπορεί η ανίχνευση πρωτεϊνών με διαφορετική σύσταση σε τέτοιες περιοχές να δώσει στοιχεία για τη λειτουργική ή άλλη αλληλεπίδραση μεταξύ πρωτεϊνών μή όμοιων μεταξύ τους.

Οι βάσεις που αναλύονται παρακάτω, επιτελούν πολύ σημαντικό ρόλο στην ταξινόμηση των αμινοξικών αλληλουχιών πρωτεϊνών σε οικογένειες. Επιπλέον δε, πρέπει να έχουμε υπόψη μας, ότι καθώς οι δομές είναι περισσότερο συντηρημένες από τις αλληλουχίες, η ύπαρξη αυτών των βάσεων βοηθάει στην εύκολη ταυτοποίηση και κατηγοριοποίηση νέων πρωτεϊνών, και στην εύκολη αναγνώριση ενός νέος πρωτεϊνικού διπλώματος. Οι βάσεις διαφέρουν μεταξύ τους, κυρίως α) στον τρόπο εύρεσης και μαθηματικής

μοντελοποίησης της περιοχής (με τοπική ομοιότητα, με pattern, με HMM κ.ο.κ.), και β) στον τρόπο με τον οποίο έχει καθοριστεί εξαρχής η περιοχή. Οι CATH και SCOP βασίζονται αποκλειστικά σε δομικά κριτήρια, ενώ οι PROSITE, PFAM, INTERPRO λαμβάνουν υπόψη κυρίως την αλληλουχία. Κατά συνέπεια, περιέχουν μεγαλύτερο αριθμό καταχωρήσεων, καθώς οι πρωτεΐνες με γνωστή δομή είναι πολύ λιγότερες. Επιπλέον δε λόγω αυτού του γεγονότος, είναι δυνατόν, σε κάποιες περιπτώσεις οι περιοχές που έχουν οριστεί να διαφέρουν.



Εικόνα 2.5: *Αναπαράσταση της ανθρώπινης κινάσης τυροσίνης HCK (Uniprot: P08631, PDB: 2HCK_A). Φαίνεται η αμινοξική αλληλουχία, και η διάρθρωση των δομικών αυτοτελών περιοχών (domains) στην τρισδιάστατη δομή. Κάτω, η ίδια πρωτεϊνη όπως την αναπαριστούν οι βάσεις PFAM και PROSITE αντίστοιχα. Καθώς οι περιοχές αυτής της πρωτεΐνης είναι δομικά αυτοτελείς, ίδια αναπαράσταση υπάρχει και στην SCOP. Σε άλλες περιπτώσεις, οι περιοχές που αναπαρίστανται στην PFAM και την PROSITE, μπορεί να μην αντιστοιχούν σε δομικά αυτοτελείς περιοχές, οπότε υπάρχει ενδεχόμενο οι βάσεις αυτές να διαφωνούν μεταξύ τους όσον αφορά στα όρια των περιοχών.*

Η **PROSITE** (http://www.expasy.ch/prosite/) αποτελεί μια βάση ταξινόμησης αμινοξικών αλληλουχιών πρωτεϊνών και αυτοτελών περιοχών αλληλουχιών (sequence domains) σε οικογένειες (Sigrist et al., 2010). Η ταξινόμηση σε οικογένειες πραγματοποιείται βάσει των ομοιοτήτων που παρουσιάζουν οι περιοχές των αλληλουχιών μεταξύ τους. Πρωτεΐνες ή περιοχές που ανήκουν στην ίδια οικογένεια έχουν πιθανότατα την ίδια λειτουργία και προέρχονται από κοινό πρόγονο. Υπάρχουν τμήματα των αμινοξικών αλληλουχιών πρωτεϊνών που είναι περισσότερο συντηρημένα στην πορεία της εξέλιξης τους και σχετίζονται άμεσα με τη λειτουργία τους και με τη δομή των πρωτεϊνών στο χώρο. Η ανάλυση αμινοξικών αλληλουχιών πρωτεϊνών που ανήκουν στην ίδια οικογένεια, μέσω μια πολλαπλής στοίχισης, είναι πιθανό να οδηγήσει σε ένα 'αποτύπωμα' χαρακτηριστικό για κάθε οικογένεια, ικανό να τη διαχωρίζει από τις πρωτεϊνικές αλληλουχίες που δεν ανήκουν σε αυτήν την οικογένεια.

Υπάρχουν γενικά δύο τρόποι για τη δημιουργία των 'αποτυπωμάτων'. Ο ένας βασίζεται στη χρήση μιας γλώσσας παρόμοιας με αυτής των "κανονικών εκφράσεων" (regular expressions), και είναι ο πιο παλιός και εύκολος στη δημιουργία, ενώ ο άλλος βασίζεται στην κατασκευή προφίλ (profiles), πίνακες με ειδικές ανά θέση πιθανότητες εμφάνισης αμινοξέων), μέθοδος η οποία είναι πιο σύνθετη αλλά και πιο ευαίσθητη. Περισσότερα για τις τεχνικές αυτές, θα αναφερθούν σε επόμενο κεφάλαιο. Μέχρι σήμερα η PROSITE περιέχει 'αποτυπώματα' για περίπου 1716 οικογένειες για καθεμία από τις οποίες συμπεριλαμβάνεται λεπτομερής ανάλυση για τη δομή και τη λειτουργία των πρωτεϊνών που την αποτελούν. Συνολικά, υπάρχουν στη βάση 1308 μοτίβα ή πρότυπα (patterns), 1107 προφίλ και 1105 "κανόνες" (αφορούν κυρίως πληροφορίες για το που θα πρέπει να βρίσκεται το μοτίβο για να θεωρηθεί έγκυρο αλλά και πληροφορίες για συνδυασμούς από μοτίβα). Προφανώς, υπάρχουν οικογένειες για τις οποίες υπάρχουν διαθέσιμα και μοτίβα και προφίλ (συνήθως, η παλαιότερες καταχωρήσεις αφορούσαν το μοτίβο). Στην βάση υπάρχουν επίσης, αναλύσεις για τις πρωτεΐνες της Uniprot που ανήκουν σε κάθε οικογένεια όσο και για τις πρωτεΐνες στις οποίες εμφανίζεται ένα "αποτύπωμα" (κυρίως όταν έχουμε να κάνουμε με μοτίβο) αλλά είναι γνωστό ότι αυτές δεν ανήκουν λειτουργικά στην οικογένεια αυτή. Τέλος, υπάρχουν εργαλεία για την αναζήτηση των μοτίβων και των προφίλ σε αλληλουχίες, όσο και εργαλεία αναπαράστασης της "σπονδυλωτής" δομής των πρωτεϊνών, δηλαδή της αναπαράστασης των περιοχών αυτών και την αποτύπωση τους πάνω σε μια δεδομένη αλληλουχία.

PFAM*:* Η βάση Pfam (http://pfam.xfam.org/) αποτελεί μια μεγάλη συλλογή πρωτεϊνικών οικογενειών (Finn et al., 2014). (Andreeva et al., 2004)Βασίζεται στην ίδια λογική με την PROSITE (ειδικά με το υποσύνολο της που βασίζεται σε profiles), αλλά η μεγάλη διαφορά είναι ότι εδώ οι οικογένειες χαρακτηρίζονται από ένα hidden Markov model (HMM), μέθοδος η οποία είναι πιο ευαίσθητη στον εντοπισμό μακρινών ομόλογων, χωρίς όμως να υστερεί σε ταχύτητα και αποτελεσματικότητα. Στην τρέχουσα έκδοση (2013), η βάση περιέχει δεδομένα για 14.831 οικογένειες παρέχοντας κάλυψη για πάνω από το 80% των πρωτεϊνικών καταχωρήσεων της UNIPROT.

Η PFAM αποτελείται από δύο υποσύνολα, την PFAM-A, και την PFAM-B. Η PFAM-A αποτελείται από καταχωρήσεις (οικογένειες) υψηλής «ποιότητας», καθώς έχουν όλες υποστεί σχολιασμό από ειδικούς, ενώ υπάρχουν αναφορές σε άλλες βάσεις δεδομένων και κυρίως σε βιβλιογραφία. Η PFAM-B είναι το υποσύνολο, το οποίο προκύπτει με αυτοματοποιημένο τρόπο εντοπίζοντας τις ομοιότητες ανάμεσα στις πρωτεϊνικές περιοχές που απομένουν όταν αφαιρεθούν οι περιοχές που αντιστοιχούν στις καταχωρήσεις της PFAM-A. Η PFAM-B είναι ιδιαίτερα χρήσιμη, γιατί με στοχευμένη ανάλυση αυτών των «οικογενειών», μπορούν να προκύψουν οικογένειες που μετέπειτα θα «προαχθούν» στην PFAM-A. Το βασικό χαρακτηριστικό της PFAM, και αυτό που την κάνει τόσο δημοφιλή, είναι ότι με τη χρήση του HMM (και ειδικά του πακέτου HMMER, βλ. στο αντίστοιχο κεφάλαιο), μπορεί να επιλεγεί για κάθε οικογένεια μία τιμή διαχωριστικού κατοφλίου στο σκορ, και κατά συνέπεια κάθε πρωτεΐνη ταξινομείται μόνο σε μία οικογένεια (σε αυτή που σκοράρει πάνω από το κατόφλι). Παρόλα αυτά, χαμηλότερη ομοιότητα μπορεί να υπάρχει μεταξύ πρωτεϊνών που ανήκουν σε διαφορετικές οικογένειες, γιαυτό και η βάση περιέχει και μια ανώτερη κατηγορία οργάνωσης, την υπερ-οικογένεια (clan).

CATH: Η CATH (http://www.biochem.ucl.ac.uk/bsm/cath_new/index.html) είναι μια βάση ιεραρχικής ταξινόμησης πρωτεϊνικών δομών που αποτελούν εγγραφές της PDB με βάση τις αυτοτελείς δομικές περιοχές (domains) που τις απαρτίζουν (Knudsen & Wiuf, 2010). Η CATH περιέχει αποκλειστικά πρωτεϊνικές δομές που είναι προσδιορισμένες σε ευκρρίνεια μεγαλύτερη των 3 Angstroms και χρησιμοποιεί κυρίως αυτοματοποιημένες μεθόδους για την ταξινόμησή τους. Σε ειδικές περιπτώσεις και όταν αυτό κρίνεται απαραίτητο χρησιμοποιούνται και ανθρώπινα κριτήρια. Η ιεραρχία αποτελείται κυρίως από τέσσερα επίπεδα: 1) την Τάξη (Class), 2) την Αρχιτεκτονική (Architecture), 3) την Τοπολογία (Οικογένεια διπλώματος) (Topology (fold family)) και 4) την Ομόλογη Οικογένεια (Homologous superfamily). Οι πρωτεΐνες που

αποτελούνται από πάνω από μία αυτοτελείς δομικές περιοχές (domains), αναλύονται στα επιμέρους στοιχεία αυτόματα με βάση ειδικούς αλγόριθμους αναγνώρισης των περιοχών. Η αυτόματη αυτή διαδικασία κατατάσσει το 53% των δομών. Οι υπόλοιπες διαχωρίζονται στις επιμέρους αυτοτελείς δομικές περιοχές με παρατηρήσεις που προκύπτουν είτε από τους αλγόριθμους αυτόματου διαχωρισμού είτε από τη βιβλιογραφία. Η ταξινόμηση πραγματοποιείται μόνο στις αυτοτελείς δομικές περιοχές. Η ανάλυση της ιεραρχίας στην CATH έχει ως εξής:

C - Τάξη (Class): Οι δομές ταξινομούνται σε 4 μεγάλες ομάδες βάσει των στοιχείων δευτεροταγούς δομής των αυτοτελών δομικών περιοχών και είναι οι: 1) mainly-alpha, όπου τα στοιχεία δευτεροταγούς δομής είναι στην συντριπτική τους πλειοψηφία α-έλικες, 2) mainly-beta, όπου τα στοιχεία δευτεροταγούς δομής είναι κυρίως β-εκτεταμένες δομές, 3) alpha-beta, όπου παρατηρούνται εναλλασσόμενες α/β και α+β δομές και 4) δομές με χαμηλό ποσοστό δευτεροταγών δομών. Η διαδικασία της ταξινόμησης γίνεται αυτόματα για το 90% των πρωτεϊνών ενώ για το υπόλοιπο 10% χρησιμοποιούνται κυρίως δεδομένα από τη βιβλιογραφία.

A - Αρχιτεκτονική (Architecture): Η ταξινόμηση πραγματοποιείται βάσει της γενικότερης δομής της αυτοτελούς δομικής περιοχής (domain), λαμβάνοντας υπόψη τον προσανατολισμό των στοιχείων δευτεροταγούς δομής αλλά όχι τον τρόπο διασύνδεσης μεταξύ τους π.χ. βαρέλια (barrels).

T - Τοπολογία (Topology): Οι δομές ομαδοποιούνται με βάση τον προσανατολισμό των στοιχείων δευτεροταγούς δομής καθώς και τον τρόπο σύνδεσής τους.

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

S - Αλληλουχία (Sequence family): Τα μέλη της εμφανίζουν ομοιότητα πάνω από 35% στο επίπεδο της αλληλουχίας με αποτέλεσμα να θεωρούνται ότι έχουν παρόμοια δομή και λειτουργία.

SCOP: Ο βασικός στόχος της βάσης SCOP (http://scop.mrc-lmb.cam.ac.uk/scop/index.html) είναι η ανάλυση των δομικών και εξελικτικών σχέσεων που παρατηρούνται μεταξύ όλων των πρωτεϊνών γνωστής δομής καταχωρημένων στην PDB (Andreeva, et al., 2004). Η ταξινόμηση των πρωτεϊνών πραγματοποιείται βάσει αυτών των δομικών και εξελικτικών σχέσεων. Τα βασικά επίπεδα ταξινόμησης είναι τέσσερα: 1) η οικογένεια (Family), 2) η υπερ-οικογένεια (Superfamily), 3) το δίπλωμα (Fold) και 4) η τάξη (Class).

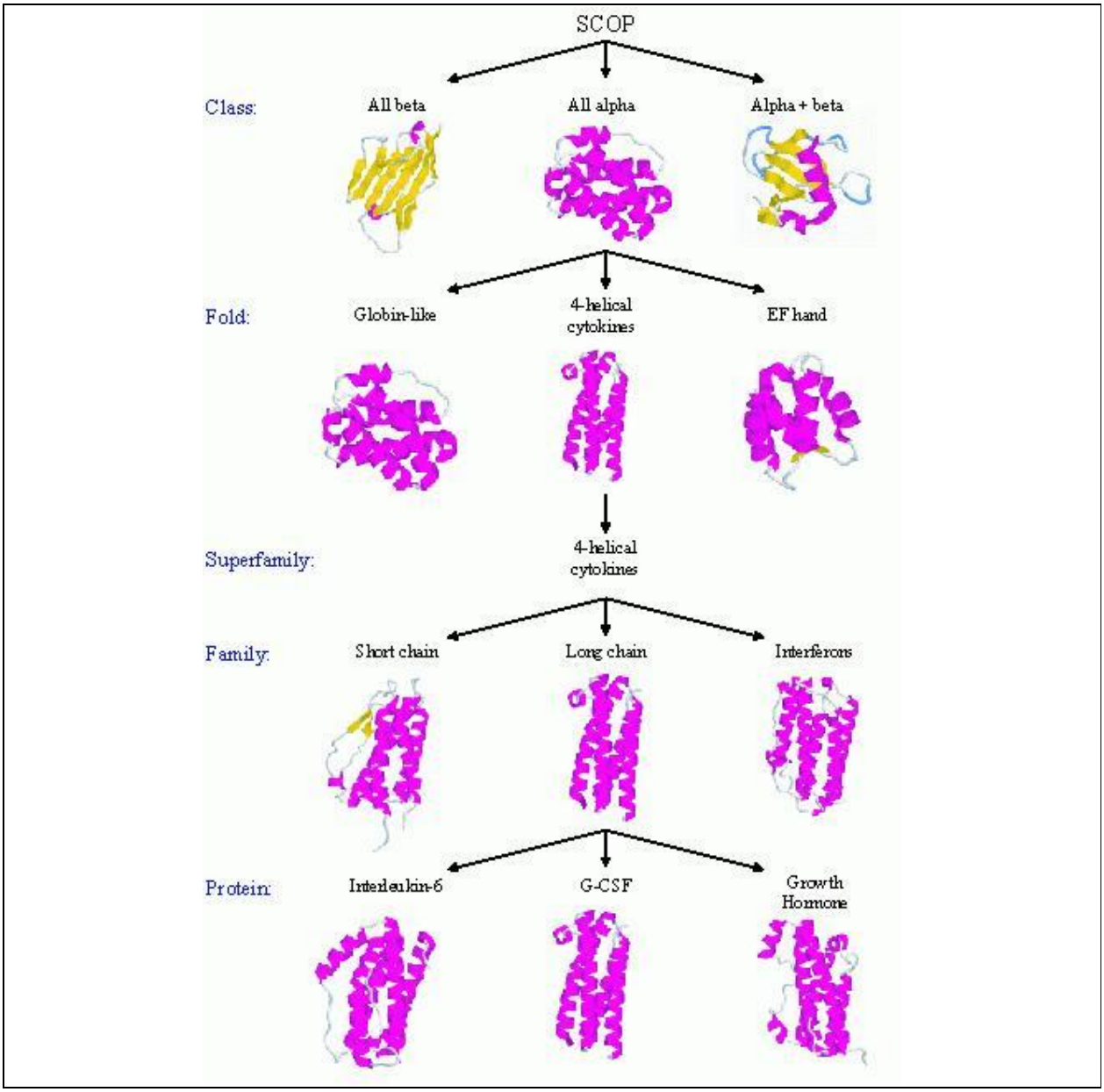
Οικογένεια (Family): Μεταξύ των μελών της οικογένειας παρατηρείται ξεκάθαρη εξελικτική σχέση. Η ομοιότητα σε επίπεδο αλληλουχίας είναι ίση ή μεγαλύτερη του 30%. Παρόλα αυτά υπάρχουν περιπτώσεις στις οποίες οι δομές και η λειτουργία είναι παρόμοιες υποδηλώνοντας κοινό πρόγονο ενώ η ομοιότητα σε επίπεδο αλληλουχίας είναι μικρότερη του 30% (σφαιρίνες, 15%).

Υπερ-οικογένεια (Superfamily): Οι πρωτεΐνες που κατατάσσονται στις υπερ-οικογένειες εμφανίζουν πολύ μικρή ομοιότητα στο επίπεδο της αλληλουχίας αλλά τα δομικά τους χαρακτηριστικά και η λειτουργία τους υποδηλώνουν ότι πιθανά προέλθει από κοινό πρόγονο.

Δίπλωμα (Fold): Σε αυτό το επίπεδο κατατάσσονται πρωτεΐνες που παρουσιάζουν ομοιότητα σε επίπεδο δομής. Οι πρωτεΐνες που εμφανίζουν το ίδιο δίπλωμα έχουν τα ίδια σε μεγάλο βαθμό χαρακτηριστικά δευτεροταγούς δομής, με κοινό προσανατολισμό και τις ίδιες τοπολογικές συνδέσεις μεταξύ τους. Πρωτεΐνες που έχουν το ίδιο δίπλωμα αλλά δεν είναι όμοιες από άποψη αμινοξικής αλληλουχίας έχουν ορισμένα περιφερειακά στοιχεία της δευτεροταγούς τους δομής και στροφές ανόμοια και όσον αφορά στο μέγεθος και όσον αφορά στη διαμόρφωση. Πρωτεΐνες που εμφανίζουν κοινό δίπλωμα δεν είναι απαραίτητο να έχουν κοινή εξελικτική προέλευση.

Τάξη (Class): Η ταξινόμηση γίνεται με βάση το δίπλωμα των στοιχείων δευτεροταγούς δομής των πρωτεϊνών σε τέσσερις κύριες δομικές κατηγορίες: 1) την all-α, όπου η δομή σχηματίζεται από α-έλικες, 2) την all-β, όπου η δομή αποτελείται από β-πτυχωτές επιφάνειες, 3) την α/β, όπου στην δομή της πρωτεΐνης εναλλάσσονται α-έλικες και β-πτυχωτές επιφάνειες και 4) την α+β, όπου σε διακριτές περιοχές της δομής βρίσκονται α-έλικες και β-πτυχωτές επιφάνειες.

Η αναγνώριση των σχέσεων καθώς και η ταξινόμηση βάσει των σχέσεων μεταξύ των πρωτεϊνών πραγματοποιείται αποκλειστικά από ειδικούς επιστήμονες μετά από λεπτομερή μελέτη και σύγκριση των πρωτεϊνικών δομών. Αυτοματοποιημένες μέθοδοι χρησιμοποιούνται μόνο για την ομοιογένεια των δεδομένων που περιέχονται στη βάση.



Εικόνα 2.6: *Παράδειγμα της ιεραρχίας στη βάση SCOP (τάξη, δίπλωμα, υπεροικογένεια, οικογένεια). Προσοχή στο γεγονός ότι για λόγους απλότητας σε κάθε επίπεδο δεν απεικονίζονται όλες οι κατηγορίες, δηλαδή στο επίπεδο του διπλώματος υπάρχουν και άλλες κατηγορίες εκτός των 3 που απεικονίζονται.*

2.2.2 Εξειδικευμένες βάσεις δεδομένων

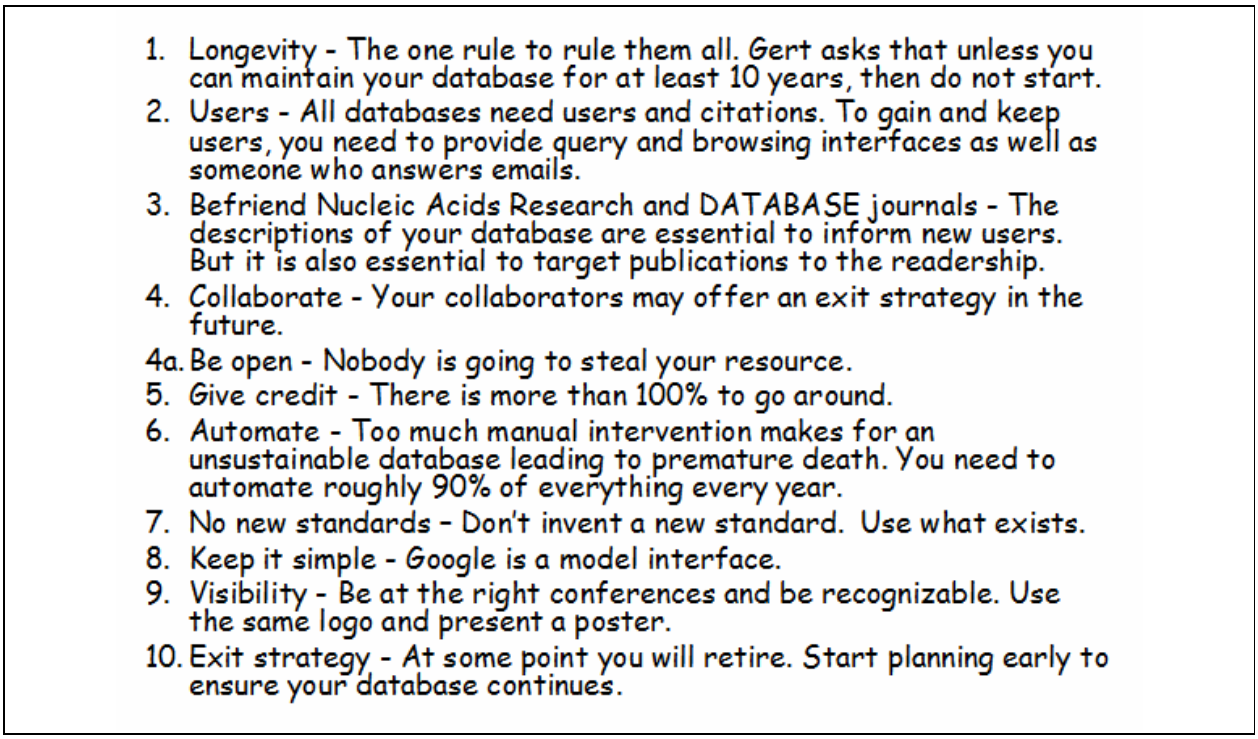
Εκτός από τις μεγάλες, δημόσια διαθέσιμες και ευρέως χρηματοδοτούμενες βάσεις δεδομένων που αναφέρθηκαν παραπάνω, σημαντικό ρόλο στην πρόοδο της βιοπληροφορικής παίζουν και οι εξειδικευμένες βάσεις δεδομένων. Συνήθως, αλλά όχι πάντα, αφορούν τις αμινοξικές αλληλουχίες πρωτεϊνων (γιατί για αυτές υπάρχει μεγάλη πληθώρα λειτουργικών δεδομένων, σε μεγάλη λεπτομέρεια, που δεν μπορεί να καλυφθεί από τις βάσεις όπως η Uniprot), και τις περισσότερες φορές, συντηρούνται από μικρές ή μεσαίου μεγέθους ερευνητικές ομάδες. Στην ενότητα αυτή θα εξεταστούν προβλήματα που αντιμετωπίζουν οι διαχειριστές αυτών των βάσεων δεδομένων και θα συζητηθούν οι λόγοι που οι επιστήμονες μπορεί να προτιμούν να δημοσιεύουν τα δεδομένα τους σε βάσεις δεδομένων αντί ιστοσελίδες ή παραδοσιακά σε άρθρα επιστημονικών περιοδικών. Τονίζεται η ανάγκη δημιουργίας, πηγών εξειδικευμένων βάσεων δεδομένων, ειδικά όταν τα δεδομένα είναι δύσκολο ή αδύνατο να παρουσιαστούν στις παραδοσιακές πηγές.

Στις 11 - 12 Αυγούστου 2014 πραγματοποιήθηκε με την χρηματοδότηση του Wellcome Trust, στο Hinxton της Αγγλίας, μία συνάντηση είκοσι ενός κύριων ερευνητών που ο καθένας διατηρεί μια εξειδικευμένη πρωτεϊνική βάση δεδομένων ή διεξάγει έρευνα σχετικά με την διατήρηση ενός τέτοιου αποθετηρίου (Specialized Protein Resources Network). Το θέμα της συνάντησης ήταν η χάραξη πολιτικής για την δημιουργία και διατήρηση πρωτεϊνικών βάσεων δεδομένων και αποτελούνταν από πέντε ενότητες: (1) βασικές προκλήσεις, (2) εισαγωγή δεδομένων, (3) βέλτιστες πρακτικές για τη διατήρηση και την επιμέλεια, (4) ροή πληροφοριών προς και από τα μεγάλα κέντρα δεδομένων, και (5) επικοινωνία και χρηματοδότηση. Στο τέλος συνοψίζονται τα συνολικά συμπεράσματα που προέκυψαν από τη αυτήν την συνάντηση (Holliday et al., 2015).

Στην συνάντηση συμμετείχαν ερευνητές που διατηρούν «εξειδικευμένες» ηλεκτρονικές βάσεις δεδομένων συγκεκριμένων ειδών πρωτεϊνών (όπως αυτές ορίζονται από τα ενζυματικά, λειτουργικά ή δομικά χαρακτηριστικά τους) αλλά και διαχειριστές μεγάλων πρωτεϊνικών αποθετηρίων (συμπεριλαμβανομένων των Pfam, RefSeq, Swiss-Prot, και UniProt). Αυτά τα μεγάλα κέντρα δεδομένων χρησιμοποιούν διάφορες προσεγγίσεις για να συντηρήσουν το περιεχόμενο των δεδομένων τους, όπως η υπολογιστική ανάλυση, η συνεργασία, η ενοποίηση δεδομένων από πολλαπλές πηγές, και η επιμέλεια από ειδικούς σχολιαστές. Όλες οι βάσεις δεδομένων υποστηρίζονται από ειδικό σχολιασμό ώστε να εξασφαλιστεί η ακρίβεια και η πληρότητα των στοιχείων που παρουσιάζονται σε κάθε μικρή ή μεγάλη πρωτεϊνική βάση δεδομένων. Ένα κοινό πρόβλημα όλων των συμμετεχόντων της συνάντησης ήταν η επιμέλεια και η ανανέωση των βάσεων, δεδομένου ότι είναι δύσκολη η ανάκτηση πληροφοριών από δημοσιευμένα άρθρα επειδή συχνά δεν αναφέρουν αναλυτικές συγκεκριμένες πληροφορίες για τον υπό μελέτη οργανισμό (ειδικά για τα strains), ή τις ακριβείς πληροφορίες της αλληλουχίας που αναλύθηκε (πχ ο κωδικός πρόσβασης στη Uniprot ή το gi). Η διεύρυνση των συνεργασιών για τη διόρθωση λαθών στις βάσεις δεδομένων και η διάδοση της γνώσης αναγνωρίστηκαν από όλους ως βασικοί τρόποι δράσης, που θα ωφελήσουν όλες τις πρωτεϊνικές πηγές αλλά και τους χρήστες τους.

Η δημιουργία μίας βάσης δεδομένων θα μπορούσε να θεωρηθεί εύκολη διαδικασία όταν υπάρχουν κάποια στοιχεία διαθέσιμα, στην πραγματικότητα όμως υπάρχουν πολλές προκλήσεις και εμπόδια που πρέπει να αντιμετωπιστούν. Κάθε βάση δεδομένων έχει τις δικές της μοναδικές προκλήσεις και προβλήματα, αλλά κάποια από αυτά είναι κοινά σε όλες τις βάσεις και μπορούν να συνδυαστούν σε ένα βασικό ερώτημα: Τι κάνει μια βάση δεδομένων σημαντική;

Κύρια πρόκληση είναι η αξιοπιστία και όχι η ποσότητα των δεδομένων. Εξαρτάται εξολοκλήρου από το πεδίο εφαρμογής και τη λειτουργία της βάσης δεδομένων. Για παράδειγμα η βάση δεδομένων ESTHER, που εξετάζει μόνο τις εστεράσες και τα άλφα-βήτα ένζυμα υδρολάσης, και η GPCRDB που εξετάζει μόνο τα GPCRs, δεν πρόκειται ποτέ να έχουν τον ίδιο αριθμό καταχωρήσεων με την UniProtKB, η οποία περιλαμβάνει όλες τις αλληλουχίες αμινοξέων που έχουν βρεθεί μέχρι τώρα. Από μία ανάλυση που πραγματοποιήθηκε το 2009 (Schnoes, Brown, Dodevski, & Babbitt, 2009) προκύπτει ότι ορισμένες βάσεις δεδομένων έχουν ποσοστό σφαλμάτων/λάθος σχολιασμών (misannotation) περίπου 80%. Αντίθετα η Swiss-Prot που είναι το τμήμα της UniProtKB στο οποίο τα σχόλια καταχωρούνται χειροκίνητα από τους διαχειριστές, είχε ποσοστό σφάλματος περίπου 0%. Υπάρχουν πολλοί διαφορετικοί τύποι σφαλμάτων που μπορούν να βρεθούν στις πηγές δεδομένων. Κάποια είναι σχετικά εύκολο να εντοπιστούν μέσω αυτοματοποιημένων διαδικασιών, όπως για παράδειγμα τα ορθογραφικά λάθη στο σχολιασμό. Σφάλματα όμως που σχετίζονται με επιστημονικές πληροφορίες είναι πολύ πιο δύσκολο να βρεθούν, ειδικά αφού η γνώση εξελίσσεται πολύ γρήγορα. Χαρακτηριστικό παράδειγμα τέτοιου είδους σφάλματος αποτελεί ο ενζυματικός μηχανισμός δράσης της λυσοζύμης. Για πάνω από 50 χρόνια ο κοινά αποδεκτός μηχανισμός περιελάμβανε ένα ενδιάμεσο ζεύγος ιόντων. Νέα πειράματα όμως έδειξαν ότι περιλαμβάνει το σχηματισμό ενός ομοιοπολικού συμπλόκου γλυκοσυλενζύμου (Kirby, 2001). Οπότε τίθενται διάφορα ερωτήματα όπως για παράδειγμα: Ο αρχικός μηχανισμός ήταν πραγματικά λάθος; Μπορούμε ποτέ να ελπίζουμε ότι θα μπορέσουμε να προσδιορίσουμε τέτοιου είδους πληροφορίες; Έχουν καταχωρηθεί και προωθηθεί οι νέες πληροφορίες σε όλες τις βάσεις δεδομένων; Πιθανόν όχι, αλλά το κλειδί για την διατήρηση των βάσεων ενημερωμένων είναι οι διαχειριστές (και/ή χρήστες) να ανατρέχουν συχνά στη βιβλιογραφία ώστε να ενημερώνονται για ό,τι νέο υπάρχει, έχει αλλάξει ή θεωρείται απαρχαιωμένο.



Εικόνα 2.7: *Ο δεκάλογος της "καλής λειτουργίας" μιας βάσης δεδομένων, όπως παρουσιάστηκε από τον Καθ. Gert Vriend*

Υπάρχουν πολλοί ακόμα τύποι σφαλμάτων, για παράδειγμα ένα συχνό σφάλμα στην ανάλυση πρωτεϊνικών αλληλουχιών σχετίζεται με την σπονδυλωτή (modular) δομή πολλών πρωτεϊνών. Συνήθως συναντάται σε ενεργοποιημένα ένζυμα από υδατάνθρακες όπου ένα μέρος της σπονδυλωτής δομής (Module) που προσδένει υδατάνθρακες (carbohydrate-binding module, CBM) βρίσκεται συχνά προσαρτημένο σε καταλυτικές περιοχές που ανήκουν σε διάφορες οικογένειες ή ακόμη και σε δομές άγνωστης λειτουργίας. Μία καλή στοίχιση στο Blast, η οποία στοιχίζει μόνο το CBM, οδηγεί συχνά σε λανθασμένο σχολιασμό των παρακείμενων δομικών περιοχών. Το ίδιο μπορεί να λειτουργήσει και με αντίθετο τρόπο, όπως για παράδειγμα όταν μία πρωτεΐνη με μία μόνο περιοχή (single domain protein) αντιστοιχίζεται σε μια πρωτεΐνη πολλαπλής δομής και ο σχολιασμός μεταφέρεται από την δομή που δεν είναι αντιστοιχισμένη (π.χ. το ένζυμο που σχετίζεται με την αμινοτρανσφεράση (UniProtKB: B8NM72)). Αυτή η πρωτεΐνη, που εμπλέκεται στη βιοσύνθεση ενός δευτερογενούς μεταβολίτη τύπου-πεπτιδίου, στο παρελθόν θεωρούνταν ότι ήταν μια μη-ριβοσωμική πεπτιδική συνθετάση (NRPS), πιθανότατα λόγω της μεταφοράς του αυτόματου σχολιασμού από τα ομόλογά της που έχουν τη δομή και λειτουργία του NRPS. Ωστόσο, με προσεκτικό και χειρωνακτικό σχολιασμό των εμπλεκόμενων πρωτεϊνών (Umemura et al., 2014), διαπιστώθηκε ότι από την πρωτεΐνη αυτή έλειπε η περιοχή NRPS και ότι στην πραγματικότητα ήταν μια ριβοσωμική πρωτεΐνη. Αυτή είναι μια περίπτωση όπου ακόμη και ένα μικρό λάθος μπορεί να οδηγήσει πολλούς ερευνητές σε λανθασμένα συμπεράσματα. Επίσης γίνεται εμφανές, γιατί ο χειρωνακτικός σχολιασμός είναι απαραίτητος στις Εξειδικευμένες Πρωτεϊνικές Βάσεις Δεδομένων (Specialist Protein Resources - SPRs).

Ένας άλλος τύπος σφάλματος προκαλείται από την υπερεκτίμηση που συνάγεται από την "απόδειξη μέσα από την επανάληψη". Αυτό ενισχύεται περαιτέρω από το γεγονός ότι η λειτουργία μιας πρωτεΐνης μπορεί να οριστεί από το μοριακό/χημικό της ρόλο (π.χ. μία κινάση σερίνης) ή από την ευρεία βιολογική διαδικασία στην οποία μεσολαβεί (π.χ. η πρωτεΐνη η οποία μεσολαβεί στην πήξη του αίματος). Γενικά, είναι αρκετά δύσκολο να αποκρυπτογραφηθεί ο βιολογικός ρόλος της πρωτεΐνης στο ενδογενές πλαίσιο χρησιμοποιώντας υπολογιστικές μεθόδους και συνεπώς, τέτοιες προβλέψεις θα πρέπει να χρησιμοποιούνται με προσοχή. Η αναζήτηση με το BLAST στη non-redundant βάση δεδομένων πρωτεϊνών του NCBI, συχνά εντοπίζει ένα μεγάλο αριθμό παρόμοιων πρωτεϊνών που προέρχονται σχεδόν αποκλειστικά από γονιδιωματικές αλληλουχίες. Εξετάζοντας προσεχτικά τα ονόματά τους παρατηρείται ότι είναι ετερογενείς και ότι γίνεται μετάβαση από την μία στην άλλη χωρίς επίβλεψη. Επιπλέον πολλά ομόλογα ένζυμων στερούνται των καθοριστικών αμινοξέων καταλοίπων στο ενεργό κέντρο, καθιστώντας τα μη ενεργά. Από την άλλη πλευρά, σφάλματα συναρμολόγησης γονίδιων (gene assembly errors) προκαλούν υποθετικές

πρωτεΐνες στις οποίες έχει ταυτοποιηθεί λάθος εναρκτήρια μεθιονίνη, ή με εξώνια που έχουν παραλειφθεί, έχοντας σαν πιθανό αποτέλεσμα την παράλειψη ενεργών κατάλοιπων. Παρά το γεγονός ότι αυτά τα λάθη μπορούν στη συνέχεια να διορθωθούν, ο έλεγχος και η διόρθωση των σχολιασμών τους αποτελεί πρόκληση για τους διαχειριστές των SPRs.

Πως μπορούν να διορθωθούν σφάλματα που έχουν εντοπιστεί στις βάσεις δεδομένων; Πολλές πηγές, όπως η UniProtKB, διαθέτουν μηχανισμούς ώστε οι χρήστες να αναφέρουν πιθανά προβλήματα. Άλλες βάσεις, όπως η PDB, δεν επιτρέπουν την διόρθωση των δεδομένων (αν και η PDB_REDO (Joosten, Long, Murshudov, & Perrakis, 2014) επιτρέπει την διόρθωση των ατομικών συντεταγμένων). Για τα άλλα είδη σφαλμάτων σχολιασμού και ιδιαίτερα εκείνων που σχετίζονται με την αλληλουχία των αμινοξέων, έχουν προταθεί μεθοδολογίες για τον εντοπισμό και την διόρθωσή τους (Nagy et al., 2008; Wong, Maurer-Stroh, & Eisenhaber, 2010).

Όταν το σφάλμα διορθωθεί, πως μπορούμε να ενημερώσουμε όλες τις βάσεις δεδομένων που χρησιμοποιούν την αρχική εγγραφή; Η προέλευση των δεδομένων είναι συχνά δύσκολο να εντοπιστεί. Οι διαχειριστές των βάσεων δεδομένων έλαβαν την πληροφορία από την UniProtKB, ή από την πρωτογενή βιβλιογραφία; Ίσως να την πήραν από το SFLD, οπότε τίθεται το ερώτημα: οι διαχειριστές του SFLD από πού την πήραν; Ορισμένες πηγές (π.χ. UniProtKB) έχουν αρχίσει να χρησιμοποιούν το ECO (Evidence Code Ontology) (Chibucos et al., 2014), στο οποίο περιλαμβάνονται και οι πηγές. Η συμπλήρωση όμως πηγών με σχολιασμό τέτοιου είδους είναι συχνά περίπλοκη διαδικασία, καθώς όλα τα δεδομένα πρέπει να διασταυρώνονται και να ελέγχονται. Ένας από τους μελλοντικούς στόχους είναι η δημιουργία κανόνων καταχώρησης σχολιασμού βάσεων όπου θα είναι σημαντική η δυνατότητα γνώσης της πηγής τους με αποτέλεσμα η χρήση του ECO ή κάποιου παρόμοιου κώδικα να είναι απαραίτητη.

Με τον συνεχώς αυξανόμενο όγκο διαθέσιμων δεδομένων, πώς θα μπορούσε να διατηρηθεί ή ακόμα και να ενισχυθεί η αξιοπιστία των πηγών των βάσεων; Ο ειδικός διαχειριστής της βάσης (άτομα που είναι εκπαιδευμένα σε έναν συγκεκριμένο τομέα) παίζει πάντα καθοριστικό ρόλο. Οι βάσεις τις οποιες τις διαχειρίζονται άνθρωποι και πραγματοποιούν διασταύρωση στοιχείων, έχουν μεγαλύτερη αξιοπιστία σε σχέση με τις βάσεις που τα δεδομένα απλά καταχωρούνται αυτόματα και συνήθως αναπαράγουν σφάλματα. Ωστόσο, ο ρόλος των χρηστών θα είναι πολύ σημαντικός στο μέλλον. Για παράδειγμα, οι χρήστες όταν εντοπίσουν κάποιο σφάλμα θα μπορούν να επικοινωνούν με τις βάση δεδομένων (δίνοντας αποδείξεις για την ύπαρξη του σφάλματος) ώστε να διορθώνεται η καταχώρηση. Υπάρχει περίπτωση βέβαια ο κατάλογος των σφαλμάτων να ξεπεράσει πολύ γρήγορα την ικανότητα της βάσης δεδομένων να τα διορθώσει. Επιπλέον, οι χρήστες θα μπορούν να προτείνουν νέες καταχωρήσεις ή ακόμα και να καταχωρούν δεδομένα. Το μεγαλύτερο εμπόδιο σε αυτή τη μέθοδο σχολιασμού είναι η εκπαίδευση των χρηστών για τον εντοπισμό των σφαλμάτων.

Ένας τρόπος (όπως εφαρμόστηκε στην διεθνή κοινότητα κρυσταλλογραφίας) είναι η εισαγωγή δεδομένων να αποτελεί προαπαιτούμενο της δημοσίευσης των αποτελεσμάτων. Ωστόσο, χωρίς την υποστήριξη και την επιβολή αυτής της απόφασης από τα περιοδικά, είναι αδύνατη η απόκτηση επαρκούς λειτουργικής πληροφορίας. Τα δεδομένα που καταχωρούνται σε πολλές περιπτώσεις δεν χρειάζεται να είναι ιδιαίτερα λεπτομερειακά. Για παράδειγμα, σημαντική πρόοδος θα μπορούσε να είναι, μαζί με τον αριθμό πρόσβασης αλληλουχιών να συμπεριλαμβάνεται και ο αριθμός Enzyme Commission (EC).

Ένας άλλος τρόπος, θα μπορούσε είναι να μέσω της Βικιπαίδεια (WikiPedia). Αυτή η μέθοδος χρησιμοποιείται ήδη από τη βάση Rfam. Οι συγγραφείς θα μπορούσαν να δημιουργούν μια σελίδα της Wikipedia, η οποία θα χρησιμοποιείται για να συμπληρωθεί η βάση δεδομένων Rfam. Ωστόσο, ποιες πηγές θα είναι οι πρωτογενείς συλλέκτες δεδομένων; Θα είναι η Swiss-Prot η μοναδική πηγή για όλους τους σχολιασμούς των αμινοξέων που θα έχουν ως αναφορά οι άλλες πηγές; Θα συμφωνήσουν όλα τα περιοδικά στην προτεινόμενη διαδικασία; Η διαδικασία σχολιασμού θα είναι αρκετά απλή και ολοκληρωμένη ώστε οι συγγραφείς να την ακολουθήσουν; Δυστυχώς δεν υπάρχει απλή απάντηση σε αυτά τα ερωτήματα, αλλά καθώς αυξάνεται ο όγκος των δεδομένων, οι δημιουργοί των SPRs, οι χρήστες και οι εκδότες θα πρέπει να τα αντιμετωπίσουν.

Για να είναι χρήσιμη μία πηγή, η γλώσσα που χρησιμοποιούν και οι δύο θα πρέπει να είναι τυποποιημένη. Παράδειγμα ενός τέτοιου εγχειρήματος αποτελεί το ερευνητικό πρόγραμμα EMBRACE (Pettifer et al., 2010). Αυτό που εννοεί μία βάση με τον όρο superfamily μπορεί να μην σημαίνει το ίδιο σε μία άλλη βάση. Για παράδειγμα, ο ορισμός SFLD απαιτεί οι πρωτεΐνες να είναι όχι μόνο εξελικτικά σχετικές, αλλά και να έχουν μια συντηρημένη χημεία. Η TIGRFAM, από την άλλη πλευρά, απαιτεί απλώς να υπάρχει εξελικτική συγγένεια. Η SFLD έχει μια ιεραρχία, η οποία αντιστοιχίζεται στο PANTHER, αλλά οι όροι που χρησιμοποιούνται είναι διαφορετικοί (μια υποομάδα SFLD είναι το ισοδύναμο μιας οικογένειας PANTHER).

Επίσης, υπάρχει το θέμα της χημείας. Η πιο κοινή μέθοδος ταξινόμησης της χημείας του ενζύμου είναι ο αριθμός Enzyme Commission (EC). Δημιουργήθηκε για να αποφευχθεί η πληθώρα εσωτερικών ονομασιών (όπως ινβερτάση, σουμπτιλισίνη, κ.λ.π.) και να συνδέσει τα ονόματα με τα μόρια (συνήθως τα υποστρώματα και τους συνολικούς χημικούς μετασχηματισμούς που συμβαίνουν, αλλά όχι την αλληλουχία και την δομή). Ακόμη και μεταξύ των βάσεων MACiE και EzCatDB, οι οποίες ταξινομούν αντιδράσεις ενζύμων, είναι πιθανό να χρειάζεται τυποποίηση της γλώσσας ή του λεξιλογίου, δεδομένου ότι διαχείρισή τους γίνεται με διαφορετικούς τρόπους. Η βάση MACiE κατατάσσει τα στάδια της αντίδρασης, ενώ η EzCatDB ταξινομεί ολόκληρες τις αντιδράσεις που αποτελούνται από ένα ή περισσότερα στάδια. Για αρκετά χρόνια γινόταν μία προσπάθεια βιοχημικού χαρακτηρισμού των ενζύμων, προσδιορίζοντας τον αντίστοιχο αριθμό EC (που χαρακτηρίζει τη γενική χημική αντίδραση που καταλύει το ένζυμο) για το κάθε ένα.

Σημαντικό πρόβλημα αποτελεί το γεγονός ότι ο αριθμός EC ορίστηκε την δεκαετία του ’50. Από τότε έχει βρεθεί ότι πολλά από τα ένζυμα που έχουν αριθμό EC είναι μη ειδικά. Παρόλα αυτά, ο αριθμός EC χρησιμοποιείται μέχρι σήμερα ενώ δημιουργούνται ακόμα και νέοι αριθμοί EC, με αποτέλεσμα να μην μπορούμε να αγνοήσουμε την ύπαρξη του ως ένα χρήσιμο εργαλείο. Ωστόσο, με την αύξηση της δυσκολίας της δημοσίευσης των χαρακτηρισμών των ενζύμων σε περιοδικά υψηλής απήχησης, η σωστή μέθοδος συσχετισμού των πρωτεϊνών με τον αντίστοιχο αριθμό EC έχει εξαλειφθεί, και σήμερα χρησιμοποιείται σχεδόν αποκλειστικά στο πεδίο της βιοπληροφορικής. Επιπλέον, ο αριθμός EC είχε ως στόχο να χαρακτηρίσει την χημική αντίδραση που εκτελείται από ένα ένζυμο, και όχι να αποδίδεται ανάλογα με την ομοιότητα των αλληλουχιών καθώς η ίδια αντίδραση μπορεί να καταλύεται από πολλές μη σχετιζόμενες οικογένειες αλληλουχιών (π.χ. οι β-λακταμάσες) και πολλά ένζυμα ομαδοποιημένα στην ίδια οικογένεια καταλυτικών αντιδράσεων που περιγράφονται από διαφορετικούς αριθμούς EC (π.χ. οι ενδονουκλεάσες). Συμπεραίνουμε επομένως ότι η απόδοση του αριθμού EC είναι πολύ πιο περίπλοκη διαδικασία σε σχέση με την απλή απόδοση του EC βάσει της καλύτερης στοίχισης στο BLAST. Τέτοιου είδους ζητήματα μεταφοράς σχολιασμού αποτελούν πρόκληση όχι μόνο για τους χρήστες των βάσεων δεδομένων, αλλά και για εμάς. Πώς ξέρουμε πότε θα πρέπει να διαδοθούν και πότε όχι οι λειτουργικές πληροφορίες; Σε μερικές περιπτώσεις, όπως η βάση δεδομένων CAZy, προτιμάται να μην διαδίδονται οι λειτουργικές πληροφορίες και απλά αναφέρουν τις λειτουργίες που έχουν προσδιοριστεί πειραματικά. Άλλες βάσεις, όπως η MACiE και η EzCatDB, απλώς αναφέρουν τους αριθμούς EC των ομολόγων, αλλά περιλαμβάνουν την ταυτότητα των συντηρημένων κατάλοιπων, έτσι ώστε οι χρήστες να μπορούν να βγάλουν τα δικά τους συμπεράσματα ως προς την εγκυρότητα των προβλέψεων.

Παρόλο που δεν είναι αποδεκτή η άποψη ότι όλες οι πηγές θα πρέπει να χρησιμοποιούν την ίδια γλώσσα (η βιολογία είναι πολύπλοκη, οπότε ένας όρος σε ένα πεδίο δεν μπορεί να μεταφραστεί με ακρίβεια σε κάποιο άλλο πεδίο), πιθανότατα θα ήταν χρήσιμο να βρεθεί ένας τρόπος να μεταφράζονται οι έννοιες. Οι οντολογίες είναι ίσως ο πιο κατάλληλος τρόπος. Παρά το γεγονός ότι η οντολογία γονιδίων (GO) είναι ίσως η πιο ευρέως γνωστή οντολογία στον τομέα της βιοπληροφορικής, πρέπει να γνωρίζουμε ότι δεν είναι η μοναδική. Μια αναζήτηση στο PubMed με τον όρο "οντολογία" στον τίτλο των εγγράφων αποδίδει περίπου 1.500 αποτελέσματα. Αν και μπορεί να μην υπάρχουν οντολογίες για κάθε πεδίο της βιοχημείας και της βιολογίας, υπάρχουν σε πολλά, και για μερικές από τις βασικές έννοιες (π.χ. ένα ένζυμο) υπάρχει η δυνατότητα σύνδεσης των δεδομένων σε όλες τις πηγές που έχουν παρόμοια στοιχεία. Είναι θεμιτό να γνωρίζουμε την οντολογία ή το λεξιλόγιο που χρησιμοποιείται, αλλά θα ήταν πιο χρήσιμο για όλους, τους χρήστες και τους διαχειριστές, και τις βάσεις οντολογίας, όπως η BioPortal (Grosjean, Soualmia, Bouarech, Jonquet, & Darmoni, 2014) και η OBO Foundry (Smith et al., 2007), η συλλογή όσο το δυνατόν περισσότερων πληροφοριών σε μια ενιαία βάση.

Εικόνα 2.8: *Φωτογραφία των συμμετεχόντων στο Protein Bioinformatics and Community Resources Retreat. Το όνομα κάθε επιστήμονα ακολουθείται από το όνομα της εξειδικευμένης βάσης δεδομένων την οποία διευθύνει.. Πίσω σειρά: David Landsman (Histone database), Dan Haft (TIGRFAMS), Bernard Henrissat (CAZy), Rob Finn (InterPro and Pfam), David Craik (ConoServer and CyBASE), Arnaud Chatonnet (ESTHER), Neil Rawlings (MEROPS); Μεσαία σειρά: Amos Bairoch (neXtProt), Gerard Manning (Kinase.com), Michael Spedding (IUPHAR), Gert Vriend (GPCRDB), Milton Saier (TCDB), Pantelis Bagos (OMPdb); Εμπρός σειρά: Narayanaswamy Srinivasan (KinG), Ramanathan Sowdhamini (PASS2), Alex Bateman (Pfam & UniProt), Patsy Babbitt (SFLD), Kim Pruitt (RefSeq), Claire O'Donovan (UniProt), Gemma Holliday (MACiE), Nozomi Nagano (EzCatDB).*

Η βάση δεδομένων του περιοδικού Nucleic Acids Research (Fernández-Suárez, Rigden, & Galperin, 2014) περιείχε το 2014 συνολικά 1.552 βάσεις δεδομένων από τις οποίες οι 58 ήταν νέες και οι 123 παλιότερες που ανανεώθηκαν. Η δημιουργία μίας βάσης δεδομένων είναι εύκολη διαδικασία. Για παράδειγμα πολλές βάσεις δεδομένων έχουν δημιουργηθεί στο πλαίσιο διδακτορικών διατριβών ή μεταπτυχιακών προγραμμάτων. Η δυσκολία έγκειται στην διατήρησή της. Μία μελέτη που πραγματοποιήθηκε το 2008 έδειξε ότι περίπου το 40% των διευθύνσεων URL των βάσεων δεδομένων που ήταν δημοσιευμένες σε επιστημονικά περιοδικά πλέον δεν ήταν διαθέσιμες (Wren, 2008). Ωστόσο, η διατήρηση της βάσης δεν αφορά μόνο την ύπαρξη μιας σταθερής διεύθυνσης URL. Το πρώτο πράγμα που χρειάζεται κάθε βάση δεδομένων είναι το προσωπικό που θα την διατηρεί και θα συνεχίσει να την αναπτύσσει. Μερικές από αυτές συντηρούνται από ειδικούς επιστήμονες που εργάζονται μόνοι τους ή/και στον ελεύθερο χρόνο τους, αλλά αυτό είναι δύσκολο να λειτουργήσει μακροπρόθεσμα. Τι συμβαίνει όταν ο επιστήμονας που διατηρεί την βάση πρέπει να προχωρήσει και δεν υπάρχει κανείς να τον αντικταστήσει; Ίσως μια λύση στο πρόβλημα αυτό είναι η ενοποίηση των βάσεων. Παράδειγμα αποτελεί η Interpro, μία πηγή που ενσωματώνει πολλές διαφορετικές πηγές. Οι βάσεις δεδομένων που την αποτελούν εξακολουθούν να διατηρούν τη δική τους ταυτότητα και να έχουν τον δικό τους ρόλο χωρίς να μπορεί να διατηρήσει τις βάσεις η Interpro. Η εξασφάλιση χρηματοδότησης είναι μια συνεχής πρόκληση για μικρές ή/και ανεξάρτητες (από μεγαλύτερες πηγές, όπως η REFSEQ ή UniProtKB) SPRs. Υπάρχουν αρκετοί τρόποι αύξησης των πόρων που διατίθενται για τις SPRs, όπως για παράδειγμα, οι επιχορηγήσεις οργανισμών, (π.χ. η SFLD αυτή τη στιγμή υποστηρίζεται από μια επιχορήγηση του NIH), χρηματοδότηση χρηστών (εμπορική), (π.χ. η KEGG (Kanehisa et al., 2014) στην οποία επιτρέπεται η πρόσβαση σε συνδρομητές), δηλαδή χρηματοδοτείται από τους χρήστες, ενώ έχουν προταθεί και άλλα πιο σύνθετα μοντέλα.

Η συνεχής επικαιροποίηση των δεδομένων των βάσεων αποτελεί ίσως την μεγαλύτερη πρόκληση που έχουμε να αντιμετωπίσουμε σήμερα. Ο όγκος των δεδομένων που διατίθενται είναι τεράστιος ενώ η αύξηση των διαθέσιμων δεδομένων είναι εκθετική. Η UniProtKB τον Νοέμβριο του 2014 είχε πάνω από 86 εκατομμύρια εγγραφές, εκ των οποίων σχολιάστηκαν χειροκίνητα ή αναθεωρήθηκαν περίπου μισό εκατομμύριο. Για κάθε πληροφορία που γνωρίζουμε για μία μόνο πρωτεΐνη, υπάρχουν ακόμη περισσότερες πρωτεΐνες για τις οποίες δεν έχουμε κανένα στοιχείο, εκτός από την πρωταρχική αλληλουχία αμινοξέων. Ο αυτόματος σχολιασμός και οι υποθέσεισείναι κρίσιμης σημασίας για να συνεχιστεί η καταχώριση της πληθώρας των πρωτογενών αλληλουχιών. Διότι, ακόμη και με αναλύσεις υψηλής απόδοσης, όπως αυτές που παρέχονται από το Structural Genomics Consortium, οι πειραματιστές δεν μπορούν να προχωρήσουν σε βιοχημικές ή ακόμα και υπολογιστικές προβλέψεις ενώ το πρωτεϊνικό δίπλωμα δεν παρέχει σχεδόν ποτέ ακριβείς πληροφορίες για την πρωτεϊνική λειτουργία ώστε να είναι απευθείας χρήσιμες σε έναν βιολόγο. Για παράδειγμα, μια πρωτεΐνη της οποίας το όνομα έχει δοθεί από την τάξη διπλώματος, όπως η «putative glycoside hydrolase» ή ένα ομόλογο πεπτιδάσης, δεν παρέχει στον χρήστη καμιά είδους ειδικότητα και άρα επαρκείς πληροφορίες ώστε να γνωρίζει ακριβώς τη λειτουργία της. Για παράδειγμα, οι κυτταρινάσες (τα ένζυμα που διασπούν την κυτταρίνη των δέντρων) και η νευραμινιδάση (που επιτρέπει στον ιό της γρίπης να ολοκληρώσει επιτυχώς τον κύκλο μόλυνσής του) είναι και οι δύο γλυκοσιδάσες (glycoside hydrolases). Είναι επίσης σημαντικό να γίνει διάκριση μεταξύ της βιοχημικής λειτουργίας (όπως η χημική διάσπαση της κυτταρίνης) και του βιολογικού ρόλου (όπως παροχή βοήθειας σε ένα ιό για να ολοκληρώσει τον κύκλο μόλυνσης του) καθώς οι βιοχημικές λειτουργίες σχετίζονται περισσότερο με πρωτεϊνικές αλληλουχίες και δομές (A. C. Martin et al., 1998) από ότι με τον βιολογικό τους ρόλο. Για παράδειγμα οι πρωτεΐνες που έχουν πάνω από μία λειτουργία (moonlighting proteins) έχουν απολύτως όμοιες αλληλουχίες και δομές, αλλά έχουν διαφορετικούς ρόλους, συχνά ανάλογα με την κυτταρική τους θέση. Ως εκ τούτου, είναι πιο δύσκολο να προσδιοριστεί η λειτουργία μιας πρωτεΐνης από ότι η τρισδιάστατη δομή της και η μεταφορά του σχολιασμού μπορεί τουλάχιστον να βοηθήσει τους χρήστες δίνοντας ένα αρχικό στοιχείο για την τεκμαιρόμενη λειτουργία της. Ωστόσο, οι χρήστες, οι διαχειριστές και οι δημιουργοί των βάσεων, πρέπει όλοι να γνωρίζουν τις διαφορές στον καθορισμό αυτών των λειτουργικών επίπεδων, όπως η σημασιολογική ακρίβεια θα βοηθήσει τους χρήστες να βρουν τις πληροφορίες που θέλουν, αλλά η αυτόματη μεταφορά του σχολιασμού εξακολουθεί να απαιτεί όχι μόνο ένα καλό μοντέλο για την πραγματοποίησή του, αλλά και υψηλής ποιότητας και όσο το δυνατόν πληρέστερα δεδομένα. Επομένως, τα μέλη της κοινότητας SPR πρέπει να εργάζονται από κοινού για να ελαχιστοποιηθεί η επικάλυψη των προσπαθειών, έχοντας ως κοινό στόχο την διατήρηση της ποιότητας αλλά και της ποσότητας των δεδομένων ώστε οι χρήστες των βάσεων να έχουν τα καλύτερα δυνατά δεδομένα. Πολύτιμη είναι επίσης και η βοήθεια των χρηστών, χωρίς τους οποίους καμία πηγή δεν μπορεί να αναπτυχθεί και να ευδοκιμήσει.

Παρακάτω, δίνεται μια σύντομη περιγραφή των εξειδικευμένων βάσεων των οποίων οι επιστημονικοί υπεύθυνοι και εκπρόσωποι συμμετείχαν στη συνάντηση του Protein Bioinformatics and Community Resources Retreat (Εικόνα 2.8).

TCDB: Πολλά από τα αποθετήρια πληροφοριών που συνήθως θεωρούνται ιστοσελίδες, στην πραγματικότητα αποτελούν σχεσιακές βάσεις δεδομένων (Stein, 2013) και επιτρέπουν την διάθεση των δεδομένων και την οργάνωση της γνώσης, ταυτόχρονα βάσει πολλαπλών κριτήριων. Οι αλληλουχίες είναι πιθανό να έχουν πολλά ονόματα και να ανήκουν σε πολλές ομάδες, καθεμία από τις οποίες αποθηκεύεται ιεραρχικά. Πλεονεκτήματα των σχεσιακών βάσεων δεδομένων (structured query language - SQL) αποτελούν η οργάνωση και η αναζήτηση των δεδομένων με πολλούς διαφορετικούς τρόπους καθώς και το γεγονός ότι συνδέονται με άλλα συστήματα (Jamison, 2003). Σε αυτή την ενότητα, θα συζητήσουμε σχετικά με το σύστημα διαχείρισης της βάση δεδομένων Transporter Classification Database (TCDB, www.tcdb.org (M. H. Saier, Reddy, Tamang, & Västermark, 2014)). Η TCDB ξεκίνησε ως μια απλή ιστοσελίδα σε HTML. Το 1998 μετατράπηκε σε μια σχεσιακή (Oracle MySQL) βάση δεδομένων με διεπαφή PHP και σήμερα στεγάζεται στο San Diego Supercomputer Center (www.sdsc.edu). Πρόκειται για μια βάση δεδομένων των πρωτεϊνικών συστημάτων μεταφοράς από όλους τους ζωντανούς οργανισμούς που κατατάσσονται σύμφωνα με την κατηγορία, υποκατηγορία, οικογένεια, υπο-οικογένεια και το σύστημα. Τα συστήματα μεταφοράς μπορεί να αποτελούνται από απλές ή σύνθετες πρωτεΐνες (multi-component), με μέγιστο έως περίπου 100 πρωτεΐνες ανά σύστημα. Η TCDB χρησιμοποιείται ως κοινό σημείο αναφοράς για τον χαρακτηρισμό άγνωστων συστημάτων. Αυτή τη στιγμή περιλαμβάνει 7 κατηγορίες, 56 υπεροικογένειες, 937 οικογένειες, 9098 συστήματα, 11806 πρωτεΐνες και 12086 βιβλιογραφικές αναφορές. Τα συστήματα έχουν καταχωρηθεί με την ημερομηνία δημοσίευσης, ενώ όλες οι καταχωρήσεις επιμελούνται και σχολιάζονται από ειδικούς. Το

σύστημα TC σχεδιάστηκε με βάση το EC (Enzyme Commission) και είναι παρόμοιο με αυτό (Bairoch, 1999), με την διαφορά ότι βασίζεται τόσο στη λειτουργία (κατηγορία και υποκατηγορία) όσο και τη φυλογένεση (οικογένεια, υπο-οικογένεια και υπερ-οικογένεια). Είναι το μόνο σύστημα εγκεκριμένο από την Διεθνή Ένωση Βιοχημείας και Μοριακής Βιολογίας (International Union of Biochemistry and Molecular Biology, IUBMB) που χρησιμοποιείται σήμερα για τα διαμεμβρανικά μοριακά συστηματα μεταφοράς (M. H. Saier, Jr., 2000). Το σύστημα διαχείρισης της TCDB έχει πολλά πλεονεκτήματα: Η γνώση ειναι ιεραρχικά δομημένη, υπάρχει πρόβλεψη για διαχείριση αρχείων ασφαλείας και ανάκτηση αυτών, ενώ έχουν αναπτυχθεί και ειδικές εφαρμογές βιοπληροφορικής πάνω στο σύστημα ταξινόμησης, όπως το TC-BLAST και λογισμικό για την ανίχνευση μακρινών φυλογενετικών σχέσεων, βάσει της Υπεροικογένειας. Τέλος, γίνονται προσπάθειες για ενοποίηση της πληροφορίας με άλλες βάσεις όπως η PFAM και η OMPdb.

OMPdb: Η βάση δεδομένων OMPdb (Tsirigos, Bagos, & Hamodrakas, 2011), διατίθεται στην ιστοσελίδα http://www.ompdb.org, είναι διαθέσιμη στο κοινό και περιέχει διαμεμβρανικά β-βαρελια της εξωτερικής μεμβράνης των κατά Gram αρνητικών βακτηρίων. Παρουσιάστηκε για πρώτη φορά το 2011 και περιείχε περίπου 70.000 εγγραφές. Μέσα στα επόμενα 3 χρόνια, περιελάμβανε περισσότερες από 500.000 καταχωρήσεις. Όλες οι πρωτεΐνες της OMPdb ταξινομούνται σε 91 οικογένειες, βάσει δομικών και λειτουργικών κριτηρίων. Κάθε οικογένεια χαρακτηρίζεται από διαφορετικό προφίλ Hidden Markov Model (pHMM), που την διαχωρίζει από τις υπόλοιπες. Οι περισσότερες από αυτές τις οικογένειες είχαν ήδη αναφερθεί στη βάση Pfam (Finn, et al., 2014), εκτενής όμως βιβλιογραφική έρευνα επέτρεψε την αναγνώριση όχι μόνο οικογενειών που δεν υπήρχαν στην αντίστοιχη Pfam clan (MBB clan - CL0193), αλλά και κάποιων που αναγνωρίζονταν ως περιοχές άγνωστης λειτουργίας (Domains of Unknown function - DUFs). Επιπλέον, συνολικά 15 οικογένειες, έλειπαν από την Pfam ή είχαν χαρακτηριστεί με αυτόματο τρόπο στην Pfam-B. Για κάθε πρωτεΐνη, ο χρήστης μπορεί να ανακτήσει πληροφορίες σχετικά με την παρουσία των σηματοδοτικών αλληλουχιών και τον σχολιασμό των διαμεμβρανικών τμημάτων. Για κάθε εγγραφή οικογένειας, και εφόσον είναι διαθέσιμη, παρατίθεται λίστα πρωτεϊνών που έχουν κρυσταλλογραφικά προσδιορισμένη δομή. Ο χαρακτηρισμός των πρωτεϊνών βάσει του προφίλ pHMM και η υιοθέτηση του συστήματος ταξινόμησης της Pfam, επιτρέπει στους επιμελητές να ακολουθήσουν ένα ημι-αυτόματο σύστημα ανάκτησης δεδομένων. Αρχικά, μια οικογένεια αναγνωρίζεται μέσω της βιβλιογραφικής αναζήτησης, στη συνέχεια δημιουργούνται μοντέλα pHMM και συγκρίνονται έναντι της βάσης Pfam, και τέλος, προσδιορίζονται τα μέλη της οικογένειας και αποθηκεύονται στη βάση δεδομένων. Το σύστημα αυτό παρουσιάζει έναν πιο πλήρη και ακριβή σχολιασμό των πρωτεϊνών δομής β-βαρελιού, λόγω της πρόσθετης αξίας του χειρωνακτικού σχολιασμού και των λεπτομερών βιβλιογραφικών αναφορών. Από την άλλη πλευρά, η σύγκριση της OMPdb με τις άλλες εξειδικευμένες βάσεων δεδομένων που περιέχουν πρωτεΐνες δομής β-βαρελιού, αποκαλύπτει ότι υπερέχει από όλες τις πλευρές, διότι διαθέτει το μεγαλύτερο αριθμό εγγραφών, πρωτεϊνών και οικογενειών. Διαθέτει τα πιο πλήρη και αποκλειστικά δεδομένα τα διαμεμβρανικά β-βαρελια, και προσφέρει την πιο ολοκληρωμένη διασύνδεση με άλλες δημόσιες βάσεις δεδομένων, βιβλιογραφικές αναφορές, εργαλεία πρόβλεψης και σχολιασμού αλληλουχιών. Η OMPdb συνεργάζεται με τους επιμελητές των βάσεων TCDB και Pfam (και οι δύο βάσεις δεδομένων περιέχουν τις οικογένειες των πρωτεϊνών δομής β-βαρελιού εξωτερικής μεμβράνης των κατά Gram αρνητικών βακτηρίων), προκειμένου να επιτύχει τη ενοποίηση των βάσεων δεδομένων με τη διασύνδεση των οικογενειών και τη διατήρηση των πληροφοριών ενημερωμένων (ανταλλαγή σχολιασμού, αναφορές κ.τ.λ.). Η διαδικτυακή εφαρμογή βασίζεται στο συνδυασμό δύο επιπέδων. Το βασικό επίπεδο είναι ένα σύστημα βάσης δεδομένων MySQL, και το δεύτερο επίπεδο είναι ένας διακομιστής εφαρμογών Apache-PHP που λαμβάνει τις αναζητήσεις των χρηστών. Παρόλο που η ιεραρχία της βάσης δεδομένων είναι μάλλον απλή (δηλαδή υπάρχει μόνο ένα επίπεδο, η οικογένεια), η βάση αποθηκεύεται σε MySQL, ώστε να διευκολυνθεί η διαδικασία των εξειδικευμένων ερωτημάτων και να γίνεται πιο εύκολη η ενημέρωση της βάσης. Η διεπαφή ιστού της OMPdb προσφέρει στο χρήστη τη δυνατότητα όχι μόνο να δει τα διαθέσιμα δεδομένα, αλλά και να υποβάλει εξειδικευμένες αναζητήσεις για την αναζήτηση ανάμεσα στις εγγραφές των πρωτεϊνών της βάσης. Η ύπαρξη ενός τόσο μεγάλου και αξιόπιστου συνόλου δεδομένων διαμεμβρανικών β-βαρελιών μπορούν να χρησιμοποιηθούν για αναλύσεις μεγάλης κλίμακας σχετικά με την ακρίβεια ταξινόμησης των υφιστάμενων προγνωστικών αλγορίθμων, για την δημιουργία νέων μεθόδων πρόβλεψης και για μελέτες μοντελοποίησης. Μακροπρόθεσμο στόχο αποτελεί η διατήρηση της OMPdb όσο το δυνατόν πιο ενημερωμένη, ακολουθώντας τις τακτικές ενημερώσεις της UniProt και κάνοντας ανασκόπηση της βιβλιογραφίας για νέες πειραματικά επαληθευμένες πρωτεΐνες δομής β-βαρελιού, προκειμένου να συμπεριληφθούν στη βάση ή να ενταχθούν σε νέες οικογένειες. Παρόμοια με άλλες βάσεις δεδομένων, η OMPdb βρίσκεται υπό εξέλιξη, και η αλληλεπίδρασή της με την κοινότητα των

χρηστών είναι ζωτικής σημασίας για την ανάπτυξη και την τελειοποίηση της. Εκτός από τη συνεργασία με τις υπόλοιπες σχετικές βάσεις δεδομένων που αναφέρθηκαν παραπάνω (Pfam και TCDB), οι διαχειριστές ενθαρρύνουν τους χρήστες να υποβάλουν στοιχεία, να διορθώσουν πιθανά λάθη, και να διατυπώσουν προτάσεις ώστε η OMPdb να αποκτήσει μεγαλύτερη χρησιμότητα για την επιστημονική κοινότητα.

CAZy: Η βάση δεδομένων CAZy (www.cazy.org) περιγράφει οικογένειες παρόμοιας καταλυτικής δομής και περιοχής πρόσδεσης υδατανθράκων (carbohydrate-binding modules) των ενζύμων, που διασπούν, τροποποιούν, ή δημιουργούν γλυκοσιδικούς δεσμούς (Cantarel et al., 2009; Lombard, Ramulu, Drula, Coutinho, & Henrissat, 2014). Η βάση δεδομένων CAZy δημοσιεύθηκε το 1991, πριν από οποιαδήποτε αλληλούχιση γονιδιώματος (Henrissat, 1991) και περιλαμβάνει την ταξινόμηση της οικογένειας αλληλουχιών των γλυκοζιδικών υδρολασών. Ξεκίνησε στις αρχές της δεκαετίας του '90 και επεκτάθηκε και σε άλλες κατηγορίες ενζύμων ενεργών υδατανθράκων όπως οι γλυκοζυλοτρανσφεράσες (Campbell, Davies, Bulone, & Henrissat, 1997). Έγινε διαθέσιμη αρχικά μέσω μιας απλής ιστοσελίδας τον Σεπτέμβριο του 1998, ενώ μετατράπηκε σε ολοκληρωμένη βάση δεδομένων τύπου SQL (το 1999), ώστε να είναι πιο εύκολη η διαχείριση τους και να βελτιωθεί ο ρυθμός συλλογής τους. Παρά την ταχεία αύξηση των δεδομένων, κάθε αλληλουχία που εμφανίζεται στην CAZy συνεχίζει να ελέγχεται από κάποιον επιμελητή, εκτός εάν η νέα αλληλουχία είναι σε συμφωνία, χωρίς κανένα κενό και με περισσότερο από 50% ταύτιση με μια ήδη ταξινομημένη αλληλουχία. Ο ανθρώπινος παράγοντας στην επιμέλεια της βάσης, που περιλαμβάνει διορθώσεις σφαλμάτων μετά από αίτημα κάποιου χρήστη, καθώς και η απόδοση αριθμών EC αποκλειστικά σε ένζυμα που έχουν χαρακτηριστεί πειραματικά χωρίς μεταφορά σχολιασμού λόγω ομοιότητας αλληλουχίας, καθιέρωσε την CAZy ως πηγή αναφοράς για τις γλυκο-επιστήμες. Ωστόσο, θα ήταν χρήσιμο, μια τέτοια βάση δεδομένων να είναι συμπληρωμένη με μια εγκυκλοπαιδική πηγή που θα είναι σε θέση να παρέχει στους ερευνητές ακριβή επισκόπηση της γνώσης για κάθε οικογένεια. Αυτή η διαπίστωση ήταν το βασικό κίνητρο για την ανάπτυξη της συμπληρωματικής ιστοσελίδας CAZypedia (http://www.cazypedia.org), που αποτελεί την λογική επέκταση της βάσης δεδομένων CAZy. Υπεύθυνος της CAZypedia είναι ο καθηγητής Harry Brumer του πανεπιστημίου British Columbia ενώ υποστηρίζεται από μια επιτροπή έμπειρων επιμελητών από όλο τον κόσμο, οι οποίοι επιζητούν υπεύθυνους επιμελητές και εξειδικευμένους συνεργάτες για να σχολιάζουν τα δεδομένα πετυχαίνοντας έτσι την συμμετοχή όλων των επιστημόνων που έχουν ως πεδίο έρευνας τις γλυκοεπιστήμες. Οι επιστήμονες αυτοί, τηρούν τις συμβάσεις ονοματοδοσίας που διέπουν το σύστημα ταξινόμησης CAZy. Κατά συνέπεια, συχνά αυτοί που ανακαλύπτουν μια νέα οικογένεια CAZymes, πριν από τη δημοσίευση της έρευνάς τους, ζητούν από τη βάση δεδομένων CAZy τον αριθμό της οικογένειας, ώστε να μπορούν να τον χρησιμοποιήσουν στην δημοσίευση. Ομοίως, όταν ανακαλυφθεί μια νέα δραστηριότητα σε μια υπάρχουσα οικογένεια, πολλοί από τους επιστήμονες ενημερώνουν τη βάση, προκειμένου να συμπληρωθεί η λειτουργική αυτή πληροφορία στην CAZy. Παρά τις προσπάθειες που έχουν γίνει μέχρι τώρα, οι πειραματικές πληροφορίες που παρουσιάζονται στην CAZy είναι αναγκαστικά ελλιπής. Οι ερευνητές μπορούν να βοηθήσουν επισημαίνοντας δεδομένα υποστρώματος / προϊόντων που έχουν δημοσιευθεί αλλά δεν έχουν ακόμα καταχωρηθεί στην CAZy. Επειδή η σύγχρονη βιοχημεία σταδιακά δημιουργεί πολύ μεγάλα σύνολα δεδομένων με τις ενεργότητες να αναφέρονται στα δημοσιευμένα άρθρα σε δεκάδες (και σύντομα χιλιάδες) ενζύμων, θα ήταν μεγάλο πλεονέκτημα αν οι ερευνητές κατέθεταν τα δεδομένα που χρησιμοποίησαν ως συμπληρωματικό υλικό σε μορφή πίνακα, που θα περιελάμβανε τη σειρά καταχώρησης στη βάση δεδομένων για κάθε χαρακτηρισμένο ένζυμο, τα υποστρώματα που χρησιμοποιήθηκαν και τα προϊόντα που ανιχνεύθηκαν. Αν τα επιστημονικά περιοδικά κάνουν υποχρεωτικό αυτόν τον απλό τρόπο καταχώρησης των δεδομένων, θα είναι δυνατή, προς όφελος όλων, μια πιο ολοκληρωμένη και αξιόπιστη συλλογή δεδομένων για τη βάση. Η πρακτική αυτή θα διευκόλυνε την λειτουργική συλλογή δεδομένων και σε άλλες βάσεις δεδομένων εκτός της CAZy.

MEROPS: Η βάση Merops (http://merops.sanger.ac.uk) αποτελεί μια βάση ταξινόμησης και ονοματολογίας πρωτεολυτικών ενζύμων και των πρωτεϊνών και μικρών μορίων αναστολέων που επηρεάζουν την ενζυματική τους δράση (Rawlings, Waller, Barrett, & Bateman, 2014). Τα πρωτεολυτικά ένζυμα έχουν πολλές βιολογικές λειτουργίες, που περιλαμβάνουν την πέψη των πρωτεϊνών, την ανακύκλωση των πρωτεϊνών, την επεξεργασία και μετατόπιση των νεοσυντιθέμενων πρωτεϊνών, την αφαίρεση των σηματοδοτικών αλληλουχιών στόχευσης, την ενεργοποίηση (και απενεργοποίηση) των ενζύμων, τις πεπτιδικές ορμόνες, τους υποδοχείς κυτταρικής επιφάνειας και τους νευροδιαβιβαστές, την αναδιαμόρφωση στις εξωκυττάριες μήτρες, την πήξη του αίματος και την ινωδόλυση. Οι πεπτιδάσες εμπλέκονται σε ευρύ φάσμα ασθενειών (ιική λοίμωξη, εισβολή παράσιτων, καρκίνο, διαβήτη τύπου ΙΙ, οστεοαρθρίτιδα στη νόσο Alzheimer), και χρησιμοποιούνται

συχνά στη βιομηχανία (βιολογικά απορυπαντικά, βυρσοδεψία, παρασκευή τυριών και σάλτσα σόγιας για εργαστηριακή χρήση στην πρωτεομική φασματοσκοπία μάζας, προσδιορισμό αλληλουχίας πρωτεϊνών). Η βάση δεδομένων δημιουργήθηκε το 1996 και σήμερα περιλαμβάνει πάνω από 400.000 αλληλουχίες πεπτιδασών. Οι αλληλουχίες που έχουν/μοιράζονται παρόμοια πρωτεϊνική αναδίπλωση οργανώνονται σε μια υπεροικογένεια (clan). Οι αλληλουχίες που έχουν ομοιότητες στην περιοχή της πεπτιδάσης οργανώνονται σε οικογένειες. Συνολικά υπάρχουν 61 υπεροικογένειες, 251 οικογένειες και 4.236 αναγνωριστικά (εκ των οποίων μόνο τα 377 περιλαμβάνονται στο *Enzyme Nomenclature*). Η συλλογή των δεδομένων έχει επεκταθεί για να συμπεριλάβει πάνω από 28.000 αναστολείς πεπτιδάσης που προέρχονται από γονιδιακά προϊόντα, καθώς και πάνω από 1.200 μικρά μόρια αναστολείς. Στη βάση περιλαμβάνονται αναφορές από πάνω από 53.000 δημοσιεύσεις και συνεργάζεται με διάφορες άλλες βάσεις δεδομένων αλληλουχίας πρωτεϊνών, συμπεριλαμβανομένων των UniProt, Pfam και Interpro. Όπως συμβαίνει και σε άλλες εξειδικευμένες βάσεις πρωτεϊνών, υπάρχει επίγνωση των λαθών στα πρωτογενή δεδομένα που διαρρέουν σε άλλες βάσεις. Είναι σχεδόν αδύνατο να διορθωθούν σφάλματα στις βάσεις πρωτογενών αλληλουχιών χωρίς τη συγκατάθεση των ατόμων που τις καταχώρησαν. Οι σχολιαστές των γονιδιωμάτων θα πρέπει να γνωρίζουν ότι τα ένζυμα, και ιδιαίτερα οι πεπτιδάσες για να μπορούν να χρησιμοποιηθούν, πρέπει να συνοδεύονται από την πλήρ γνώση των κατάλοιπων του ενεργού κέντρου, και ότι αν σε κάποια καταχώρηση αλληλουχίας λείπει οποιοδήποτε από αυτά δεν πρέπει να σχολιάζεται ως ενεργό ένζυμο. Ο μεγάλος και διαρκώς αυξανόμενος όγκος δημοσιευμένων δεδομένων, καθιστά απαραίτητη τη συμμετοχή όλων των μελών της επιστημονικής κοινότητας στο σχολιασμό. Τα οφέλη για τον ερευνητή που συμβάλλει σε μία βιολογική βάση δεδομένων είναι: η αναγνώριση της συνεισφοράς του, η προβολή των δημοσιεύσεών του, η βελτίωση των συλλογών δεδομένων και η διόρθωση λαθών, ενώ επίσης βοηθά άλλους ερευνητές που χρησιμοποιούν τα δεδομένα.

neXtProt: Η neXtProt (http://www.nextprot.org/) είναι μια διαδικτυακή βάση δεδομένων πρωτεϊνών του ανθρώπινου οργανισμού (Lane et al., 2012). Ενσωματώνει πληροφορίες που προέρχονται από την UniProtKB/Swiss-Prot με μια πληθώρα άλλων στοιχείων που προέρχονται από τα αποθετήρια και βάσεις δεδομένων που περιέχουν αποτελέσματα πειραμάτων υψηλής απόδοσης στον τομέα της πρωτεομικής, μεταγραφομικής και γονιδιωματικής. Υπάρχουν μια σειρά από δυσκολίες για τη διατήρηση παρόμοιων πηγών με την neXtProt. Η πρώτη έγκειται στην επιλογή και αξιολόγηση της ποιότητας των πληροφοριών που καταχωρούνται στη βάση. Ένας από τους στόχους της ομάδας επιμέλειας της neXtProt είναι η ταξινόμηση των πειραματικών αποτελέσματων σε τρεις κατηγορίες: «χάλκινο» (> 5% ποσοστό σφάλματος), «αργυρό» (1-5% ποσοστό σφάλματος) και «χρυσό» (λιγότερο από 1% ποσοστό σφάλματος). Τα χάλκινα δεδομένα δεν ενσωματώνονται στην neXtProt. Η αξιολόγηση της ποιότητας των πειραματικών αποτελεσμάτων δεν είναι γενικά πολύ εύκολο να επιτευχθεί, δεδομένου ότι συχνά η καταχώρηση μελετών ή δεδομένων στα repositories δεν παρέχουν τα απαραίτητα κριτήρια για να αξιολογηθεί αντικειμενικά η ποιότητα της πειραματικής διάταξης αλλά ούτε και των αποτελεσμάτων. Στην ιδανική περίπτωση, οι εκτιμήσεις αυτές θα πρέπει να επανεξετάζονται σε τακτά χρονικά διαστήματα, όταν οι τεχνικές αλλάζουν και έχουν αντικατασταθεί από καλύτερες και πιο ακριβείς μεθόδους. Μια άλλη σημαντική πρόκληση για πηγές παρόμοιες με την neXtProt που προσπαθούν να ενσωματώσουν μεγάλη ποικιλία πληροφοριών που προέρχονται από πολλές ετερογενείς πηγές είναι η συνεχής ανάγκη τροποποίησης και ενημέρωσης της πληροφορίας που παρέχεται από τη βάση δεδομένων. Η neXtProt προσπαθεί να ακολουθήσει ένα μηνιαίο χρονοδιάγραμμα έκδοσης δημοσιεύσεων αλλά αυτό μερικές φορές διαταράσσεται από αλλαγές σε τουλάχιστον μία από τις ενσωματωμένες πηγές. Σημαντικό πρόβλημα δημιουργούν αλλαγές στην μορφή των δεδομένων. Τέλος, σημαντικό πρόβλημα αποτελεί ότι οι πηγές που ενσωματώνονται στην βάση δεν έχουν όλες την ίδια προτυποποίηση. Η παραγωγή και η διατήρηση πινάκων αντιστοίχισης μεταξύ διαφορετικών οντολογιών ή ελεγχόμενων λεξιλόγιων είναι γενικά απαραίτητη. Αυτό το πρόβλημα είναι ιδιαίτερα οξύ στην περίπτωση των ανθρώπινων ασθενειών καθώς υπάρχουν πάνω από 10 οντολογίες που χρησιμοποιούνται από τις επιστημονικές κοινότητες της ιατρικής και των επιστημών ζωής.

PASS2: Η βάση PASS2 περιέχει στοιχίσεις δομών πρωτεϊνικών αλληλουχιών σε επίπεδο υπερ-οικογενειών (Protein sequence Alignments of Structural Superfamilies). Η πρώτη έκδοση της βάσης αναφέρεται ως «CAMPASS» (Sowdhamini et al., 1998). Η πολλαπλή στοίχιση αλληλουχιών μελών μιας υπερ-οικογένειας πρωτεϊνών που διαφέρουν μεταξύ τους, αποτελεί δύσκολη διαδικασία εξαιτίας την μικρής ομοιότητας των αλληλουχιών, παρόλο που μπορεί να υπάρχουν αναμφισβήτητες εξελικτικές συνδέσεις, λειτουργικές και δομικές ομοιότητες. Οι προηγούμενες εκδόσεις της PASS2 ήταν σε μορφή HTML, ενώ η τρέχουσα έκδοση της (PASS2.4) (Gandhimathi, Nair, & Sowdhamini, 2012), λειτουργεί σε μια πλατφόρμα MYSQL με διεπαφή

σε PHP. Αυτή η έκδοση της βάσης, η οποία είναι σε άμεση αντιστοιχία με την SCOP 1.75 (Murzin, Brenner, Hubbard, & Chothia, 1995) για τον ορισμό των μελών της υπερ-οικογένειας, αυτή τη στιγμή προσφέρει στοιχίσεις αλληλουχιών βάσει της δομής 1961 υπεροικογενειών. Σημαντική πρόκληση για τη διατήρηση και την ενημέρωση της βάσης, λαμβάνοντας υπόψη τη συνεχή συσσώρευση πρόσθετων μελών και υπερ-οικογενειών, είναι η μείωση της χειρωνακτικής παρέμβασης, και η αυτοματοποίηση όσο το δυνατόν περισσότερο της διαδικασίας, διατηρώντας όμως την ποιότητα των δεδομένων σε υψηλό επίπεδο. Αυτό είναι πράγματι δύσκολο, δεδομένου ότι η εξέλιξη φέρνει μαζί της διαφοροποιήσεις, και αυτό σημαίνει ότι θα μπορούσαν να υπάρχουν αρκετά «outliers» (Gandhimathi, et al., 2012), τα οποία είναι δύσκολο να εντοπιστούν κατά τη διάρκεια της αυτόματης στοίχισης των πρωτεϊνικών αυτοτελών δομικών περιοχών των υπερ-οικογενειών. Η μελέτη των λειτουργικών αποκλίσεων των καταλοίπων και της ειδικής κατηγορίας της φύσης των διατηρημένων κατάλοιπων ή μοτίβων από τους πειραματιστές σε ένα ελεγχόμενο λεξιλόγιο στις αναφορές προσδιορισμού της δομής τους, θα μπορούσε να καταστήσει δυνατή την έγκαιρη αναγνώριση των outliers.

KinG: Η βάση δεδομένων Kinases in Genomes (KinG) αποτελεί μία πηγή κινασών Ser/Thr /Tyr που κωδικοποιούνται στα πλήρως αλληλουχημένα γονιδιώματα των προκαρυωτικών, ιικών και ευκαρυωτικών κυττάρων (Krupa, Abhinandan, & Srinivasan, 2004). Το πλήρες ρεπερτόριο των κινασών πρωτεϊνών σε διάφορα πλήρως αλληλουχημένα γονιδιώματα παρουσιάζεται στο δίκτυο Garuda India στην ιστοσελίδα http://megha.garudaindia.in/king/. Το δίκτυο παρέχει λεπτομερή κατάλογο των Ser/Thr/Tyr και άτυπων κινασών πρωτεϊνών διάφορων οργανισμών συνοδευόμενα από χαρακτηριστικά, όπως η ταξινόμηση σε υπο-οικογένειες πρωτεϊνικών κινασών και η οργάνωση των αυτοτελών δομικών περιοχών. Η βάση επιτρέπει επίσης την ανάκτηση των κινασών πρωτεϊνών που ανήκουν σε καθορισμένη υπο-οικογένεια ή σε συγκεκριμένους συνδυασμούς αυτοτελών δομικών περιοχών. Ο χρήστης μπορεί αναζητήσει συγκεκριμένες αλληλουχίες ώστε να προσδιορίσει την καταλυτική περιοχή της κινάσης και τα διάφορα λειτουργικά κατάλοιπα στην καταλυτική περιοχή. Στην πρώτη έκδοση της KinG που δημοσιεύθηκε το 2004 (Krupa, et al., 2004), δημοσιεύθηκαν κινάσες μόνο από 40 οργανισμούς. Η KinG ανανεώνεται κάθε χρόνο. Οι Κινάσες εκφράζονται έντονα, ειδικά στους ευκαρυωτικούς οργανισμούς. Επιπλέον, καθώς ο αριθμός των πλήρως αλληλουχημένων γονιδιωμάτων αυξάνεται με γρήγορο ρυθμό, σε κάθε ανανέωση της βάσης αυξάνεται και ο αριθμός των κινασών που πρέπει να διαχειριστούν. Στην τρέχουσα έκδοση της KinG μελετώνται 12200 ομάδες δεδομένων γονιδιωματικής με αποτέλεσμα τον εντοπισμό και την ταξινόμηση 131.921 κινασών. Εκτός από το ότι πρέπει η βάση να συμβαδίζει με την αύξηση του αριθμού των κινασών, υπάρχουν μερικές επιπλέον ενδιαφέρουσες προκλήσεις που πρέπει να αντιμετωπιστούν. Η ταξινόμηση των κινασών σε υπο-οικογένειες στην βάση KinG πραγματοποιείται σύμφωνα με το σύστημα ταξινόμησης των Hanks και Hunter's (Hanks & Hunter, 1995) προσαρμοσμένο με μια προσέγγιση πολλαπλών ειδικών ανά θέση πινάκων (multiple position-specific scoring matrices - PSSM) (Gowri, Krishnadev, Swamy, & Srinivasan, 2006). Η ομαδοποίηση των αλληλουχιών κινάσης σε αυτές τις υπο-οικογένειες οδηγεί σε αναγνώριση γνήσιων υπο-οικογενειών που δεν περιέχονται στο αρχικό πλαίσιο ταξινόμησης. Ως εκ τούτου, τα συστήματα ταξινόμησης αναδιοργανώνονται, προκειμένου να συμπεριληφθούν όσο το δυνατόν περισσότερες κινάσες. Μια άλλη πρόκληση είναι η ασυμφωνία μεταξύ της ταξινόμησης σε υπο-οικογένειες, η οποία βασίζεται αποκλειστικά στις αλληλουχίες των καταλυτικών περιοχών, και τους συνδυασμούς των περιοχών των κινασών. Πρόσφατες αναλύσεις (Deshmukh, Anamika, & Srinivasan, 2010; Rakshambikai, Gnanavel, & Srinivasan, 2014) επέτρεψαν την αναγνώριση της εμφάνισης υβριδικών (hybrid) κινασών οι οποίες χαρακτηρίζονται από μια υπο-οικογένεια κινάσης, που αναγνωρίζεται με βάση μόνο την αλληλουχία των καταλυτικών περιοχών, και η οποία χαρακτηρίζεται σε μια άλλη υπο-οικογένεια κινασών με βάση την αρχιτεκτονική της καταλυτικής περιοχής. Αυτή η περίπλοκη κατάσταση δεν επιτρέπει την ταξινόμηση της κινάσης σε καμία από τις δύο υπο-οικογένειες, και ως εκ τούτου προτείνεται να ταξινομούνται ως υβριδικές κινάσες με χαρακτηριστικά των δύο διαφορετικών υπο-οικογενειών κινάσης. Επιπλέον, δυσκολία στην ταξινόμηση προκαλείται από την εμφάνιση κινασών που η καταλυτική περιοχή τους συνδέεται με μία συγκεκριμένη οικογένεια κινασών αλλά η συν-ύπαρξη των περιοχών δεν χαρακτηρίζεται από κάποια υπο-οικογένεια. Γι' αυτό τα χαρακτηριστικά τους αποκλίνουν από αυτά των υπόλοιπων μελών της υπο-οικογένειας. Αυτές οι κινάσες ονομάζονται κινάσες rogue (Deshmukh, et al., 2010; Rakshambikai, et al., 2014). Ο σχεδιασμός του προτεινόμενου συστήματος ταξινόμησης των κινασών διευκολύνεται από την ανάλυση δεδομένων υψηλής απόδοσης που προκύπτουν συνεχώς. Τα δεδομένα και το σύστημα ταξινόμησης τροφοδοτούν το ένα το άλλο, καθώς βελτιώνεται συνεχώς το σύστημα ταξινόμησης (Bhaskara et al., 2014; Gnanavel et al., 2014; J. Martin, Anamika, & Srinivasan, 2010) θα συνεχίσει η ενημέρωση της βάσης δεδομένων KinG. Η παρούσα έκδοση της KinG έχει

δημιουργηθεί χρησιμοποιώντας το NetBeans IDE σε πυρήνα Java, JSP, Servlets, AJAX, Jquery, XML, HTML και CSS ενώ το περιβάλλον της είναι φιλικό προς το χρήστη, και οι αναζητήσεις πραγματοποιούνται γρήγορα.

EzCatDB: Η βάση EzCatDB (http://ezcatdb.cbrc.jp/EzCatDB/) δημιουργήθηκε το 2004 με στόχο να αποτελέσει έναν οδηγό κατάταξης ενζυμικών αντιδράσεων, των δομών των ενεργών κέντρων των ενζύμων, αλλά και των καταλυτικών τους μηχανισμών. Βασίζεται σε πληροφορίες από την βιβλιογραφία (Nagano, 2005; Nagano et al., 2014) και διαφέρει από την Enzyme Commission (E.C.) (NC-IUBMB; http://www.chem.qmul.ac.uk/iubmb/enzyme/) η οποία ταξινομεί τα ένζυμα με βάση τις χημικές δομές των υποστρωμάτων και των προϊόντων (Fleischmann et al., 2004; McDonald, Boyce, & Tipton, 2009; Tipton, 1994). Αν και η ταξινόμηση της ιεραρχικής αντίδρασης (RLCP) στην EzCatDB αρχικά περιελάμβανε μόνο αντιδράσεις των πυρηνόφιλων υποκαταστάσεων (nucleophilic substitution reactions), όπως υδρολύσεις και αντιδράσεις μεταφοράς, στη συνέχεια επεκτάθηκε και σε άλλες αντιδράσεις όπως προσθήκης, αφαίρεσης, ισομερισμού, μεταφοράς υδριδίου και μεταφοράς ηλεκτρονίων (Nagano, et al., 2014).

Η EzCatDB περιλαμβάνει τις τριτοταγείς δομές των ενζύμων της Protein Data Bank (PDB) (Rose et al., 2013) και τα αντίστοιχα δεδομένα αλληλουχίας αμινοξέων της UniProt, ιδιαίτερα με την αντίστοιχη ταξινόμηση CATH (Cuff et al., 2011). Εκτός από αυτές τις βάσεις, για τα δεδομένα ενώσεων που σχετίζονται με τα ένζυμα λαμβάνεται υπόψη και η βάση KEGG (Kanehisa, et al., 2014). Στην EzCatDB χρησιμοποιείται το σύστημα διαχείρισης βάσεων δεδομένων PostgreSQL. Συνεπώς η αναζήτηση των δεδομένων ενός ενζύμου μπορεί να γίνει με διάφορους τρόπους (Nagano, 2005) όπως για παράδειγμα χρησιμοποιώντας τον αριθμό E.C., τα IDs από άλλες βάσεις δεδομένων, τους τύπους των αμινοξέων που βρίσκονται στο ενεργό κέντρο των ενζύμων και τους τύπους των προσδετών που μπορούν να συνδυαστούν για την αναζήτηση (Nagano, 2005). Επιπλέον, για κάθε εγγραφή είναι δυνατή η δημιουργία ενός πίνακα σχολιασμού των προσδετών για τα δεδομένα της PDB, στον οποίο τα μόρια προσδέτη που συνδέονται με τις δομές του ενζύμου έχουν περιγραφεί χειροκίνητα ως συμπαράγοντες, υποστρώματα, προϊόντα ή ενδιάμεσοι, φυσικοί προσδέτες ή ανάλογοι προσδέτες. Επίσης δημιουργείται ένας πίνακας με πληροφορίες σχετικά με τα αμινοξικά κατάλοιπα που βρίσκονται στο ενεργό κέντρο του ενζύμου (Nagano, 2005). Τα δεδομένα αυτά είναι απαραίτητα για την κατανόηση των καταλυτικών μηχανισμών του ενζύμου.

Όλες οι διαδικασίες που σχετίζονται με τον χειρωνακτικό σχολιασμό είναι χρονοβόρες, λόγω του μεγάλου όγκου των δεδομένων και της δυσκολίας αναζήτησης στη βιβλιογραφία. Συγκεκριμένα, η εξαγωγή και ανάλυση πληροφορίας από την βιβλιογραφία είναι η πιο χρονοβόρα και απαιτεί τοπική αποθήκευση της λίστας των δημοσιεύσεων, παραγγελία του πλήρους κειμένου, αναζήτηση των λέξεων κλειδιών κλπ. Η EzCatDB περιέχει σήμερα 871 εγγραφές ένζυμων, που αφορούν 1.610 αλληλουχίες της UniProtKB και 6.704 εγγραφές της PDB. Είναι επομένως φανερό ότι τα διαθέσιμα δεδομένα ενζύμων είναι περιορισμένα. Επιπλέον βρίσκονται στο στάδιο της επεξεργασίας 300 εγγραφές που όμως αποτελεί δύσκολη διαδικασία λόγω των περιορισμών στο ανθρώπινο δυναμικό και στη χρηματοδότηση.

MACiE: Η MACiE (Mechanism, Annotation and Classification in Enzymes, http://www.ebi.ac.uk/thornton-srv/databases/MACiE), είναι μια βάση δεδομένων που περιέχει μηχανισμούς αντίδρασης ενζύμων (Holliday et al., 2012). Στην MACiE συγκεντρώνονται και αποθηκεύονται πληροφορίες σχετικά με τα ένζυμα, τους συνολικούς χημικούς μετασχηματισμούς τους, τους μηχανισμούς αντίδρασης, τους συμπαράγοντες και τα καταλυτικά κατάλοιπα. Κάθε εγγραφή της MACiE αντιστοιχίζεται σε τουλάχιστον μια κρυσταλλική δομή στην PDB και σε έναν καλά καθορισμένο μηχανισμό από την πρωτογενή βιβλιογραφία. Τα δεδομένα της MACiE μπορούν να θεωρηθούν ως μια εννοιολογική ιεράρχηση, η οποία προκύπτει από το γεγονός ότι ένα ένζυμο μπορεί να οριστεί σαφώς από τα στοιχεία του.

Με αυτήν την ιεράρχηση ένα ένζυμο μπορεί να οριστεί κατά την πιο απλή του μορφή, σαν ένα βιοπολυμερές που έχει μια πρωτοταγή αμινοξική αλληλουχία και καταλύει έναν συνολικό χημικό μετασχηματισμό (ο ορισμός αυτός δεν περιλαμβάνει τα ριβοένζυμα). Ένας συνολικός χημικός μετασχηματισμός πρέπει να αποτελείται από τουλάχιστον ένα υπόστρωμα και ένα προϊόν, και να έχει ένα μηχανισμό που όμως είναι πιθανό να μην γνωστός με κάθε λεπτομέρεια. Το γεγονός ότι τα δεδομένα αυτά μπορούν να διαταχθούν ιεραρχικά αναδεικνύει την σχέση που υπάρχει μεταξύ τους καθώς και ότι είναι δυνατή η περιγραφή τους σε μια σχεσιακή βάση δεδομένων. Η MACiE χρησιμοποιεί την ανοιχτή βάση δεδομένων MySQL. Αυτό το επίπεδο των σχεσιακών πληροφοριών επιτρέπει την γρήγορη εκτέλεση σύνθετων αναζητήσεων. Τέτοιες ιεραρχικές αναζητήσεις έχουν ήδη υλοποιηθεί πολλές φορές στην ιστοσελίδα της MACiE (Holliday et al., 2007).

Με τη χρήση των σχέσεων αυτών είναι εύκολο να περάσουμε από το ένα δεδομένο στο άλλο, με την προϋπόθεση ότι και τα δύο είναι στοιχεία που απαιτούνται για τον καθορισμό του ενζύμου. Η MACiE περιέχει μεγάλο αριθμό μετα-δεδομένων που μπορούν να συνδεθούν με τα ένζυμα και τους μηχανισμούς αντίδρασης τους. Τα βασικά στοιχεία περιλαμβάνουν (1) τις λεπτομερείς λειτουργίες κάθε καταλυτικού αμινοξικού κατάλοιπου στην θέση κατάλυσης, (2) την παρουσία καταλυτικών δυάδων ή τριάδων, (3), τη μηχανιστική περιγραφή του κάθε σταδίου αντίδρασης, (4) τις μεταβολές των δεσμών, (5) τα κέντρα των αντιδράσεων, τους συμπαράγοντες και τις λειτουργίες τους, και (6) τα στοιχεία σύνδεσης με εξωτερικές βάσεις δεδομένων, όπως για παράδειγμα ο αριθμός EC, τα αναγνωριστικά της UniProtKB, CATH, και οι κωδικοί PDB.

Μία τελική βασική συνιστώσα της δομής των δεδομένων της βάσης MACiE είναι η χρήση ενός αυστηρά ελεγχόμενου λεξιλογίου και ο εκτενής έλεγχος των σφαλμάτων κατά την καταχώρηση. Ελέγχεται αν η πρωτεΐνη έχει ήδη καταχωρηθεί στην MACiE και αν τα σχολιασμένα αμινοξικά κατάλοιπα υπάρχουν στην κρυσταλλική δομή. Εξαιρέσεις γίνονται όταν η καταλυτική μονάδα, δηλαδή η μικρότερη μονάδα που απαιτείται για να συμβεί η κατάλυση, δεν αντιστοιχεί στην ασύμμετρη μονάδα στο αρχείο PDB. Επίσης ελέγχεται εάν τα σχόλια ταιριάζουν με την ιεράρχηση, π.χ., όταν ένα κατάλοιπο είναι σχολιασμένο ως αντιδρόν, αλλά δεν έχει σχολιασμένη λειτουργία. Οι έλεγχοι βοηθούν στην ελαχιστοποίηση των ανθρώπινων λαθών, στην ανάθεση της λειτουργίας και του μηχανισμού των εγγραφών, που είναι πάντα δυνατή με τον χειροκίνητι σχολιασμό.

ESTHER:Η βάση δεδομένων ESTHER (ESTerases and alpha/beta-Hydrolase Enzymes and Relatives) περιέχει την ανάλυση πρωτεϊνών που ανήκουν στην υπερ-οικογένεια των α/β-υδρολάσων (www.bioweb.supagro.inra.fr/esther). Οι α/β υδρολάσες αποτελούν μια από τις μεγαλύτερες και πιο πολυποίκιλες υπερ-οικογένειες πρωτεϊνών που χαρακτηρίζονται ένα μονο είδος διπλώματος. Μέχρι στιγμής περιλαμβάνει περισσότερες από 800.000 αλληλουχίες (που αντιστοιχοπυν σε 42.000 μη ομόλογες εγγραφές) ομαδοποιημένες σε 175 υπο-οικογένειες (Lenfant, Hotelier, Bourne, Marchot, & Chatonnet, 2013). Κάθε υπο-οικογένεια έχει δημιουργηθεί σύμφωνα με ένα προφίλ HMM (Lenfant et al., 2013). Μέλη της υπερ-οικογένειας έχουν βασικό ρόλο σχεδόν σε όλες τις φυσιολογικές διαδικασίες και αποτελούν στόχους φαρμάκων για την θεραπεία ασθενειών όπως ο διαβήτης, η παχυσαρκία, και οι νευροεκφυλιστικές διαταραχές. Παρά τις προτεινόμενες κοινές ονομασίες τους, πολλές από αυτές τις πρωτεΐνες δεν είναι ένζυμα, καθώς κάποιες από αυτές έχουν χάσει όλα τα αναγκαία κατάλοιπα που δυνητικά μπορούν να αποτελέσουν ένα ενεργό κέντρο (Lenfant, Hotelier, Bourne, Marchot, & Chatonnet, 2014; Marchot & Chatonnet, 2012). Οι λειτουργίες λίγων εκπροσώπων αυτής της τελευταίας ομάδας είναι γνωστές: ενδοκυτταρικοί υποδοχείς μικρών μορίων, πρόδρομα μόρια ορμονών, αλληλεπίδραση με μόρια σε κυτταρικές επιφάνειες κ.ά. Ένας από τους στόχους της βάσης είναι η σύνδεση των βιολογικών δεδομένων με τις διαφορετικές υπο-οικογένειες, προκειμένου να συμβάλλει στον καθορισμό της λειτουργίας τους. Η βάση ESTHER δημιουργήθηκε το 1994 σε εξυπηρετητή Gopher και γρήγορα πέρασε στο WWW (Cousin, Hotelier, Lievin, Toutant, & Chatonnet, 1996). Το σύστημα στο οποίο βασίζεται είναι το ACeDB.

Η βάση δεδομένων περιέχει επίσης μικρά μόρια που αλληλεπιδρούν με εστεράσες ως υποστρώματα, αναστολείς ή ενεργοποιητές και άλλα συναφή κινητικά δεδομένα (Chatonnet, Cousin, & Robinson, 2001). Στην παρούσα φάση γίνεται προσπάθεια επέκτασης αυτής της ενότητας. Έχει ενσωματωθεί το πακέτο R, προκειμένου να επιτραπεί η στατιστική σύγκριση των κινητικών παραμέτρων των διαφόρων ενζύμων ή μεταλλακτών με διάφορα υποστρώματα και / ή των αναστολέων κάτω από διαφορετικές πειραματικές συνθήκες.

ConoServer: Το δηλητήριο του σαλιγκαριού Cone είναι πιθανά μια μεγάλη πηγή αρκετών εκατοντάδων χιλιάδων ενεργών πεπτιδίων εξαιρετικά επιλεκτικών για τους υποδοχείς και μεταφορείς του νευρικού συστήματος με εφαρμογές σε νευρολογικούς ανιχνευτές και φάρμακα (Akondi et al., 2014; Terlau & Olivera, 2004). Η ποικιλομορφία αυτών των δηλητηρίων (Davis, Jones, & Lewis, 2009) έχει αναλυθεί σε μελέτες γενετικής (Biggs et al., 2010; Chang & Duda, 2012; Puillandre, Koua, Favreau, Olivera, & Stocklin, 2012) και οικολογικές (Duda, Chang, Lewis, & Lee, 2009; Duda & Lee, 2009) μελέτες. Τον Δεκέμβριο του 2014, η βάση δεδομένων ConoServer (Kaas, Yu, Jin, Dutertre, & Craik, 2012) περιείχε περισσότερα από 2000 κονο-πεπτίδια, ενώ βοηθά στη συστηματοποίηση των τριών συστημάτων ταξινόμησης που περιγράφουν την εξέλιξη του κονοπεπτιδίου, τρισδιάστατες δομές και μοριακούς στόχους (Kaas, Westermann, & Craik, 2010).

Η βάση ConoServer (www.conoserver.org) δημιουργήθηκε (όπως και ο σχολιασμός) με στόχο την όσο το δυνατόν μικρότερη απαίτηση ανθρώπινων πόρων, δηλαδή ένα άτομο. Ο στόχος αυτός επετεύχθηκε με

την εφαρμογή ενός επιπλέον επιπέδου. Επινοήθηκε ένας ψευδο-πίνακας που συνδέει τα δεδομένα που είναι αποθηκευμένα στην σχεσιακή βάση δεδομένων MySQL σε περιβάλλον PHP. Τα πρωτογενή δεδομένα που υποστηρίζουν την ConoServer προέρχονται από την GenBank (Benson, et al., 2014), την UniProt-KB (UniProt, 2014) και την PDB (Berman, Henrick, Nakamura, & Markley, 2007), καθώς και από τη βιβλιογραφία ή των υποβολών από τους συγγραφείς. Οι πρόσφατα ανακτημένες αλληλουχίες, πριν δημοσιευτούν, σχολιάζονται από διάφορα κείμενα που εισάγουν δεδομένα στη βάση MySQL, τα δεδομένα αυτά στη συνέχεια αναθεωρούνται και τροποποιούνται χειροκίνητα μέσω μιας διεπαφής σχολιασμού.

Ένας αριθμός αλληλουχιών κονοπεπτιδίων είναι διαθέσιμος μόνο σε πίνακες, σχήματα ή συμπληρωματικές πληροφορίες από έγκριτα επιστημονικά άρθρα, και έχουν εισαχθεί χειροκίνητα μέσω μιας διεπαφής του διαδικτύου. Οι κονοπεπτιδικές αλληλουχίες των πρωτεϊνών και των προδρόμων αλληλουχιών νουκλεϊκών οξέων σχολιάζονται βάσει των χαρακτηριστικών των αλληλουχιών και ταξινομούνται σύμφωνα με τρία τυποποιημένα συστήματα ταξινόμησης. Το ConoPrec (Kaas, et al., 2012) είναι ένα διαδικτυακό εργαλείο το οποίο επιτρέπει τη χρήση αυτής της διαδικασίας σχολιασμού, βοηθώντας τους χρήστες να αναλύσουν αλληλουχίες χρησιμοποιώντας πρότυπα της ConoServer πριν από τη δημοσίευση τους, απλοποιώντας έτσι αργότερα την είσοδό τους στην ConoServer.

CyBase: Οι ριβοσωμικώς συντιθέμενες κυκλικές πρωτεΐνες, έχουν παρατηρηθεί σε όλα τα βασίλεια της ζωής (Craik, 2006; Kedarisetti, Mizianty, Kaas, Craik, & Kurgan, 2014). Η κυκλοποίηση της κύριας αλυσίδας καθιστά τις πρωτεΐνες αδιαπέραστες στις εξωπρωτεάσες και οδηγεί σε δραματική βελτίωση της σταθερότητάς τους έναντι της ενζυματικής αποικοδόμησης καθώς και της θερμικής ή χαοτρόπου αποδιάταξης (Trabi & Craik, 2002). Η υψηλή σταθερότητα των κυκλικών πεπτιδίων και πρωτεϊνών έχει προσελκύσει έντονα το ενδιαφέρον των σχεδιαστών φαρμάκων για τη σταθεροποίηση βιοδραστικών πεπτιδικών επιτόπων (Poth, Chan, & Craik, 2013). Η βάση δεδομένων CyBase (www.cybase.org.au) είναι μια βάση που παρέχει πρόσβαση σε πληροφορίες σχετικά με την κωδικοποίηση των γονιδίων, την κύρια αλυσίδα και τις κυκλικές πρωτεΐνες (Wang, Kaas, Chiche, & Craik, 2008). Από τον Δεκέμβριο του 2014 η CyBase περιέχει πληροφορίες για περίπου 420 φυσικά δημιουργημένες και περίπου 160 συνθετικές κυκλικές πρωτεΐνες. Αυτές οι πρωτεΐνες έχουν ταξινομηθεί σε εννέα κύριες κατηγορίες, η μεγαλύτερη των οποίων είναι η κατηγορία cyclotide, με 282 εγγραφές. Οι στρατηγικές καταχώρησης και σχολιασμού της CyBase είναι παρόμοιες με αυτές που περιγράφονται στην βάση ConoServer. Στην CyBase πραγματοποιείται αναζήτηση αλγόριθμων που έχουν προσαρμοστεί για τον χειρισμό των κυκλικών πρωτεϊνών, ενώ τα εργαλεία που χρησιμοποιούνται συχνότερα είναι η στοίχιση αλληλουχίας και η φασματομετρία μάζας στις αναζητήσεις αποτυπωμάτων (Wang, et al., 2008). Ένα ιδιαίτερο χαρακτηριστικό της CyBase είναι η σε βάθος περιγραφή του κειμένου βιολογικής ανάλυσης και φυσικοχημικών χαρακτηρισμών της κάθε κυκλικής πρωτεΐνης.

GPCRDB: Οι υποδοχείς που είναι συζευγμένοι με G-πρωτεΐνες (G protein-coupled receptors, GPCRs) αποτελούν τη μεγαλύτερη οικογένεια μεμβρανικών πρωτεϊνών σε ορισμένους ευκαρυωτικούς οργανισμούς. Ρυθμίζουν μια πληθώρα φυσιολογικών διεργασιών που εκτείνονται από το νευρικό και ενδοκρινικό σύστημα, μέχρι και την αίσθηση των οσμών, της γεύση και του φωτός (Bockaert & Pin, 1999; Lagerstrom & Schioth, 2008). Αποτελούν τους στόχους περίπου του 30% των φαρμάκων της αγοράς, αν και μέχρι σήμερα έχουν αξιοποιηθεί θεραπευτικά λίγοι μόνο από τους υποδοχείς (Garland, 2013; Overington, Al-Lazikani, & Hopkins, 2006). Η βάση δεδομένων GPCR, η GPCRDB (http://gpcrdb.org), ξεκίνησε το 1993. Εκείνη την εποχή ταυτοποιήθηκε μέσω κλωνοποίησης γονιδίων ένας μεγάλος αριθμός αλληλουχιών υποδοχέων, και, καθώς δεν είχαν ακόμη έχουν δημιουργηθεί οι περιηγητές του διαδικτύου, η GPCRDB ήταν αρχικά ένα αυτόματο σύστημα απόκρισης ηλεκτρονικών μηνυμάτων το οποίο μπορούσε να στέλνει αλληλουχίες, στοιχίσεις και μοντέλα ομολογίας. Μετά από δύο δεκαετίες, η GPCRDB εξελίχθηκε σε ένα ολοκληρωμένο πληροφοριακό σύστημα (Horn et al., 2003; Horn et al., 1998; Vroling et al., 2011). Το 2013, η GPCRDB μεταφέρθηκε στην ομάδα Gloriam του Πανεπιστημίου της Κοπεγχάγης, η οποία υποστηρίζεται από το EU COST GPCR Action 'GLISTEN'. Σήμερα, η GPCRDB στοχεύει σε ένα διεπιστημονικό κοινό αντί να αποτελεί πηγή κυρίως για βιοπληροφορικούς. Αυτό περιλαμβάνει την δημοσίευση νέων δεδομένων φιλικών προς το χρήστη, διαγράμματα και εργαλεία, καθώς και αναφορές με σημαντικές συμπληρωματικές βάσεις δεδομένων (Isberg et al., 2014).

Η GPCRDB περιέχει τις μεγαλύτερες συλλογές in vitro μεταλλάξεων στα γονίδια των υποδοχέων οι οποίες δημιουργήθηκαν μετά από αρκετά χρόνια επιμέλειας της επιστημονικής βιβλιογραφίας. Αποτελεί μια ανοιχτή βάση δεδομένων και επιτρέπει τη συνεισφορά δεδομένων μεταλλαξιγένεσης από τους ερευνητές ώστε να αυξηθεί η διάδοση των δεδομένων και να είναι δυνατή η σύγκρισή τους με δεδομένα που έχουν ήδη

δημοσιευθεί. Οι μεταλλάξεις συχνά παρατίθενται μέσα στα διαγράμματα κατάλοιπων των υποδοχέων, τα οποία μπορούν να ανακτηθούν και να χρησιμοποιηθούν σε δημοσιεύσεις ή παρουσιάσεις. Η GPCRDB επίσης διατηρεί συλλογή επιμελημένων αναφορών όλων των κρυσταλλικών δομών GPCR, οι οποίες έχουν αυξηθεί εκθετικά σε αριθμό, λόγω των πρόσφατων τεχνολογικών ανακαλύψεων (Katritch, Cherezov, & Stevens, 2013). Οι δομές μπορούν να αναζητηθούν και να φιλτραριστούν από δεδομένα προσδετών και υποδοχέων και από μέτρα ομοιότητας του στόχου-πρότυπου αλληλουχιών. Πρόσθετα εργαλεία του εξυπηρετητή του διαδικτύου επιτρέπουν τη διαχείριση δομών, για παράδειγμα υπέρθεση στη συνολική δομή ή υπο-περιοχές κατάλοιπων που συνιστούν περιοχές πρόσδεσης του προσδέτη.

Η GPCRDB βρίσκεται στη διαδικασία μετάβασης σε πιο σύγχρονες τεχνολογίες διαδικτύου. Η νέα διεπαφή χρησιμοποιεί τεχνολογίες HTML5 (συμπεριλαμβανομένου του CSS3) και JavaScript ώστε να παρέχει στον χρήστη ένα διαδραστικό περιβάλλον. Για τη δημιουργία διαδραστικών διαγράμματων χρησιμοποιούνται Scalable Vector Graphics (SVG), που μπορούν να αποθηκευτούν σε υψηλή ανάλυση και να μπορούν παρουσιαστούν σε δημοσιεύσεις. Τα δεδομένα αποθηκεύονται χρησιμοποιώντας τη σχεσιακή βάση MySQL, ενώ ο εξυπηρετητής διαδικτύου Apache χρησιμοποιείται για τις ιστοσελίδες των χρηστών. Η GPCRDB προσφέρει υπηρεσίες διαδικτύου SOAP για πρόσβαση μέσω προγραμματισμού, και έχει σαν στόχο να καταστήσει περισσότερη πληροφορία/περιεχόμενο προσβάσιμο μέσω άλλων διαδικτυακών ιστότοπων.

IUPHAR/BPS Guide to PHARMACOLOGY: Η βάση IUPHAR/BPS Guide to PHARMACOLOGY (GtoPdb, (Pawson et al., 2014)) έχει αναπτυχθεί από κοινού από τη Διεθνή Ένωση Βασικής και Κλινικής Φαρμακολογίας (Union of Basic and Clinical Pharmacology, IUPHAR) και τη Βρετανική Φαρμακολογική Εταιρία (BPS) για να παρέχει πρόσβαση σε υψηλής ποιότητας πληροφορίες για φαρμακευτικούς στόχους. Η GtoPdb καταγράφει τιμές συγγένειας με την χαρτογράφηση βιοδραστικών χημικών δομών των πρωτεϊνών. Η GtoPdb (http://www.guidetopharmacology.org/) είναι μια βάση που συγκεντρώνει προηγουμένως διαφορετικές αλλά συμπληρωματικές πληροφορίες που είχαν αρχικά καταχωρηθεί στη βάση IUPHAR (IUPHAR-DB, (Harmar et al., 2009) και στον Οδηγό Υποδοχέων και Καναλιών (Guide to Receptors and Channels, GRAC), που αποτελεί σειρά δημοσιεύσεων στο περιοδικό BPS, British Journal of Pharmacology (π.χ., (Alexander, Mathie, & Peters, 2011)).
Η GtoPdb έχει σαν στόχο:

- Την παροχή πρόσβασης σε δεδομένα σχετικά με όλους τους γνωστούς βιολογικούς στόχους καθώς και με τους υποδοχείς/κανάλια ιόντων
- Να προτείνει προσδέτες για χαρακτηρισμό των εν λόγω στόχων
- Την παροχή ενός σημείου εισόδου στην βιβλιογραφία της φαρμακολογίας
- Την παροχή μιας ολοκληρωμένης πηγής εκπαίδευσης με υψηλή ποιότητα εξάσκησης στις αρχές της βασικής και κλινικής φαρμακολογίας καθώς και τις τεχνικές της
- Την προώθηση καινοτόμων φαρμακευτικών ανακαλύψεων

Μερικά προβλήματα στην τρέχουσα ανακάλυψη φαρμάκων αφορούν τον αριθμό των μεταβλητών που εμπλέκονται στις αλληλεπιδράσεις φαρμάκου-υποδοχέα, τις νέες περιοχές ncRNAs (σε εξέλιξη, με HGNC), στην επιγενετική (Tough, Lewis, Rioja, Lindon, & Prinjha, 2014), την εναλλακτική συρραφή (Bonner, 2014), το allostery (Christopoulos et al., 2014), και τις ανοσολογικές αντιδράσεις (σε εξέλιξη), που συμβάλλουν σημαντικά στις διεργασίες μιας νόσου (Spedding, 2011). Για τον λόγο αυτό η συμβουλή εξειδικευμένων επιστημόνων μπορεί να βοηθήσει σημαντικά στην επίλυση δυσκολιών που προκύπτουν.

Η GtoPdb περιλαμβάνει σήμερα πάνω από 2.700 επιβεβαιωμένους ή πιθανούς στόχους φαρμάκων και σχετικών πρωτεϊνών (υποδοχείς συζευγμένοι με G-πρωτεΐνη συμπεριλαμβανομένων των ορφανών GPCRs, κανάλια ιόντων, υποδοχείς πυρηνικών ορμονών, καταλυτικοί υποδοχείς, κινάσες, πρωτεάσες, μεταφορείς κλπ), μαζί με προσδέτες, οι οποίοι είναι είτε φάρμακα που διατίθενται ήδη στην αγορά ή πιθανά φάρμακα για ανάπτυξη, ή τα καλύτερα διαθέσιμα πειραματικά εργαλεία για την αξιολόγηση των εν λόγω στόχων. Οι περαιτέρω στόχοι περιλαμβάνουν τη σύντομη εισαγωγή στην φαρμακολογία, ανάγνωση υποβάθρου, και δημοσιευμένα δεδομένα σχετικά με τη συγγένεια των προσδετών και των στόχων τους. Για τα υποσύνολα πολύ σημαντικών στόχων παρέχονται λεπτομερείς σχολιασμοί όπως λειτουργία, φυσιολογία, και βιολογικές ή κλινικές σχετικές παραλλαγές.

Η GtoPdb μπορεί να μην έχει την έκταση της ChEMBL (Gaulton et al., 2012), ωστόσο συμπληρώνει τις προσεγγίσεις μεγάλης κλίμακας με το να είναι μια εστιασμένη βάση δεδομένων σε προσεκτικά επιλεγμένους προσδέτες και στόχους και που ο σχολιασμός της γίνεται από εξειδικευμένους επιστήμονες. Παρέχει βασικές πληροφορίες και σχόλια που προστίθενται στο γενικό πλαίσιο. Επιπλέον, παρέχονται

σύνδεσμοι σε αντίστοιχες εγγραφές σε άλλες πηγές, π.χ. UniProt, Ensembl, Entrez Gene, KEGG, OMIM. Οι πληροφορίες σχετικά με τους προσδέτες περιλαμβάνουν χημικές δομές, αλληλουχίες πεπτιδίων, κλινικά δεδομένα και ονοματολογία, που συνδέονται με βασικές πηγές, συμπεριλαμβανομένων των PubChem, DrugBank και ChEMBL. Τέλος, υπάρχει ένας δημοσιευμένος οδηγός φαρμακολογίας 'Concise Guide to PHARMACOLOGY', που δημιουργήθηκε στην GtoPdb από περιλήψεις οικογενειών στόχων, και χρησιμεύει ως οδηγός γρήγορης αναφοράς. Δημοσιεύεται ανά διετία στο British Journal of Pharmacology και αντικαθιστά την GRAC (Alexander et al., 2013).

Kinase.com: Η Kinase.com διερευνά τις λειτουργίες και την εξέλιξη των πρωτεϊνικών κινασών, οι οποίες αποτελούν βασικούς ρυθμιστές των περισσότερων βιοχημικών μονοπατιών και είναι ιδιαίτερα σημαντικές για την υγεία και τις ασθένειες (Manning, Whyte, Martinez, Hunter, & Sudarsanam, 2002). Επικεντρώνεται στο "kinome", δηλαδή στο σύνολο των κινασών σε σε ένα γονιδίωμα. Η ιστοσελίδα της βάσης, KinBase, είναι διαδραστική και περιλαμβάνει πληροφορίες για πάνω από 7.000 γονίδια πρωτεϊνικών κινασών που βρίσκονται στο ανθρώπινο γονιδίωμα, καθώς επίσης 14 επιπλέον γονιδιώματα (Bradham et al., 2006; Caenepeel, Charydczak, Sudarsanam, Hunter, & Manning, 2004; Eisen et al., 2006; Goldberg et al., 2006; Srivastava et al., 2010; Stajich et al., 2010). Οι κινάσες κατατάσσονται ιεραρχικά σε 10 ομάδες, 287 οικογένειες και 356 υποοικογένειες. Η αναζήτηση στην KinBase μπορεί να γίνει βάσει του ονόματος των γονιδίων, των συμπληρωματικών δομικών μοτίβων, ή σύμφωνα με την ταξινόμηση. Επιπλέον, η ιστοσελίδα παρέχει την υπηρεσία BLAST ώστε η αναζήτηση των κινασών να μπορεί να πραγματοποιηθεί με βάση την ομοιότητα αλληλουχίας.

Κάθε κινάση έχει τη δική της σελίδα που περιέχει την ταξινόμηση, την αλληλουχία, τον σχολιασμό της από εξωτερικές πηγές, το γράφημα του συδιασμού των δομικών μοτίβων, και συγκεκριμένα, τη σύγκριση με το αντίστοιχο προφίλ HMM των ομάδων κινάσης, των οικογενειών και των υπεροικογενειών. Κάθε κατηγορία κινάσης (ομάδα, οικογένεια και υποοικογένεια) έχει τη δική της σελίδα που περιέχει την στοίχηση αλληλουχίας, το προφίλ HMM και το φυλογενετικό δέντρο των πρωτεϊνικών αλληλουχιών και των μοτίβων κινασών. Εκτός από την ιστοσελίδα KinBase, η Kinase.com περιέχει ένα δημόσιο σύστημα wiki, το WiKinome, που εστιάζει στην εξέλιξη και τη λειτουργία των κινασών. Ο απώτερος στόχος είναι η δημιουργία μίας σελίδας wiki για κάθε οικογένεια και υπο-οικογένεια κινάσης.

Η βάση Kinase.com δημιουργήθηκε το 1999 για την υποστήριξη της δημοσιευμένης ανάλυσης της εταιρείας σχεδιασμού φαρμάκων Sugen σχετικά με τις κινάσες του Caenorhabditis elegans (Bingham, Plowman, & Sudarsanam, 2000; Manning, 2005; Plowman, Sudarsanam, Bingham, Whyte, & Hunter, 1999) και του Saccharomyces cerevisiae (Hunter & Plowman, 1997). Η βάση δεδομένων KinBase δημιουργήθηκε το 2002 για την υποστήριξη των περαιτέρω εργασιών ανθρώπινων πρωτεϊνικών κινασών (Manning, Whyte, et al., 2002) και των μυγών των φρούτων (Manning, Plowman, Hunter, & Sudarsanam, 2002). Έχει αναπτυχθεί χρησιμοποιώντας βάση δεδομένων MySQL και τη γλώσσα προγραμματισμού Perl. Η ιστοσελίδα της έχει ανανεωθεί πρόσφατα με τη χρήση σύγχρονων τεχνολογιών ανάπτυξης ιστοσελίδων συμπεριλαμβανομένων των Model–view–controller web framework, HTML5, CSS5 και JavaScript.

Βασική δυσκολία αποτελεί η ενημέρωση της βάσης δεδομένων με στοιχεία υψηλής ποιότητας. Παρά το γεγονός ότι έχουν αλληλουχηθεί πάνω από 6.000 ευκαρυωτικά γονιδιώματα, η Kinase.com περιλαμβάνει τα kinomes μόνο των 15 γονιδιωμάτων. Έχει γίνει προσπάθεια αυτόματης εύρεσης και καταχώρησης πρωτεϊνικών κινασών για όλα τα γονιδιωμάτα, αλλά δεν έχει επιτευχθεί η ίδια ποιότητα, από πλευράς μοντέλου γονιδίου ή/και ταξινόμησης, όπως για τα kinomes των 15 γονιδιωμάτων που αναφέρθηκαν. Χωρίς την χειροκίνητη επιμέλεια και άλλων ερευνητών, είναι δύσκολο να πραγματοποιούνται τακτικές και συχνές ενημερώσεις υψηλής ποιότητας των kinomes.

Structure-Function Linkage Database: Η βάση δεδομένων Structure-Function Linkage Database (SFLD; http://sfld.rbvi.ucsf.edu/django/), (Akiva et al., 2014; Pegg et al., 2006), παρέχει την ιεραρχική ταξινόμηση των λειτουργικά διαφορετικών υπερ-οικογενειών των ενζύμων και συνδέει τις αλληλουχίες και τα δομικά χαρακτηριστικά για κάθε ένζυμο. Δημιουργήθηκε για να διευκολύνει την εννοιολογική κατανόηση του τρόπου που εκπροσωπούνται οι διάφορες αντιδράσεις σε υπερ-οικογένειες που εξελίσσονται (Gerlt & Babbitt, 2001). Η SFLD είναι η μοναδική από τις πηγές που σχολιάζουν πρωτεΐνες η οποία χρησιμοποιεί ως βάση για την συσχέτιση των αλληλουχιών, της δομής και των καταλυτικών χαρακτηριστικών το "χημικά - περιορισμένο" μοντέλο (Babbitt & Gerlt, 1997).

Τέτοιες υπερ-οικογένειες συναντώνται στη φύση και εκτιμάται ότι αντιπροσωπεύουν τουλάχιστον το ένα τρίτο του συνόλου των υπερ-οικογενειών των ενζύμων (Almonacid & Babbitt, 2011). Όλα τα μέλη της

υπερ-οικογένειας εμφανίζουν συντήρηση των λειτουργικά σημαντικών καταλοίπων του ενεργού κέντρου, ενώ τα υποστρώματα τους, τα προϊόντα και ακόμη και οι συνολικές αντιδράσεις μπορεί να είναι ουσιωδώς διαφορετικά. Στο ανώτερο επίπεδο της ιεραρχίας (επίπεδο υπερ-οικογένειας), η SFLD συνδέει αυτά τα συντηρημένα μοτίβα του ενεργού κέντρου με χημικά χαρακτηριστικά που είναι κοινά σε όλα τα μέλη. Για παράδειγμα, στην περίπτωση της ιεραρχίας της υπερ-οικογένειας ενολάσης, όλα τα ένζυμα που χαρακτηρίζονται μέλη της υπερ-οικογένειας έχουν κοινή μια παρόμοια αρχιτεκτονική ενεργού κέντρου που συνδέεται με μια συγκεκριμένη μερική αντίδραση, την αφαίρεση ενός πρωτονίου προς ένα υπόστρωμα καρβοξυλικού άλατος, και τον σχηματισμό ενός κοινού τύπου ενδιαμέσου ενολικού ανιόντος (Babbitt et al., 1996; Gerlt, Babbitt, & Rayment, 2005).

Η βάση δεδομένων παρέχει υψηλής ποιότητας επιμέλεια των λειτουργικών ιδιοτήτων σε επίπεδο υπερ-οικογένειας, υποομάδας και οικογένειας για ένα μικρό σύνολο μεγάλων και ποικίλων υπερ-οικογενειών (Core SFLD), μαζί με πολλούς τύπους σχετικών μετα-δεδομένων και αποτελεσμάτων ανάλυσης. Τα δεδομένα και οι πληροφορίες για κάθε μία από τις υπερ-οικογένειες, υποομάδες, και τα επίπεδα οικογενειών διατίθενται ελεύθερα μέσω ενός εξελιγμένου γραφικού περιβάλλοντος διεπαφής του χρήστη (user interface). Το διαθέσιμο υλικό περιλαμβάνει αρχείο της υπερ-οικογένειας των αλληλουχιών, των HMMs, των σχολιασμένων πολλαπλών στοιχίσεων, απεικονίσεων των 3D δομών και σχολιασμένων ενεργών κέντρων που μπορεί να επεξεργαστεί χρησιμοποιώντας το ελεύθερα διαθέσιμο λογισμικό Chimera (Pettersen et al., 2004), καθώς και συνδέσμους με πολλές άλλες σχετικές πηγές. Η ενότητα Extended SFLD παρέχει λιγότερο επιμελημένες πληροφορίες για ένα μεγαλύτερο σύνολο λειτουργικά διαφορετικών υπεροικογενειών ενζύμων.

Για τα δεδομένα που διατίθενται στους χρήστες, η SFLD χρησιμοποιεί το Django, το οποίο είναι ένα πλαίσιο υψηλού επιπέδου του Python Web, για να δημιουργήσει το διαδικυακό περιβάλλον. Η χρήση αυτού του πλαισίου διευκολύνει σημαντικά την ανάπτυξη της διεπαφής που αλληλεπιδρούν οι χρήστες, καθώς και την καταχώρηση δεδομένων από τους επιμελητές μέσω ενός εξελιγμένου γραφικού περιβάλλοντος χρήστη (GUI), το οποίο επιτρέπει επίσης τον έλεγχο λαθών κατά την καταχώρηση των δεδομένων.

Καθώς τα δεδομένα πρωτεϊνικών αλληλουχιών συνεχίζουν να αυξάνονται με εκθετικό ρυθμό, έχουν αυξηθεί και τα μέλη των υπερ-οικογενειών που σε ορισμένες περιπτώσεις έχουν ξεπεράσει τις 100.000 αλληλουχίες. Για την αντιμετώπιση αυτής της πρόκλησης και την παροχή υποστήριξης για την εφαρμογή του ιεραρχικού μοντέλου SFLD σε αυτές τις μεγάλες και διαφορετικές υπερ-οικογένειες του Core και Extended SFLD, χρησιμοποιούνται τα δίκτυα ομοιότητας πρωτεϊνών (Atkinson, Morris, Ferrin, & Babbitt, 2009).

Histone Database: Η βάση Histone (http://research.nhgri.nih.gov/histones/) ιδρύθηκε το 1996 (Baxevanis & Landsman, 1996), ως αποτέλεσμα έρευνας αναφορικά με το δίπλωμα των ιστόνων, πρωτεϊνών που προσδένουν το DNA (Baxevanis, Arents, Moudrianakis, & Landsman, 1995). Όλα αυτά τα χρόνια έχουν χρησιμοποιηθεί διάφορα εργαλεία για τον εντοπισμό των ιστόνων στις βάσεις δεδομένων αλληλουχιών, συμπεριλαμβανομένων των πρόσφατων εκδόσεων των PSI-BLAST (Altschul et al., 1997) και HMMER (Eddy, 2009). Μετά την ταυτοποίηση, τα εργαλεία αυτά χρησιμοποιούνται για τον έλεγχο της στοίχισης των αλληλουχιών και των εσφαλμένων καταχωρήσεων στις βάσεις δεδομένων. Τα περισσότερα από αυτά τα λάθη σχετίζονται με την εσφαλμένη τοποθέτηση των κωδικονίων έναρξης. Δεδομένου ότι οι ιστόνες είναι άκρως συντηρημένες και ότι το δίπλωμα των ιστόνων περιγράφεται και σχολιάζεται επαρκώς, είναι αρκετά εύκολο να ταυτοποιηθούν οι ιστόνες που είναι εσφαλμένα σχολιασμένες στις δημόσιες βάσεις δεδομένων. Οι στοιχίσεις της κάθε οικογένειας (H1, H2A, H2B, H3, H4) είναι διαθέσιμες για μεταφόρτωση, ενώ υπάρχει επίσης μία σελίδα αναζήτησης για την εξαγωγή μόνο των αλληλουχιών για τις οποίες ενδιαφέρεται ο χρήστης. Οι στοιχίσεις αυτές έχουν χρησιμοποιηθεί για την ταυτοποίηση τέτοιων πρωτεϊνών (Baxevanis, et al., 1995), καθώς και την πρόσφατη προσθήκη της οικογένειας των ιστονών των Αρχαιοβακτηρίων (Marino-Ramirez et al., 2011). Αυτές οι τελευταίες συλλογές πρωτεϊνών ταξινομούνται στη βάση Histone Database ως ξεχωριστές οικογένειες. Η Histone Database περιέχει επίσης λίστες προσδιορισμένων τρισδιάστατων δομών ιστονών που εξάγονται από την Protein Data Bank (PDB) (Rose et al., 2014), ως επί το πλείστον με τη μορφή του νουκλεοσωματικών δομών που προσδιορίζονται με κρυσταλλογραφία ακτίνων-Χ.

Οι μελλοντικές προκλήσεις περιλαμβάνουν την ενημέρωση της βάσης με νέες αλληλουχίες από τις δημόσιες βάσεις δεδομένων και την υποδιαίρεση κάθε οικογένειας των ιστονών σε διάφορες υπο-οικογένειες ή υπο-τύπους (π.χ.κεντρομερικές ιστόνες (CENP-A και CSE4), ιστόνες H3.3, H2A.B , H2A.Z, H2B.Z, macroH2A (136) και, ενδεχομένως, διάφορους υπο-τύπους ιστόνης H1).

Άλλες εξειδικευμένες βάσεις δεδομένων

Παρόλο που, όπως αναφέραμε ήδη, οι περισσότερες εξειδικευμένες βάσεις αφορούν πρωτεϊνικές αλληλουχίες, υπάρχουν δεκάδες άλλες δημόσια διαθέσιμες βάσεις δεδομένων που είναι δυνατόν να ανήκουν σε οποιαδήποτε από τις κατηγορίες των πρωτογενών βάσεων που αναφέρθηκαν παραπάνω.

Κάποιες από αυτές μπορεί να είναι εξειδικευμένες με την έννοια ότι συλλέγουν όλη τη διαθέσιμη πληροφορία για το γονιδίωμα ενός οργανισμού και τις πρωτεΐνες που αυτό κωδικοποιεί (όμοια με την nextProt που αναφέραμε παραπάνω). Τέτοια παραδείγματα είναι η **SubtiList** (http://genolist.pasteur.fr/SubtiList/) (Moszer, Jones, Moreira, Fabry, & Danchin, 2002) για τον *Bacillus subtilis* και η **EcoCyc** (http://ecocyc.org/) για την *Escherichia coli* K-12 (Karp et al., 2002). Παρόμοιες βάσεις υπάρχουν και για άλλους οργανισμούς, ενώ η **Genome Online Database** (GOLD), περιέχει κατάλογο με όλους τους οργανισμούς με πλήρως προσδιορισμένο γονιδίωμα (https://gold.jgi-psf.org/) (Reddy et al., 2015).

Για τα δεδομένα γονιδιακής έκφρασης, υπάρχουν επίσης αρκετές εξειδικευμένες βάσεις δεδομένων. Για παράδειγμα η **ONCOMINE,** είναι βάση δεδομένων που περιέχει πειράματα μικροσυστοιχιών που αφορούν διαφόρους τύπους καρκίνου. Επίσης παρέχει στο χρήστη εργαλεία διαχείρισης των δεδομένων για την αποδοτικότερη εύρεση των επιθυμητών πειραμάτων και γονιδίων, http://www.oncomine.org/, (Rhodes et al., 2004). Το **RNA-Seq Atlas** είναι μια δημόσια βάση δεδομένων για δεδομένα από αλληλούχιση RNA (RNA-Seq). Περιέχει δεδομένα έκφρασης για 11 διαφορετικούς ιστούς από υγιείς ανθρώπους. Η βάση περιέχει επίσης εργαλεία για τη σύγκριση μεταξύ των ιστών καθώς και για την εύρεση γονιδίων με εντοπισμένη έκφραση σε κάποιον ιστό (Krupp et al., 2012). Το **Next Generation Sequencing Catalog (NGS Catalog)** είναι μια δημόσια βάση δεδομένων για τη συλλογή δεδομένων έκφρασης από μελέτες **Next Generation Sequencing σε ανθρώπους** και βασίζεται σε συλλογή δεδομένων από τη βιβλιογραφία. Η βάση περιέχει βιβλιογραφικά δεδομένα, βιολογικές πληροφορίες όπως πληροφορίες για την ασθένεια ή τον πληθυσμό και τεχνικές λεπτομέρειες για τη διαδικασία αλληλούχισης (Xia et al., 2012).

Ειδική αναφορά αξίζει στις όλο και περισσότερο αναπτυσσόμενες τα τελευταία χρόνια, βάσεις δεδομένων γενετικής συσχέτισης. Οι βάσεις αυτές περιέχουν πληροφορίες που εμπλέκουν γονίδια και παραλλαγές των γονιδίων (τους πολυμορφισμούς δηλαδή) με ασθένειες. Παραδοσιακά, υπήρχε η **OMIM (Online Mendelian Inheritance in Man),** η οποία περιέχει κυρίως πληροφορίες για νοσήματα μονογονιδιακής αιτιολογίας. Τα τελευταία χρόνια όμως, με την ανάπτυξη της γενετικής επιδημιολογίας και των μελετών γενετικής συσχέτισης, έχουν αρχίσει να αναπτύσσονται και οι αντίστοιχες βάσεις δεδομένων, οι οποίες βασίζονται κυρίως σε ανάλυση των δημοσιευμένων εργασιών. Το πρώτο παράδειγμα ήταν η **GAD** (Genetic Association Database, http://geneticassociationdb.nih.gov/) η οποία συνέλλεγε όλες τις σχετικές δημοσιεύσεις από την PubMed αλλά πλέον σταμάτησε τη λειτουργία της (Becker, Barnes, Bright, & Wang, 2004), ενώ το **Catalog of Published Genome-Wide Association Studies (**http://www.genome.gov/gwastudies/) και η **GWASdb** (http://jjwanglab.org/gwasdb) επικεντρώνονται στις ευρυγονιδιωματικές μελέτες (genomewide association studies), οι οποίες στηρίζονται σε μια τεχνολογία υψηλής απόδοσης ανάλογης με αυτήν των μικροσυστοιχιών DNA. Επιπλέον δε, έχουν αναπτυχθεί και μικρότερες βάσεις δεδομένων, οι οποίες συλλέγουν και επεξεργάζονται δεδομένα ειδικά για μια συγκεκριμένη ασθένεια, όπως για παράδειγμα η **Epilepsy Genetic Association Database (epiGAD)** για την επιληψία (Tan & Berkovic, 2010), η **Cancer GAMAdb** (Schully et al., 2011) για τον καρκίνο, ή η **AlzGene** για τη νόσο Alzheimer (Bertram, McQueen, Mullin, Blacker, & Tanzi, 2007).

Τέλος, παρόλο που στην ενότητα για τις εξειδικευμένες βάσεις δεδομένων που συμμετείχαν στο δίκτυο SPRN αναφέρθηκαν και περιγράφηκαν μια σειρά από τέτοιες βάσεις, είναι προφανές πως υπάρχουν δεκάδες άλλες βάσεις που περιέχουν σημαντικά δεδομένα και αξίζουν περιγραφή. Ένα καλό σημείο αναφοράς, είναι η συλλογή Database Collection του περιοδικού Nucleic Acids Research, στο οποίο κάθε χρόνο σε ειδικό τεύχος δημοσιεύονται άρθρα που περιγράφουν βάσεις βιολογικών δεδομένων (http://www.oxfordjournals.org/our_journals/nar/database/cap/). Μερικές τέτοιες βάσεις που αξίζουν ειδικής αναφοράς είναι η **PDBTM** που περιέχει τις τοπολογίες από τις τρισδιάστατες δομές διαμεβρανικών πρωτεϊνών, http://pdbtm.enzim.hu/ (Kozma, Simon, & Tusnady, 2013), η **ExTopoDB** η οποία περιέχει πειραματικά δεδομένα για την τοπολογία μεμβρανικών πρωτεϊνών με όχι γνωστή τρισδιάστατη δομή, http://bioinformatics.biol.uoa.gr/ExTopoDB (Tsaousis et al., 2010), και η **gpDB** η οποία περιέχει δεδομένα για τους GPCRs και τις αλληλεπιδράσεις τους με τις G-πρωτεΐνες http://bioinformatics.biol.uoa.gr/gpDB (Theodoropoulou, Bagos, Spyropoulos, & Hamodrakas, 2008). Η **DBPTM** είναι μια βάση δεδομένων που συλλέγει διαφόρων τύπων δεδομένα για μετα-μεταφραστικές τροποποιήσεις πρωτεϊνών http://dbptm.mbc.nctu.edu.tw/ (Lu et al., 2013), ενώ η **DIP** περιέχει δεδομένα, προερχόμενα με διαφορετικούς τρόπους, για αλληλεπιδράσεις πρωτεϊνών-πρωτεϊνών http://dip.doe-mbi.ucla.edu/dip/Main.cgi (Xenarios et

al., 2002). Τέλος, η **bioGrid** περιέχει γενικά δεδομένα και εργαλεία ανάλυσης για βιολογικές αλληλεπιδράσεις http://thebiogrid.org/ (Stark et al., 2006).

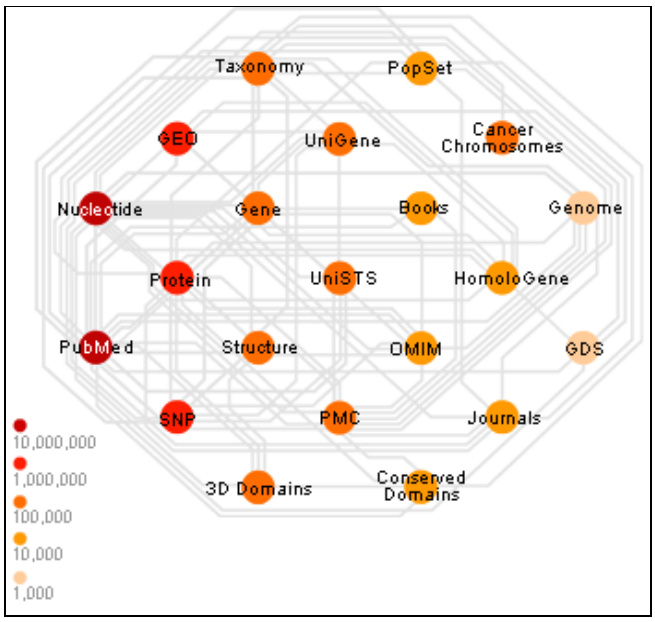
Όσον αφορά τα νουκλεϊκά οξέα (DNA και RNA), τα πράγματα είναι επίσης παρόμοια. Υπάρχουν δεκάδες διαθέσιμες εξειδικευμένες βάσεις δεδομένων και αξίζει εδώ να αναφέρουμε τουλάχιστον αυτές που περιέχουν miRNA και στόχους αυτών, όπως η **MiRBase** http://www.mirbase.org/ (Griffiths-Jones, Grocock, van Dongen, Bateman, & Enright, 2006), η **MirTarBase** http://mirtarbase.mbc.nctu.edu.tw/ (Hsu et al., 2011) και η **TarBase** http://diana.imis.athena-innovation.gr/DianaTools/index.php?r=tarbase/index/ (Sethupathy, Corda, & Hatzigeorgiou, 2006). Άλλες σημαντικές βάσεις δεδομενων, είναι αυτές που περιέχουν δεδομένα για εσώνια-εξώνια όπως η **EID** http://bpg.utoledo.edu/~afedorov/lab/eid.html (Shepelev & Fedorov, 2006), αλλά και αυτές που ασχολούνται με τους υποκινητές των γονιδίων όπως η **EPD** http://epd.vital-it.ch/ (Dreos, Ambrosini, Cavin Perier, & Bucher, 2013) και η **MMPROMdb** http://mpromdb.wistar.upenn.edu/ (Sun et al., 2006). Φυσικά, η λίστα δεν τελειώνει εδώ, και για εξειδικευμένες αναζητήσεις, οι χρήστες θα πρέπει να παρακολουθούν τη βιβλιογραφία και να ενημερώνονται για δημοσιεύσεις που περιγράφουν νέες βάσεις δεδομένων.

2.3 Ολοκληρωμένα συστήματα ανάκτησης πληροφοριών από βάσεις δεδομένων.

Το **SRS**, είναι ένα ειδικό λογισμικό που διατίθεται από την εταιρία LION Bioscience και αποτελεί ένα ισχυρό και εύχρηστο σύστημα διαχείρισης βιολογικών δεδομένων. Είναι μεν εμπορικό λογισμικό, αλλά διατίθεται δωρεάν για ακαδημαϊκή χρήση. Παρέχει την δυνατότητα αναζήτησης και ανάκτησης δεδομένων σε ένα φιλικό προς τον χρήστη γραφικό περιβάλλον και σε περισσότερες από 400 βάσεις δεδομένων οι οποίες μπορεί να είναι αποθηκευμένες στον ίδιο κεντρικό υπολογιστή. Το βασικό πλεονέκτημα του SRS είναι η δυνατότητα ταυτόχρονης αναζήτησης πληροφοριών σε περισσότερες από μία βάσεις οι οποίες είναι πιθανό να περιέχουν πληροφορίες διαφορετικού είδους καθώς και η μορφοποίηση των δεδομένων σε καθεμιά να είναι διαφορετική. Επιπλέον, λαμβάνοντας υπόψη τον τεράστιο όγκο πληροφορίας και τον μεγάλο αριθμό βάσεων που μπορεί να διαχειρίζεται ταυτόχρονα, σημαντικό πλεονέκτημα αποτελεί η ταχύτητα με την οποία εκτελούνται οι αναζητήσεις. Τέλος δίνεται η δυνατότητα στον κάτοχο του συστήματος να ενσωματώνει σε αυτό και βάσεις που έχει δημιουργήσει ο ίδιος ή προγράμματα για κάθε είδος υπολογιστική ανάλυση χωρίς να επηρεάζεται η απόδοση του συστήματος. Πάνω στο SRS είχαν χτιστεί παλιότερα οι βάσεις του EBI και άλλων μεγάλων ερευνητικών ινστιτούτων. Παρόλα αυτά, πλέον θεωρείται παροχημένο και οι σύγχρονες βάσεις όπως η Uniprot χρησιμοποιούν ειδικά κατασκευασμένα συστήματα βάσεων δεδομένων για την αποθήκευση του όλο και μεγαλύτερου όγκου των δεδομένων.

Το **Entrez** αποτελεί ένα σύστημα διαχείρισης δεδομένων για την αναζήτηση και ανάκτηση πληροφοριών όλων των βάσεων δεδομένων που περιέχονται στο NCBI (National Center for Biotechnology Information) των ΗΠΑ. Το Entrez είναι ανάλογο του SRS και παρέχει στον χρήστη τη δυνατότητα αναζήτησης σε βάσεις δεδομένων νουκλεοτιδικών και πρωτεϊνικών αλληλουχιών, δομές βιομορίων και γονιδιωμάτων. Επιπλέον, μέσω του ίδιου γραφικού περιβάλλοντος, παρέχει την δυνατότητα αναζήτησης στη βάση βιβλιογραφίας PUBMED καθώς και πιο πολύπλοκες αναζητήσεις ανάμεσα στα στοιχεία τους. Βασικό μειονέκτημα αποτελεί το γεγονός ότι περιορίζεται μόνο στις βάσεις δεδομένων του NCBI και ότι δεν επιτρέπει ιδιαίτερα πολύπλοκες αναζητήσεις. Παρόλα αυτά, αποτελεί για χρόνια τώρα την διεπαφή όλων των βάσεων δεδομένων του NCBI, και επιτρέπει με τον ίδιο απλό τρόπο ο χρήστης να πραγματοποιήσει αναζητήσεις σε τελείως διαφορετικές βάσεις δεδομένων.

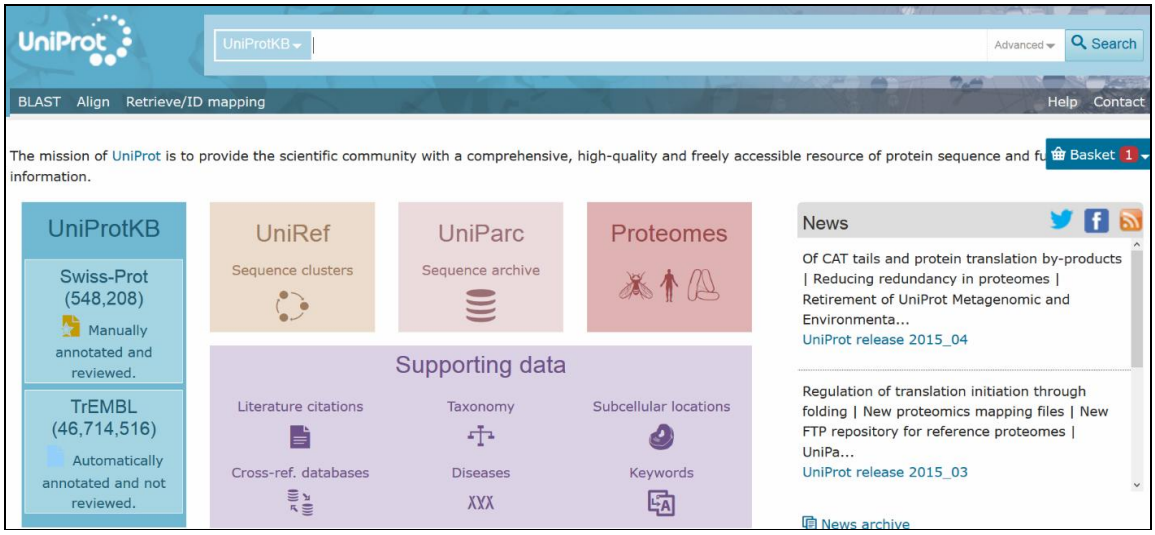
Αξίζει να αναφερθεί, ότι μία από τις διαπιστώσεις της συνάντησης του δικτύου SPRN, όσον αφορά τις εξειδικευμένες βάσεις δεδομένων, ήταν ότι στην συντριπτική τους πλειοψηφία, οι βάσεις αυτές στηρίζονται σε κάποιο γενικό σύστημα βάσης δεδομένων όπως η MySQL σε συνδυασμό με PHP. Όπως αναφέρθηκε, παρόλο που στις περισσότερες περιπτώσεις η ιεραρχία ήταν απλή, και θα αρκούσε και μια απλή ιστοσελίδα, το σύστημα αδιαχείρισης και μόνο (πχ για να γίνονται γρήγορες ανανεώσεις της βάσης ή αντίγραφα ασφαλείας κλπ), ήταν αρκετό για τους ερευνητές για να επιλέξουν αυτόν τον σχεδιασμό. Σε άλλες περιπτώσεις με πιο πολύπλοκη ιεραρχία, η SQL προσδίδει επίσης τα απαραίτητα χαρακηριστικά στους διαχειριστές, και έτσι φαίνεται ότι αυτό το μοντέλο είναι αρκετά διαδεδομένο (αν και δεν υπάρχουν πλήρη δεδομένα για όλες τις μικρές βάσεις που έχουν δημοσιευτεί).



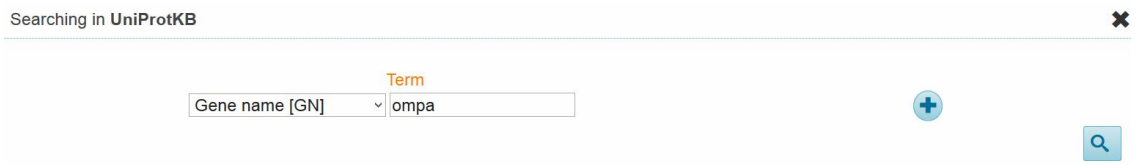
Εικόνα 2.9: *Διαγραμματική απεικόνιση της διασύνδεσης των διαφορετικών βάσεων του NCBI οι οποίες στηρίζονται στο Entrez. Τα διαφορετικά χρώματα, αντιστοιχούν σε διαφορετικό αριθμό καταχωρήσεων. Το NCBI διαθέτει ένα ολοκληρωμένο σύστημα που καλύπτει όλο το εύρος των δημόσιων βάσεων δεδομένων, ακόμα και αυτών που στηρίζονται σε άλλες πηγές. Για παράδειγμα, η Conserved Domains Database είναι αντίστοιχη της PROSITE ενώ η Structure (MMDB) είναι αντίστοιχη της PDB.*

ΠΡΑΚΤΙΚΟ ΜΕΡΟΣ

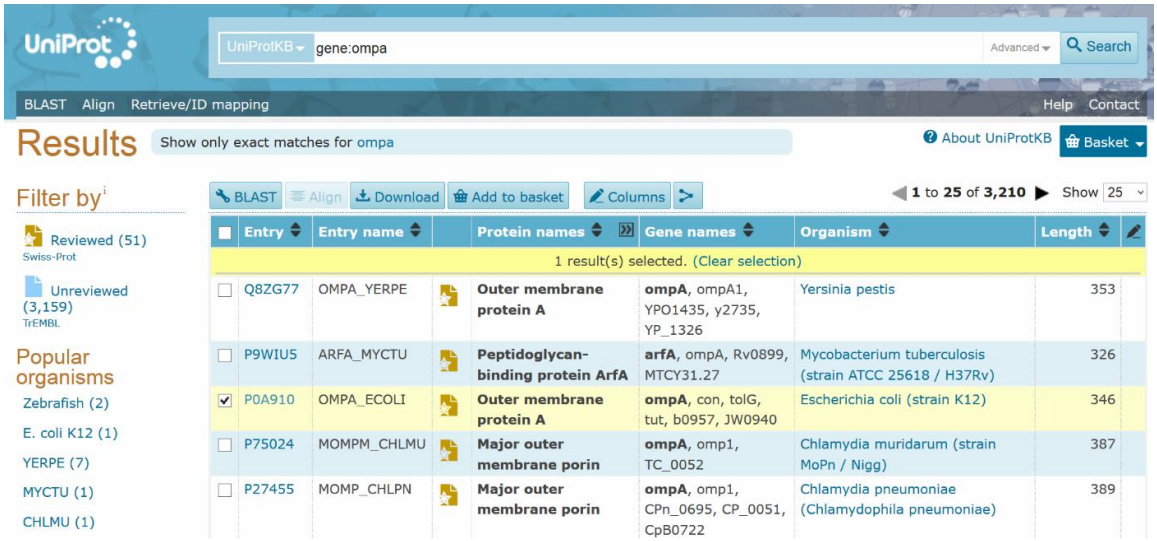
1. Να γίνει αναζήτηση της πρωτεϊνικης αλληλουχίας ompA του βακτηρίου Escherichia coli στη UNIPROT, με βάση το πεδίο Gene Name και να ανακτήσετε την εγγραφή της UNIPROT.



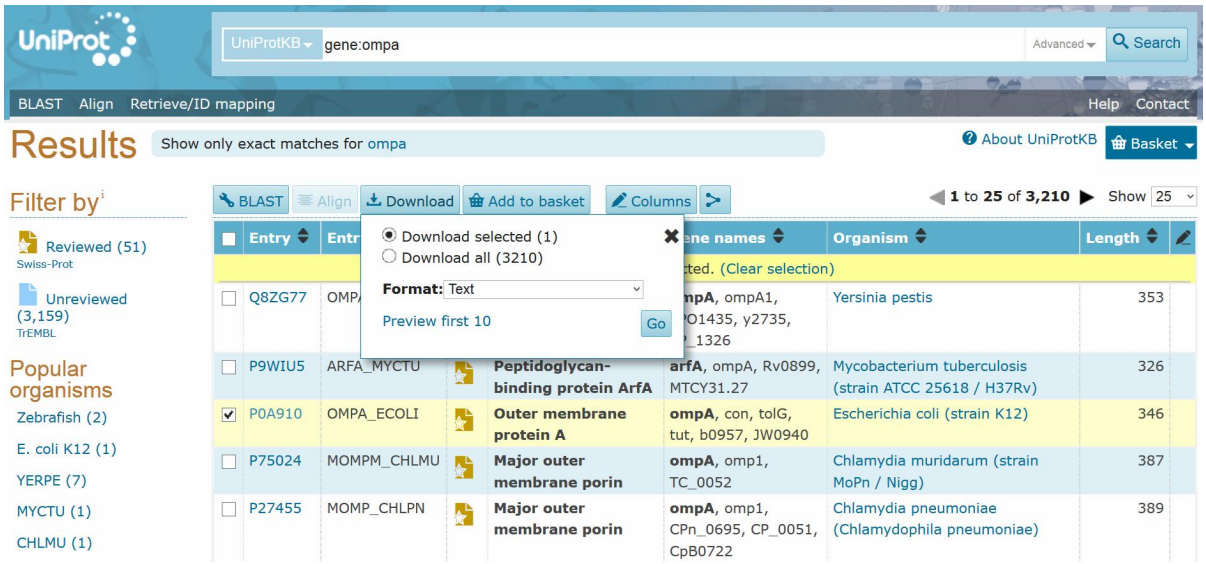
Εικόνα 2.10: *Η αρχική σελίδα της Uniprot*



Εικόνα 2.11: *Επιλέγουμε το όνομα του γονιδίου στο αντίστοιχο πεδίο*

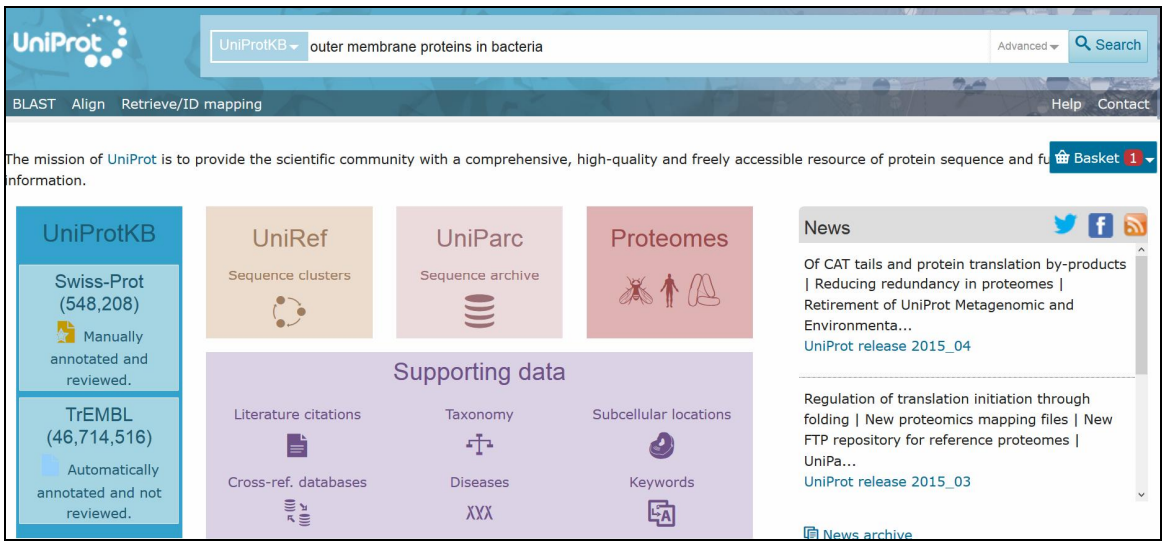


Εικόνα 2.12: *Επιλέγουμε την πρωτεΐνη του οργανισμού που θέλουμε*

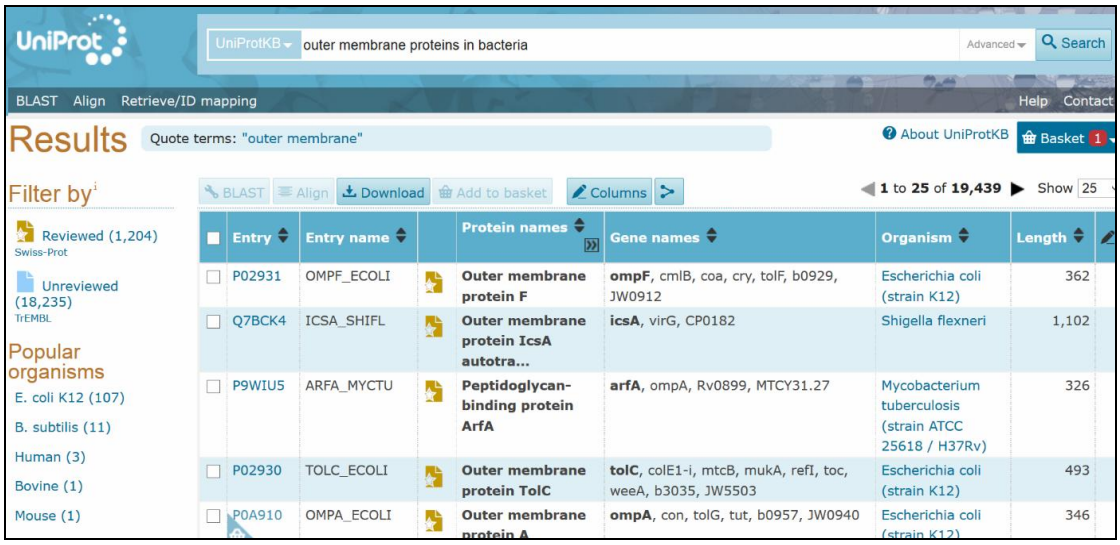


Εικόνα 2.13: *Η επιλογή για να μεταφορτώσουμε όλη την καταχώρηση*

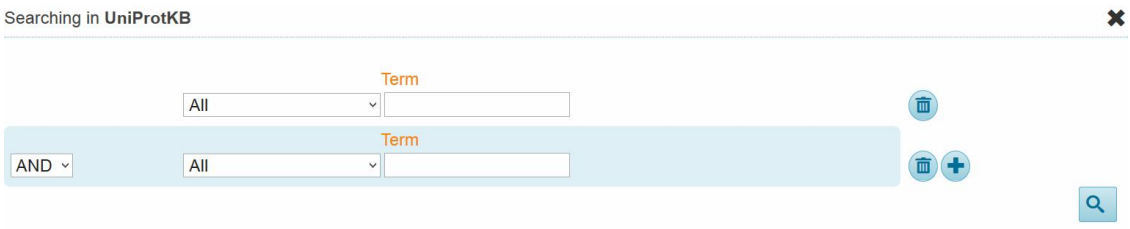
2. Να γίνει αναζήτηση στη βάση δεδομένων Uniprot με σκοπό την ανεύρεση των πρωτεινων της εξωτερικής μεμβράνης (outer membrane) των βακτηρίων με γνωστή προσδιορισμένη δομή.
(η συνολική επερώτηση είναι: taxonomy:"Bacteria [2]" existence:"evidence at protein level" database:(type:pdb) locations:(location:"Cell outer membrane [SL-0040]") keyword:"Cell outer membrane [KW-0998]")



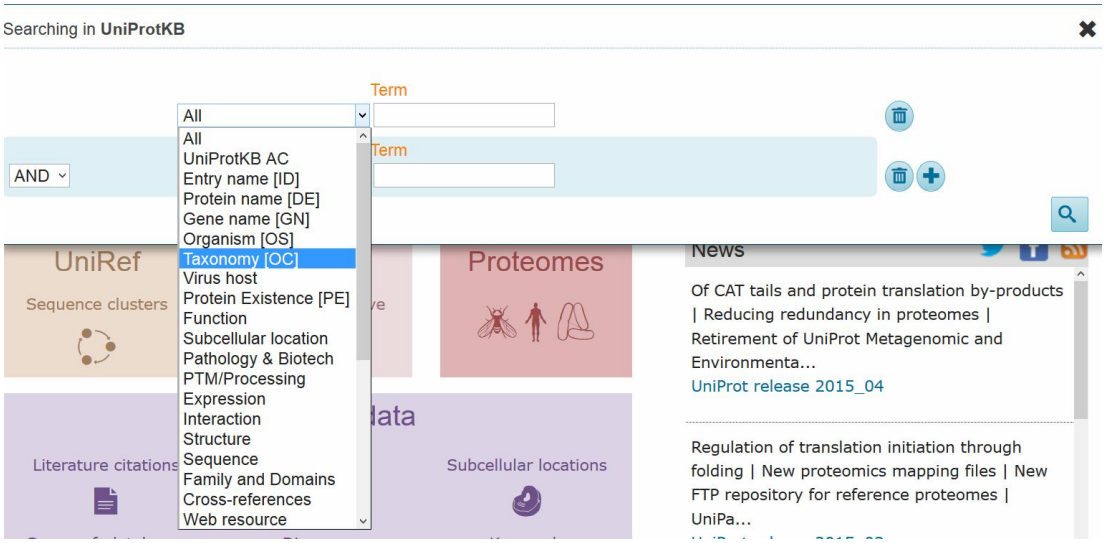
Εικόνα 2.14: *Στην αρχική σελίδα της Uniprot αν κάνουμε μια γενική επερώτηση (σε όλα τα πεδία), θα πάρουμε και πολλές άσχετες απαντήσεις*



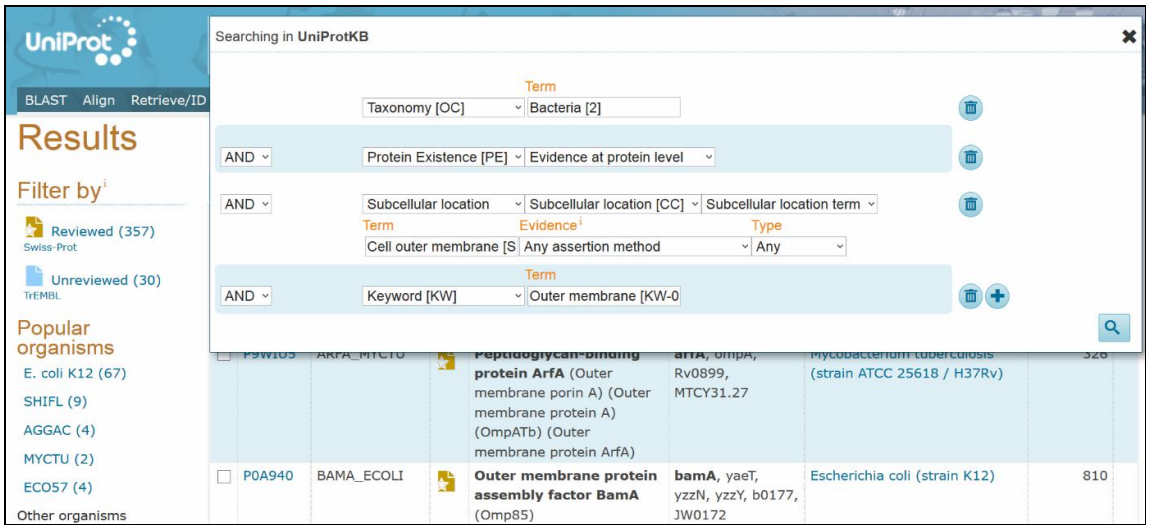
Εικόνα 2.15: *Στην αρχική σελίδα της Uniprot αν κάνουμε μια γενική επερώτηση (σε όλα τα πεδία), θα πάρουμε και πολλές άσχετες απαντήσεις*



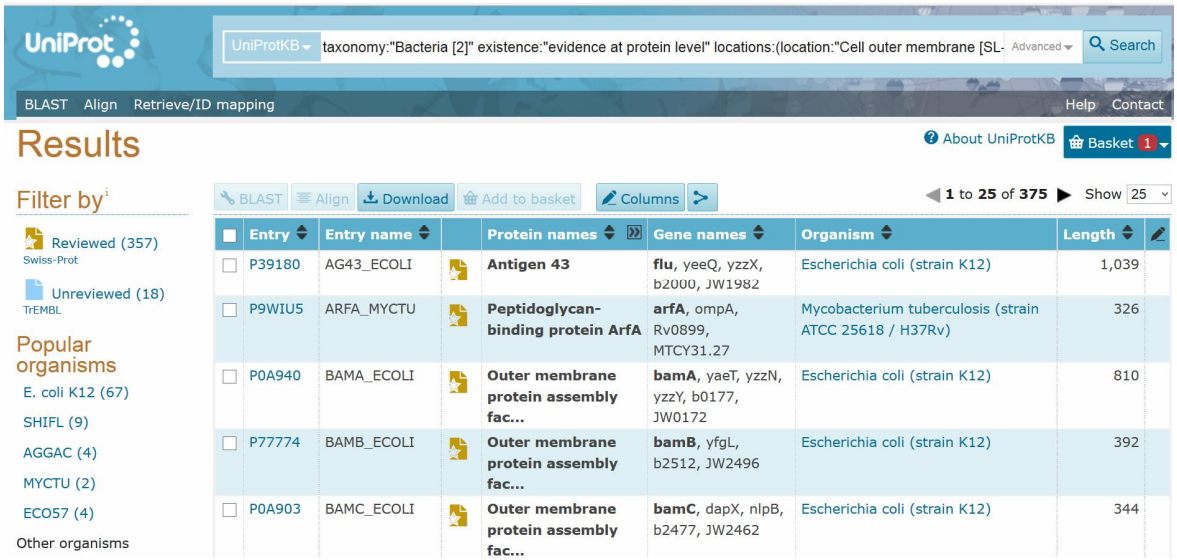
Εικόνα 2.16: *Θα πρέπει να επιλέξουμε να κάνουμε επερωτήσεις για κάθε πεδίο ξεχωριστά*



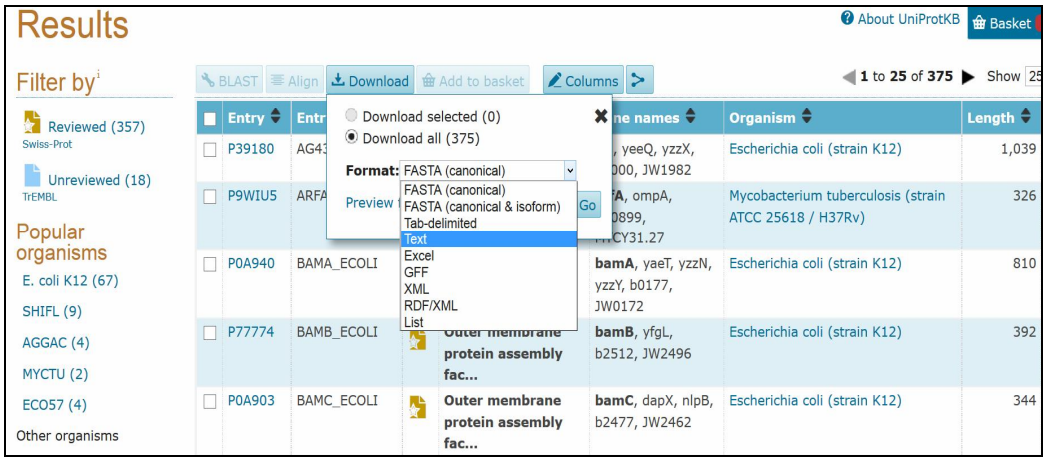
Εικόνα 2.17: *Θα πρέπει να επιλέξουμε να κάνουμε επερωτήσεις για κάθε πεδίο ξεχωριστά. Εδώ, διαλέγουμε την ταξινομική βαθμίδα του οργανισμού*



Εικόνα 2.18: *Θα πρέπει να επιλέξουμε να κάνουμε επερωτήσεις για κάθε πεδίο ξεχωριστά. Εδώ, φαίνονται και τα υπόλοιπα πεδία συμπληρωμένα*



Εικόνα 2.19: *Τα αποτελέσματα της αναζήτησης*



Εικόνα 2.20: *Η επιλογή όλων για μεταφόρτωση σε μορφή κειμένου (υπάρχουν και άλλες επιλογές)*

3. Να γίνει αναζήτηση στη βάση δεδομένων Uniprot με σκοπό την ανεύρεση ανθρώπινων υποδοχέων συζευγμένων με G-πρωτεΐνες οι οποίοι έχουν γνωστή (προσδιορισμένη) τρισδιάστατη δομή:

Η συνολική επερώτηση είναι:
taxonomy:"keyword:"G-protein coupled receptor [KW-0297]" AND organism:"Human [9606]" AND existence:"evidence at protein level" AND database:(type:pdb)

Ποιες από τις πρωτεΐνες έχουν δομή στη βάση δεδομένων PDB; Για τις παραπάνω πρωτεΐνες, να σημειωθεί ποιες αντιστοιχούν στην περιοχή της πρωτεΐνης στην οποία βρίσκονται το σύνολο των διαμεμβρανικών τμημάτων, ποιες αντιστοιχούν σε μέρος της περιοχής των διαμεμβρανικών τμημάτων και ποιες δεν περιλαμβάνουν κανένα τμήμα της αλληλουχίας το οποίο να αντιστοιχεί σε αλληλουχία διαμεμβρανικών τμημάτων.

ΠΑΡΑΡΤΗΜΑ (Παραδείγματα από τις βάσεις δεδομένων)

1. Εγγραφή της GENBANK για το γονίδιο της πρωτεΐνης Outer membrane protein A (ompA) από τον οργανισμό Escherichia coli.

```
LOCUS       NC_000913               1041 bp    DNA     linear   CON 16-DEC-2014

DEFINITION  Escherichia coli str. K-12 substr. MG1655, complete genome.

ACCESSION   NC_000913 REGION: complement(1019013..1020053)
VERSION     NC_000913.3  GI:556503834
DBLINK      BioProject: PRJNA57779
            BioSample: SAMN02604091
KEYWORDS    RefSeq.
SOURCE      Escherichia coli str. K-12 substr. MG1655
  ORGANISM  Escherichia coli str. K-12 substr. MG1655
            Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales;
            Enterobacteriaceae; Escherichia.
REFERENCE   1  (bases 1 to 1041)
  AUTHORS   Riley,M., Abe,T., Arnaud,M.B., Berlyn,M.K., Blattner,F.R.,
            Chaudhuri,R.R., Glasner,J.D., Horiuchi,T., Keseler,I.M., Kosuge,T.,
            Mori,H., Perna,N.T., Plunkett,G. III, Rudd,K.E., Serres,M.H.,
            Thomas,G.H., Thomson,N.R., Wishart,D. and Wanner,B.L.
  TITLE     Escherichia coli K-12: a cooperatively developed annotation
            snapshot--2005
  JOURNAL   Nucleic Acids Res. 34 (1), 1-9 (2006)
   PUBMED   16397293
  REMARK    Publication Status: Online-Only
REFERENCE   2  (bases 1 to 1041)
  AUTHORS   Hayashi,K., Morooka,N., Yamamoto,Y., Fujita,K., Isono,K., Choi,S.,
            Ohtsubo,E., Baba,T., Wanner,B.L., Mori,H. and Horiuchi,T.
  TITLE     Highly accurate genome sequences of Escherichia coli K-12 strains
            MG1655 and W3110
  JOURNAL   Mol. Syst. Biol. 2, 2006 (2006)
   PUBMED   16738553
REFERENCE   3  (bases 1 to 1041)
  AUTHORS   Blattner,F.R., Plunkett,G. III, Bloch,C.A., Perna,N.T., Burland,V.,
            Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F.,
            Gregor,J., Davis,N.W., Kirkpatrick,H.A., Goeden,M.A., Rose,D.J.,
            Mau,B. and Shao,Y.
  TITLE     The complete genome sequence of Escherichia coli K-12
  JOURNAL   Science 277 (5331), 1453-1462 (1997)
   PUBMED   9278503
REFERENCE   4  (bases 1 to 1041)
  AUTHORS   Arnaud,M., Berlyn,M.K.B., Blattner,F.R., Galperin,M.Y.,
            Glasner,J.D., Horiuchi,T., Kosuge,T., Mori,H., Perna,N.T.,
            Plunkett,G. III, Riley,M., Rudd,K.E., Serres,M.H., Thomas,G.H. and
            Wanner,B.L.
  TITLE     Workshop on Annotation of Escherichia coli K-12
  JOURNAL   Unpublished
  REMARK    Woods Hole, Mass., on 14-18 November 2003 (sequence corrections)
REFERENCE   5  (bases 1 to 1041)
  AUTHORS   Glasner,J.D., Perna,N.T., Plunkett,G. III, Anderson,B.D.,
            Bockhorst,J., Hu,J.C., Riley,M., Rudd,K.E. and Serres,M.H.
  TITLE     ASAP: Escherichia coli K-12 strain MG1655 version m56
  JOURNAL   Unpublished
  REMARK    ASAP download 10 June 2004 (annotation updates)
REFERENCE   6  (bases 1 to 1041)
  AUTHORS   Hayashi,K., Morooka,N., Mori,H. and Horiuchi,T.
  TITLE     A more accurate sequence comparison between genomes of Escherichia
```

```
            coli K12 W3110 and MG1655 strains
  JOURNAL   Unpublished
  REMARK    GenBank accessions AG613214 to AG613378 (sequence corrections)
REFERENCE   7  (bases 1 to 1041)
  AUTHORS   Perna,N.T.
  TITLE     Escherichia coli K-12 MG1655 yqiK-rfaE intergenic region, genomic
            sequence correction
  JOURNAL   Unpublished
  REMARK    GenBank accession AY605712 (sequence corrections)
REFERENCE   8  (bases 1 to 1041)
  AUTHORS   Rudd,K.E.
  TITLE     A manual approach to accurate translation start site annotation: an
            E. coli K-12 case study
  JOURNAL   Unpublished
REFERENCE   9  (bases 1 to 1041)
  CONSRTM   NCBI Genome Project
  TITLE     Direct Submission
  JOURNAL   Submitted (26-AUG-2014) National Center for Biotechnology
            Information, NIH, Bethesda, MD 20894, USA
REFERENCE   10 (bases 1 to 1041)
  AUTHORS   Blattner,F.R. and Plunkett,G. III.
  TITLE     Direct Submission
  JOURNAL   Submitted (30-JUL-2014) Laboratory of Genetics, University of
            Wisconsin, 425G Henry Mall, Madison, WI 53706-1580, USA
  REMARK    Protein update by submitter
REFERENCE   11 (bases 1 to 1041)
  AUTHORS   Blattner,F.R. and Plunkett,G. III.
  TITLE     Direct Submission
  JOURNAL   Submitted (15-NOV-2013) Laboratory of Genetics, University of
            Wisconsin, 425G Henry Mall, Madison, WI 53706-1580, USA
  REMARK    Protein update by submitter
REFERENCE   12 (bases 1 to 1041)
  AUTHORS   Blattner,F.R. and Plunkett,G. III.
  TITLE     Direct Submission
  JOURNAL   Submitted (26-SEP-2013) Laboratory of Genetics, University of
            Wisconsin, 425G Henry Mall, Madison, WI 53706-1580, USA
  REMARK    Sequence update by submitter
REFERENCE   13 (bases 1 to 1041)
  AUTHORS   Rudd,K.E.
  TITLE     Direct Submission
  JOURNAL   Submitted (06-FEB-2013) Department of Biochemistry and Molecular
            Biology, University of Miami Miller School of Medicine, 118 Gautier
            Bldg., Miami, FL 33136, USA
  REMARK    Sequence update by submitter
REFERENCE   14 (bases 1 to 1041)
  AUTHORS   Rudd,K.E.
  TITLE     Direct Submission
  JOURNAL   Submitted (24-APR-2007) Department of Biochemistry and Molecular
            Biology, University of Miami Miller School of Medicine, 118 Gautier
            Bldg., Miami, FL 33136, USA
  REMARK    Annotation update from ecogene.org as a multi-database
            collaboration
REFERENCE   15 (bases 1 to 1041)
  AUTHORS   Plunkett,G. III.
  TITLE     Direct Submission
  JOURNAL   Submitted (07-FEB-2006) Laboratory of Genetics, University of
            Wisconsin, 425G Henry Mall, Madison, WI 53706-1580, USA
  REMARK    Protein updates by submitter
REFERENCE   16 (bases 1 to 1041)
  AUTHORS   Plunkett,G. III.
  TITLE     Direct Submission
```

```
  JOURNAL   Submitted (10-JUN-2004) Laboratory of Genetics, University of
            Wisconsin, 425G Henry Mall, Madison, WI 53706-1580, USA
  REMARK    Sequence update by submitter
REFERENCE   17 (bases 1 to 1041)
  AUTHORS   Plunkett,G. III.
  TITLE     Direct Submission
  JOURNAL   Submitted (13-OCT-1998) Laboratory of Genetics, University of
            Wisconsin, 425G Henry Mall, Madison, WI 53706-1580, USA
REFERENCE   18 (bases 1 to 1041)
  AUTHORS   Blattner,F.R. and Plunkett,G. III.
  TITLE     Direct Submission
  JOURNAL   Submitted (02-SEP-1997) Laboratory of Genetics, University of
            Wisconsin, 425G Henry Mall, Madison, WI 53706-1580, USA
REFERENCE   19 (bases 1 to 1041)
  AUTHORS   Blattner,F.R. and Plunkett,G. III.
  TITLE     Direct Submission
  JOURNAL   Submitted (16-JAN-1997) Laboratory of Genetics, University of
            Wisconsin, 425G Henry Mall, Madison, WI 53706-1580, USA
COMMENT     REVIEWED REFSEQ: This record has been curated by NCBI staff. The
            reference sequence is identical to U00096.
            On Nov 3, 2013 this sequence version replaced gi:49175990.
            RefSeq Category: Reference Genome
                        FGS: First Genome sequenced
                        MOD: Model Organism
                        PHY: Based on Phylogenetics
                        UPR: UniProt Genome
            Current U00096 annotation updates are derived from EcoGene
            http://ecogene.org. Suggestions for updates can be sent to Dr.
            Kenneth Rudd (krudd@miami.edu). These updates are being generated
            from a collaboration that also includes ASAP/ERIC, the Coli Genetic
            Stock Center, EcoliHub, EcoCyc, RegulonDB and UniProtKB/Swiss-Prot.
            COMPLETENESS: full length.
FEATURES             Location/Qualifiers
     source          1..1041
                     /organism="Escherichia coli str. K-12 substr. MG1655"
                     /mol_type="genomic DNA"
                     /strain="K-12"
                     /sub_strain="MG1655"
                     /db_xref="taxon:511145"
     gene            1..1041
                     /gene="ompA"
                     /locus_tag="b0957"
                     /gene_synonym="con; ECK0948; JW0940; tolG; tut"
                     /db_xref="EcoGene:EG10669"
                     /db_xref="GeneID:945571"
     CDS             1..1041
                     /gene="ompA"
                     /locus_tag="b0957"
                     /gene_synonym="con; ECK0948; JW0940; tolG; tut"
                     /function="membrane; Outer membrane constituents"
                     /GO_component="GO:0009279 - cell outer membrane;
                     GO:0009274 - peptidoglycan-based cell wall"
                     /note="outer membrane protein 3a (II*;G;d)"
                     /codon_start=1
                     /transl_table=11
                     /product="outer membrane protein A (3a;II*;G;d)"
                     /protein_id="NP_415477.1"
                     /db_xref="GI:16128924"
                     /db_xref="ASAP:ABE-0003240"
                     /db_xref="UniProtKB/Swiss-Prot:P0A910"
                     /db_xref="EcoGene:EG10669"
```

```
                     /db_xref="GeneID:945571"
                     /translation="MKKTAIAIAVALAGFATVAQAAPKDNTWYTGAKLGWSQYHDTGF
                     INNNGPTHENQLGAGAFGGYQVNPYVGFEMGYDWLGRMPYKGSVENGAYKAQGVQLTA
                     KLGYPITDDLDIYTRLGGMVWRADTKSNVYGKNHDTGVSPVFAGGVEYAITPEIATRL
                     EYQWTNNIGDAHTIGTRPDNGMLSLGVSYRFGQGEAAPVVAPAPAPAPEVQTKHFTLK
                     SDVLFNFNKATLKPEGQAALDQLYSQLSNLDPKDGSVVVLGYTDRIGSDAYNQGLSER
                     RAQSVVDYLISKGIPADKISARGMGESNPVTGNTCDNVKQRAALIDCLAPDRRVEIEV
                     KGIKDVVTQPQA"
ORIGIN
        1 atgaaaaaga cagctatcgc gattgcagtg gcactggctg gtttcgctac cgtagcgcag
       61 gccgctccga aagataacac ctggtacact ggtgctaaac tgggctggtc ccagtaccat
      121 gacactggtt tcatcaacaa caatggcccg acccatgaaa accaactggg cgctggtgct
      181 tttggtggtt accaggttaa cccgtatgtt ggctttgaaa tgggttacga ctggttaggt
      241 cgtatgccgt acaaaggcag cgttgaaaac ggtgcataca aagctcaggg cgttcaactg
      301 accgctaaac tgggttaccc aatcactgac gacctggaca tctacactcg tctgggtggc
      361 atggtatggc gtgcagacac taaatccaac gtttatggta aaaaccacga caccggcgtt
      421 tctccggtct tcgctggcgg tgttgagtac gcgatcactc ctgaaatcgc tacccgtctg
      481 gaataccagt ggaccaacaa catcggtgac gcacacacca tcggcactcg tccggacaac
      541 ggcatgctga gcctgggtgt ttcctaccgt ttcggtcagg gcgaagcagc tccagtagtt
      601 gctccggctc cagctccggc accggaagta cagaccaagc acttcactct gaagtctgac
      661 gttctgttca acttcaacaa agcaaccctg aaaccggaag gtcaggctgc tctggatcag
      721 ctgtacagcc agctgagcaa cctggatccg aaagacggtt ccgtagttgt tctgggttac
      781 accgaccgca tcggttctga cgcttacaac cagggtctgt ccgagcgccg tgctcagtct
      841 gttgttgatt acctgatctc caaaggtatc ccggcagaca agatctccgc acgtggtatg
      901 ggcgaatcca acccggttac tggcaacacc tgtgacaacg tgaaacagcg tgctgcactg
      961 atcgactgcc tggctccgga tcgtcgcgta gagatcgaag ttaaaggtat caaagacgtt
     1021 gtaactcagc cgcaggctta a
//
```

Επεξηγήσεις των σημαντικότερων πεδίων μιας εγγραφής στην GENBANK

LOCUS: Περιέχει ένα μικρό όνομα για τον χαρακτηρισμό της εγγραφής.

DEFINITION: Μια λεπτομερής περιγραφή της αλληλουχίας.

ACCESSION: Κωδικός που αποκτά μια νεοεισερχόμενη εγγραφή χαρακτηριστικός για την GENBANK. Ο κωδικός παραμένει σταθερός

VERSION: Ειδικός κωδικός που απαρτίζεται από το πρωταρχικό Accession Number, ακολουθεί το σύμβολο της τελείας και στη συνέχεια ένας αριθμός που δηλώνει την έκδοση της παρούσας εγγραφής.

KEYWORDS: Χαρακτηριστικές λέξεις-κλειδιά που σχετίζονται με την νουκλεοτιδική αλληλουχία και τις ιδιότητες των προϊόντων της.

SOURCE: Βιολογική πηγή της αλληλουχίας όπου αναφέρεται ο οργανισμός από των οποίο έχει απομονωθεί με τα ιδιαίτερα χαρακτηριστικά του (πιθανές μεταλλάξεις, πλασμίδια κ.α.).

ORGANISM: Οργανισμός απ' όπου προήλθε η αλληλουχία. Ακολουθείται η διώνυμη ονομασία κατά Λινναίο. Επίσης παρατίθεται και η συστηματική ταξινόμηση του οργανισμού.

- Τα παρακάτω πεδία σχετίζονται με την δημοσιευμένη εργασία στην οποία αναφέρεται ο προσδιορισμός της παρούσας αλληλουχίας.

REFERENCE: Περιέχει τον αριθμό της αναφοράς καθώς και το μήκος της αλληλουχίας που έχει προσδιοριστεί στην παρούσα εργασία.

AUTHORS: Αναφέρονται οι συμμετέχοντες στην διεξαγωγή της παρούσας εργασίας.

TITLE: Τίτλος της δημοσιευμένης εργασίας.

JOURNAL: Περιέχει λεπτομέρειακα στοιχεία για την αναζήτηση της αναφοράς όπως είναι ο τίτλος του περιοδικού που εκδόθηκε, τεύχος, ημερομηνία έκδοσης και σελίδες που καταλαμβάνει στο συγκεκριμένο τεύχος.

MEDLINE: Κωδικός για την βιβλιογραφική αναφορά στην βάση δεδομένων MEDLINE.

COMMENT: Περιέχει κάποιες γενικές παρατηρήσεις, ή αναφορές και σε άλλες βάσεις.

FEATURES: Πίνακας που περιέχει πληροφορίες σχετικά με τα προϊόντα της αλληλουχίας όπως πολυπεπτιδικές αλυσίδες (από μετάφραση) και RNA (από μεταγραφή) και στοιχεία από πειραματικά δεδομένα που καταδεικνύουν τη βιολογική της σημασία.

BASE COUNT: Αριθμητική ανάλυση της αλληλουχίας στα επιμέρους συστατικά της. Περιέχει το σύνολο καταλοίπων Αδενίνης, Γουανίνης, Κυτοσίνης, Θυμίνης.

ORIGIN: Θέση της πρώτης βάσης της κατατεθειμένης αλληλουχίας σε σχέση με το γονιδίωμα από το οποίο έχει απομονωθεί.

Ακριβώς από κάτω παρατίθεται η αλληλουχία της παρούσας εγγραφής.

Η αναπαράσταση της αλληλουχίας είναι της μορφής:

```
ORIGIN
        1 atgaaaaaga cagctatcgc gattgcagtg gcactggctg gtttcgctac cgtagcgcag
       61 gccgctccga aagataacac ctggtacact ggtgctaaac tgggctggtc ccagtaccat
      121 gacactggtt tcatcaacaa caatggcccg acccatgaaa accaactggg cgctggtgct
      181 tttggtggtt accaggttaa cccgtatgtt ggctttgaaa tgggttacga ctggttaggt
      241 cgtatgccgt acaaaggcag cgttgaaaac ggtgcataca aagctcaggg cgttcaactg
      301 accgctaaac tgggttaccc aatcactgac gacctggaca tctacactcg tctgggtggc
      361 atggtatggc gtgcagacac taaatccaac gtttatggta aaaaccacga caccggcgtt
      421 tctccggtct tcgctggcgg tgttgagtac gcgatcactc ctgaaatcgc tacccgtctg
      481 gaataccagt ggaccaacaa catcggtgac gcacacacca tcggcactcg tccggacaac
      541 ggcatgctga gcctgggtgt ttcctaccgt ttcggtcagg gcgaagcagc tccagtagtt
      601 gctccggctc cagctccggc accggaagta cagaccaagc acttcactct gaagtctgac
      661 gttctgttca acttcaacaa agcaaccctg aaaccggaag gtcaggctgc tctggatcag
      721 ctgtacagcc agctgagcaa cctggatccg aaagacggtt ccgtagttgt tctgggttac
      781 accgaccgca tcggttctga cgcttacaac cagggtctgt ccgagcgccg tgctcagtct
      841 gttgttgatt acctgatctc caaaggtatc ccggcagaca agatctccgc acgtggtatg
      901 ggcgaatcca acccggttac tggcaacacc tgtgacaacg tgaaacagcg tgctgcactg
      961 atcgactgcc tggctccgga tcgtcgcgta gagatcgaag ttaaaggtat caaagacgtt
     1021 gtaactcagc cgcaggctta a
//
```

- Τα νουκλεοτίδια απεικονίζονται με τον κώδικα ενός γράμματος ανάλογα με την αζωτούχο βάση την οποία αποτελούνται.

- Κάθε αλληλουχία αποτελείται από 60 αμινοξικά κατάλοιπα ανά γραμμή, σε ομάδες των δέκα αμινοξικών καταλοίπων, ξεκινώντας πάντα από την θέση 11 της γραμμής. Οι ομάδες των 10 καταλοίπων χωρίζονται μεταξύ τους με κενό διάστημα.

- Από τη θέση 9 της γραμμής και προς τα αριστερά υπάρχει ένας αριθμός που δείχνει την αρίθμηση του πρώτου καταλοίπου κάθε γραμμής.

//: Λήξη της εγγραφής.

2. Εγγραφή της Uniprot για την πρωτεϊνική αλληλουχία της Outer membrane protein A (ompA) από τον οργανισμό Escherichia coli.

```
ID   OMPA_ECOLI              Reviewed;         346 AA.
AC   P0A910; P02934;
DT   20-JUL-1986, integrated into UniProtKB/Swiss-Prot.
DT   20-JUL-1986, sequence version 1.
DT   06-JAN-2015, entry version 99.
DE   RecName: Full=Outer membrane protein A;
DE   AltName: Full=Outer membrane protein II*;
DE   Flags: Precursor;
GN   Name=ompA; Synonyms=con, tolG, tut; OrderedLocusNames=b0957, JW0940;
OS   Escherichia coli (strain K12).
OC   Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales;
OC   Enterobacteriaceae; Escherichia.
OX   NCBI_TaxID=83333;
RN   [1]
RP   NUCLEOTIDE SEQUENCE [GENOMIC DNA].
RC   STRAIN=K12;
RX   PubMed=6253901; DOI=10.1093/nar/8.13.3011;
RA   Beck E., Bremer E.;
RT   "Nucleotide sequence of the gene ompA coding the outer membrane
RT   protein II of Escherichia coli K-12.";
RL   Nucleic Acids Res. 8:3011-3027(1979).
RN   [2]
RP   NUCLEOTIDE SEQUENCE [GENOMIC DNA].
RC   STRAIN=K12;
RX   PubMed=6260961; DOI=10.1016/0022-2836(80)90193-X;
RA   Movva N.R., Nakamura K., Inouye M.;
RT   "Gene structure of the OmpA protein, a major surface protein of
RT   Escherichia coli required for cell-cell interaction.";
RL   J. Mol. Biol. 143:317-328(1979).
RN   [3]
RP   NUCLEOTIDE SEQUENCE [LARGE SCALE GENOMIC DNA].
RC   STRAIN=K12 / W3110 / ATCC 27325 / DSM 5911;
RX   PubMed=8905232; DOI=10.1093/dnares/3.3.137;
RA   Oshima T., Aiba H., Baba T., Fujita K., Hayashi K., Honjo A.,
RA   Ikemoto K., Inada T., Itoh T., Kajihara M., Kanai K., Kashimoto K.,
RA   Kimura S., Kitagawa M., Makino K., Masuda S., Miki T., Mizobuchi K.,
RA   Mori H., Motomura K., Nakamura Y., Nashimoto H., Nishio Y., Saito N.,
RA   Sampei G., Seki Y., Tagami H., Takemoto K., Wada C., Yamamoto Y.,
RA   Yano M., Horiuchi T.;
RT   "A 718-kb DNA sequence of the Escherichia coli K-12 genome
RT   corresponding to the 12.7-28.0 min region on the linkage map.";
RL   DNA Res. 3:137-155(1995).
RN   [4]
RP   NUCLEOTIDE SEQUENCE [LARGE SCALE GENOMIC DNA].
RC   STRAIN=K12 / MG1655 / ATCC 47076;
RX   PubMed=9278503; DOI=10.1126/science.277.5331.1453;
RA   Blattner F.R., Plunkett G. III, Bloch C.A., Perna N.T., Burland V.,
RA   Riley M., Collado-Vides J., Glasner J.D., Rode C.K., Mayhew G.F.,
RA   Gregor J., Davis N.W., Kirkpatrick H.A., Goeden M.A., Rose D.J.,
RA   Mau B., Shao Y.;
RT   "The complete genome sequence of Escherichia coli K-12.";
RL   Science 277:1453-1462(1996).
RN   [5]
RP   NUCLEOTIDE SEQUENCE [LARGE SCALE GENOMIC DNA].
RC   STRAIN=K12 / W3110 / ATCC 27325 / DSM 5911;
RX   PubMed=16738553; DOI=10.1038/msb4100049;
RA   Hayashi K., Morooka N., Yamamoto Y., Fujita K., Isono K., Choi S.,
RA   Ohtsubo E., Baba T., Wanner B.L., Mori H., Horiuchi T.;
```

```
RT   "Highly accurate genome sequences of Escherichia coli K-12 strains
RT   MG1655 and W3110.";
RL   Mol. Syst. Biol. 2:E1-E5(2005).
RN   [6]
RP   PROTEIN SEQUENCE OF 22-346.
RC   STRAIN=K12;
RX   PubMed=7001461; DOI=10.1073/pnas.77.8.4592;
RA   Chen R., Schmidmayr W., Kramer C., Chen-Schmeisser U., Henning U.;
RT   "Primary structure of major outer membrane protein II (ompA protein)
RT   of Escherichia coli K-12.";
RL   Proc. Natl. Acad. Sci. U.S.A. 77:4592-4596(1979).
RN   [7]
RP   PROTEIN SEQUENCE OF 22-34.
RC   STRAIN=K12 / EMG2;
RX   PubMed=9298646; DOI=10.1002/elps.1150180807;
RA   Link A.J., Robison K., Church G.M.;
RT   "Comparing the predicted and observed properties of proteins encoded
RT   in the genome of Escherichia coli K-12.";
RL   Electrophoresis 18:1259-1313(1996).
RN   [8]
RP   PROTEIN SEQUENCE OF 22-32.
RC   STRAIN=K12 / W3110 / ATCC 27325 / DSM 5911;
RA   Pasquali C., Sanchez J.-C., Ravier F., Golaz O., Hughes G.J.,
RA   Frutiger S., Paquet N., Wilkins M., Appel R.D., Bairoch A.,
RA   Hochstrasser D.F.;
RL   Submitted (AUG-1994) to UniProtKB.
RN   [9]
RP   PROTEIN SEQUENCE OF 22-26.
RC   STRAIN=K12 / W3110 / ATCC 27325 / DSM 5911;
RX   PubMed=9629924; DOI=10.1002/elps.1150190539;
RA   Molloy M.P., Herbert B.R., Walsh B.J., Tyler M.I., Traini M.,
RA   Sanchez J.-C., Hochstrasser D.F., Williams K.L., Gooley A.A.;
RT   "Extraction of membrane proteins by differential solubilization for
RT   separation using two-dimensional gel electrophoresis.";
RL   Electrophoresis 19:837-844(1997).
RN   [10]
RP   MUTANTS RESISTANT TO PHAGE ENTRY.
RX   PubMed=6086577;
RA   Morona R., Klose M., Henning U.;
RT   "Escherichia coli K-12 outer membrane protein (OmpA) as a
RT   bacteriophage receptor: analysis of mutant genes expressing altered
RT   proteins.";
RL   J. Bacteriol. 159:570-578(1983).
RN   [11]
RP   MUTANTS RESISTANT TO PHAGE ENTRY.
RX   PubMed=3902787;
RA   Morona R., Kramer C., Henning U.;
RT   "Bacteriophage receptor area of outer membrane protein OmpA of
RT   Escherichia coli K-12.";
RL   J. Bacteriol. 164:539-543(1984).
RN   [12]
RP   PORIN ACTIVITY.
RC   STRAIN=K12;
RX   PubMed=1370823;
RA   Sugawara E., Nikaido H.;
RT   "Pore-forming activity of OmpA protein of Escherichia coli.";
RL   J. Biol. Chem. 267:2507-2511(1991).
RN   [13]
RP   SUBCELLULAR LOCATION.
RX   PubMed=7813480; DOI=10.1111/j.1432-1033.1994.00891.x;
RA   Kuhn A., Kiefer D., Koehne C., Zhu H.-Y., Tschantz W.R., Dalbey R.E.;
```

RT "Evidence for a loop-like insertion mechanism of pro-Omp A into the
RT inner membrane of Escherichia coli.";
RL Eur. J. Biochem. 226:891-897(1993).
RN [14]
RP TOPOLOGY.
RX PubMed=8106193;
RA Gromiha M.M., Ponnuswamy P.K.;
RT "Prediction of transmembrane beta-strands from hydrophobic
RT characteristics of proteins.";
RL Int. J. Pept. Protein Res. 42:420-431(1992).
RN [15]
RP IDENTIFICATION BY 2D-GEL.
RX PubMed=9298644; DOI=10.1002/elps.1150180805;
RA VanBogelen R.A., Abshire K.Z., Moldover B., Olson E.R.,
RA Neidhardt F.C.;
RT "Escherichia coli proteome analysis using the gene-protein database.";
RL Electrophoresis 18:1243-1251(1996).
RN [16]
RP TOPOLOGY.
RX PubMed=10368142;
RA Koebnik R.;
RT "Structural and functional roles of the surface-exposed loops of the
RT beta-barrel membrane protein OmpA from Escherichia coli.";
RL J. Bacteriol. 181:3688-3694(1998).
RN [17]
RP DIMERIZATION, AND SUBCELLULAR LOCATION.
RC STRAIN=BL21-DE3;
RX PubMed=16079137; DOI=10.1074/jbc.M506479200;
RA Stenberg F., Chovanec P., Maslen S.L., Robinson C.V., Ilag L.,
RA von Heijne G., Daley D.O.;
RT "Protein complexes of the Escherichia coli cell envelope.";
RL J. Biol. Chem. 280:34409-34419(2004).
RN [18]
RP SUBCELLULAR LOCATION.
RC STRAIN=K12 / MG1655 / ATCC 47076;
RX PubMed=21778229; DOI=10.1074/jbc.M111.245696;
RA Fontaine F., Fuchs R.T., Storz G.;
RT "Membrane localization of small proteins in Escherichia coli.";
RL J. Biol. Chem. 286:32464-32474(2010).
RN [19]
RP X-RAY CRYSTALLOGRAPHY (2.5 ANGSTROMS) OF 22-192.
RX PubMed=9808047; DOI=10.1038/2983;
RA Pautsch A., Schulz G.E.;
RT "Structure of the outer membrane protein A transmembrane domain.";
RL Nat. Struct. Biol. 5:1013-1017(1997).
RN [20]
RP X-RAY CRYSTALLOGRAPHY (1.65 ANGSTROMS).
RX PubMed=10764596; DOI=10.1006/jmbi.2000.3671;
RA Pautsch A., Schulz G.E.;
RT "High-resolution structure of the OmpA membrane domain.";
RL J. Mol. Biol. 298:273-282(1999).
RN [21]
RP STRUCTURE BY NMR OF 22-197.
RX PubMed=11276254; DOI=10.1038/86214;
RA Arora A., Abildgaard F., Bushweller J.H., Tamm L.K.;
RT "Structure of outer membrane protein A transmembrane domain by NMR
RT spectroscopy.";
RL Nat. Struct. Biol. 8:334-338(2000).
RN [22]
RP MASS SPECTROMETRY.
RX PubMed=10757971; DOI=10.1021/bi000150m;

```
RA   le Coutre J., Whitelegge J.P., Gross A., Turk E., Wright E.M.,
RA   Kaback H.R., Faull K.F.;
RT   "Proteomics on full-length membrane proteins using mass
RT   spectrometry.";
RL   Biochemistry 39:4237-4242(1999).
CC   -!- FUNCTION: Required for the action of colicins K and L and for the
CC       stabilization of mating aggregates in conjugation. Serves as a
CC       receptor for a number of T-even like phages. Also acts as a porin
CC       with low permeability that allows slow penetration of small
CC       solutes.
CC   -!- SUBUNIT: Homodimer.
CC   -!- INTERACTION:
CC       P0C0V0:degP; NbExp=5; IntAct=EBI-371347, EBI-547165;
CC       P0A850:tig; NbExp=3; IntAct=EBI-371347, EBI-544862;
CC   -!- SUBCELLULAR LOCATION: Cell outer membrane
CC       {ECO:0000269|PubMed:16079137, ECO:0000269|PubMed:21778229,
CC       ECO:0000269|PubMed:7813480}; Multi-pass membrane protein
CC       {ECO:0000269|PubMed:16079137, ECO:0000269|PubMed:21778229,
CC       ECO:0000269|PubMed:7813480}.
CC   -!- MASS SPECTROMETRY: Mass=35177; Method=Electrospray; Range=22-346;
CC       Evidence={ECO:0000269|PubMed:10757971;
CC   -!- SIMILARITY: Belongs to the OmpA family. {ECO:0000305}.
CC   -!- SIMILARITY: Contains 1 OmpA-like domain. {ECO:0000255|PROSITE-
CC       ProRule:PRU00473}.
CC   ---------------------------------------------------------------------
CC   Copyrighted by the UniProt Consortium, see http://www.uniprot.org/terms
CC   Distributed under the Creative Commons Attribution-NoDerivs License
CC   ---------------------------------------------------------------------
DR   EMBL; V00307; CAA23588.1; -; Genomic_DNA.
DR   EMBL; U00096; AAC74043.1; -; Genomic_DNA.
DR   EMBL; AP009048; BAA35715.1; -; Genomic_DNA.
DR   PIR; A93707; MMECA.
DR   RefSeq; NP_415477.1; NC_000913.3.
DR   RefSeq; YP_489229.1; NC_007779.1.
DR   PDB; 1BXW; X-ray; 2.50 A; A=21-192.
DR   PDB; 1G90; NMR; -; A=22-197.
DR   PDB; 1QJP; X-ray; 1.65 A; A=22-192.
DR   PDB; 2GE4; NMR; -; A=22-197.
DR   PDB; 2JMM; NMR; -; A=23-197.
DR   PDB; 3NB3; EM; -; A/B/C=1-346.
DR   PDBsum; 1BXW; -.
DR   PDBsum; 1G90; -.
DR   PDBsum; 1QJP; -.
DR   PDBsum; 2GE4; -.
DR   PDBsum; 2JMM; -.
DR   PDBsum; 3NB3; -.
DR   ProteinModelPortal; P0A910; -.
DR   SMR; P0A910; 22-192, 209-346.
DR   DIP; DIP-31879N; -.
DR   IntAct; P0A910; 11.
DR   MINT; MINT-1308131; -.
DR   STRING; 511145.b0957; -.
DR   TCDB; 1.B.6.1.1; the ompa-ompf porin (oop) family.
DR   SWISS-2DPAGE; P0A910; -.
DR   PaxDb; P0A910; -.
DR   PRIDE; P0A910; -.
DR   EnsemblBacteria; AAC74043; AAC74043; b0957.
DR   EnsemblBacteria; BAA35715; BAA35715; BAA35715.
DR   GeneID; 12931038; -.
DR   GeneID; 945571; -.
DR   KEGG; ecj:Y75_p0929; -.
```

```
DR   KEGG; eco:b0957; -.
DR   PATRIC; 32117133; VBIEscCol129921_0991.
DR   EchoBASE; EB0663; -.
DR   EcoGene; EG10669; ompA.
DR   eggNOG; COG2885; -.
DR   HOGENOM; HOG000274199; -.
DR   InParanoid; P0A910; -.
DR   KO; K03286; -.
DR   OMA; EYALTKN; -.
DR   OrthoDB; EOG6PP9QB; -.
DR   BioCyc; EcoCyc:EG10669-MONOMER; -.
DR   BioCyc; ECOL316407:JW0940-MONOMER; -.
DR   EvolutionaryTrace; P0A910; -.
DR   PRO; PR:P0A910; -.
DR   Proteomes; UP000000318; Chromosome.
DR   Proteomes; UP000000625; Chromosome.
DR   Genevestigator; P0A910; -.
DR   GO; GO:0009279; C:cell outer membrane; IDA:EcoliWiki.
DR   GO; GO:0016021; C:integral component of membrane; IDA:EcoliWiki.
DR   GO; GO:0016020; C:membrane; IDA:EcoliWiki.
DR   GO; GO:0019867; C:outer membrane; IDA:EcoliWiki.
DR   GO; GO:0046930; C:pore complex; IEA:UniProtKB-KW.
DR   GO; GO:0015288; F:porin activity; IDA:EcoCyc.
DR   GO; GO:0005198; F:structural molecule activity; IEA:InterPro.
DR   GO; GO:0006974; P:cellular response to DNA damage stimulus; IEP:EcoliWiki.
DR   GO; GO:0000746; P:conjugation; IMP:EcoliWiki.
DR   GO; GO:0009597; P:detection of virus; IMP:EcoliWiki.
DR   GO; GO:0034220; P:ion transmembrane transport; IDA:EcoCyc.
DR   GO; GO:0006811; P:ion transport; IDA:EcoliWiki.
DR   GO; GO:0006810; P:transport; IDA:EcoliWiki.
DR   GO; GO:0046718; P:viral entry into host cell; IMP:EcoliWiki.
DR   Gene3D; 2.40.160.20; -; 1.
DR   Gene3D; 3.30.1330.60; -; 1.
DR   InterPro; IPR011250; OMP/PagP_b-brl.
DR   InterPro; IPR006664; OMP_bac.
DR   InterPro; IPR002368; OmpA.
DR   InterPro; IPR006690; OMPA-like_CS.
DR   InterPro; IPR000498; OmpA-like_TM_dom.
DR   InterPro; IPR006665; OmpA/MotB_C.
DR   Pfam; PF00691; OmpA; 1.
DR   Pfam; PF01389; OmpA_membrane; 1.
DR   PRINTS; PR01021; OMPADOMAIN.
DR   PRINTS; PR01022; OUTRMMBRANEA.
DR   SUPFAM; SSF103088; SSF103088; 1.
DR   SUPFAM; SSF56925; SSF56925; 1.
DR   PROSITE; PS01068; OMPA_1; 1.
DR   PROSITE; PS51123; OMPA_2; 1.
PE   1: Evidence at protein level;
KW   3D-structure; Cell outer membrane; Complete proteome; Conjugation;
KW   Direct protein sequencing; Disulfide bond; Ion transport; Membrane;
KW   Porin; Reference proteome; Repeat; Signal; Transmembrane;
KW   Transmembrane beta strand; Transport.
FT   SIGNAL        1     21       {ECO:0000269|PubMed:7001461,
FT                                ECO:0000269|PubMed:9298646,
FT                                ECO:0000269|PubMed:9629924,
FT                                ECO:0000269|Ref.8}.
FT   CHAIN        22    346       Outer membrane protein A.
FT                                /FTId=PRO_0000020094.
FT   TOPO_DOM     22     26       Periplasmic.
FT   TRANSMEM     27     37       Beta stranded.
FT   TOPO_DOM     38     54       Extracellular.
```

```
FT   TRANSMEM     55     66       Beta stranded.
FT   TOPO_DOM     67     69       Periplasmic.
FT   TRANSMEM     70     78       Beta stranded.
FT   TOPO_DOM     79     95       Extracellular.
FT   TRANSMEM     96    107       Beta stranded.
FT   TOPO_DOM    108    111       Periplasmic.
FT   TRANSMEM    112    124       Beta stranded.
FT   TOPO_DOM    125    137       Extracellular.
FT   TRANSMEM    138    151       Beta stranded.
FT   TOPO_DOM    152    155       Periplasmic.
FT   TRANSMEM    156    163       Beta stranded.
FT   TOPO_DOM    164    181       Extracellular.
FT   TRANSMEM    182    190       Beta stranded.
FT   TOPO_DOM    191    346       Periplasmic.
FT   REPEAT      201    202       1.
FT   REPEAT      203    204       2.
FT   REPEAT      205    206       3.
FT   REPEAT      207    208       4.
FT   DOMAIN      210    338       OmpA-like. {ECO:0000255|PROSITE-
FT                                ProRule:PRU00473}.
FT   REGION      197    208       Hinge-like.
FT   REGION      201    208       4 X 2 AA tandem repeats of A-P.
FT   DISULFID    311    323
FT   STRAND       27     37       {ECO:0000244|PDB:1QJP}.
FT   STRAND       41     43       {ECO:0000244|PDB:1G90}.
FT   STRAND       46     48       {ECO:0000244|PDB:1G90}.
FT   STRAND       50     53       {ECO:0000244|PDB:2GE4}U.
FT   STRAND       55     67       {ECO:0000244|PDB:1QJP}.
FT   STRAND       70     81       {ECO:0000244|PDB:1QJP}.
FT   STRAND       93    128       {ECO:0000244|PDB:1QJP}.
FT   STRAND      130    132       {ECO:0000244|PDB:1QJP}.
FT   STRAND      134    153       {ECO:0000244|PDB:1QJP}.
FT   STRAND      156    165       {ECO:0000244|PDB:1QJP}.
FT   TURN        172    175       {ECO:0000244|PDB:1G90}.
FT   STRAND      182    190       {ECO:0000244|PDB:1QJP}.
SQ   SEQUENCE   346 AA;  37201 MW;  195147734CDF8B04 CRC64;
     MKKTAIAIAV ALAGFATVAQ AAPKDNTWYT GAKLGWSQYH DTGFINNNGP THENQLGAGA
     FGGYQVNPYV GFEMGYDWLG RMPYKGSVEN GAYKAQGVQL TAKLGYPITD DLDIYTRLGG
     MVWRADTKSN VYGKNHDTGV SPVFAGGVEY AITPEIATRL EYQWTNNIGD AHTIGTRPDN
     GMLSLGVSYR FGQGEAAPVV APAPAPAPEV QTKHFTLKSD VLFNFNKATL KPEGQAALDQ
     LYSQLSNLDP KDGSVVVLGY TDRIGSDAYN QGLSERRAQS VVDYLISKGI PADKISARGM
     GESNPVTGNT CDNVKQRAAL IDCLAPDRRV EIEVKGIKDV VTQPQA
//
```

<u>Επεξηγήσεις των σημαντικότερων πεδίων μιας εγγραφής UNIPROT</u>

ID (Identification):

Είναι της μορφής *Entry_name data_class; molecule_type; sequence length*

Entry_name: Το όνομα της αλληλουχίας χαρακτηριστικό για τη βάση UNIPROT.

π.χ. OMPA_ECOLI. Το πρώτο τμήμα υποδηλώνει το όνομα της αλληλουχίας όπως είναι κατατεθειμένο στην βάση. Μπορεί να έχει μήκος μέχρι 4 χαρακτήρες. Το δεύτερο καθορίζει το είδος από το οποίο προέρχεται η αλληλουχία. Μπορεί να έχει μήκος μέχρι 5 χαρακτήρες.

data_class: Δηλώνει αν η εγγραφή έχει σχολιαστεί ή όχι με βάση τα κριτήρια της βάσης UNIPROT.

molecule_type: Δηλώνει σε ποια ομάδα μακρομορίων ανήκει η αλληλουχία. Για τις εγγραφές της UNIPROT είναι PRT (Protein).

sequence length: Το μήκος της αλληλουχίας σε αμινοξικά κατάλοιπα (AA).

AC (Accession number): Είναι ένας χαρακτηριστικός κωδικός που αποκτά μια πολυπεπτιδική αλυσίδα όταν κατατίθεται στην βάση. Χρησιμεύει στην αναγνώριση εγγραφών ανάμεσα στις διαφορετικές εκδόσεις της βάσης όπως αυτή ανανεώνεται ανά τακτά χρονικά διαστήματα.

DT (Date): Αναγραφή ημερομηνίας για τη δημιουργία της παρούσας εγγραφής, τελευταίας τροποποιήσης, προσθήκης σχολίων.

DE (Description): Γενική περιγραφή για την αλληλουχία.

GN (Gene name): Γονίδιο από το οποίο με μετάφραση προέκυψε η αμινοξική αλληλουχία.

OS (Organism Species): Οργανισμός απ' όπου προήλθε η αλληλουχία. Ακολουθείται η διώνυμη ονομασία κατά Λινναίο.

OG (Organelle): Επεξηγεί αν το γονίδιο που κωδικοποιεί την συγκεκριμένη αλληλουχία εδράζεται σε μιτοχόνδρια, χλωροπλάστες ή πλασμίδιο.

OC (Organism Classification): Συστηματική ταξινόμηση του οργανισμού απ'όπου προήλθε η αλληλουχία.

OX (Organism taxonomy cross-reference): Παραπομπή σε βάση δεδομένων συστηματικής ταξινόμησης των οργανισμών.

- **RN, RP, RC, RX, RA, RT, RL :** Τα παρακάτω πεδία σχετίζονται με βιβλιογραφικές αναφορές σχετικές με την παρούσα εγγραφή.

RN (Reference number): Αύξων αριθμός αναφοράς σχετικής με την παρούσα εγγραφή.

RP (Reference Position): Περιέχει λίγες πληροφορίες σχετικές με το τι πραγματεύεται η συγκεκριμένη αναφορά.

RX (Reference cross-reference): Παραπομπές σε βιβλιογραφικές βάσεις δεδομένων π.χ. PUBMED.

RA (Reference author): Λίστα με τους συγγραφείς της παρούσας αναφοράς.

RT (Reference title): Τίτλος της παρούσας εργασίας όπως δημοσιεύτηκε σε επιστημονικά περιοδικά.

RL (Reference Location): Περιοδικό ή βιβλίο όπου δημοσιεύτηκε η παρούσα εργασία.

CC (Comments): Το πεδίο αυτό περιέχει μία σειρά από πληροφορίες πάσης φύσεως σχετικές με την αλληλουχία. Χωρίζεται σε υπο-πεδία όπως:

CATALYTIC ACTIVITY: Περιγραφή της αντίδρασης που καταλύεται αν η αλληλουχία είναι ένζυμο.

ALTERNATIVE PRODUCTS: Αναφέρεται αν υπάρχουν σχετικές με αυτή αλληλουχίες που έχουν προκύψει από εναλλακτικό μάτισμα.

FUNCTION: Σύντομη περιγραφή της λειτουργίας που συμμετέχει η αλληλουχία.

SUBCELLULAR LOCATION: Θέση της αλληλουχίας στο κύτταρο.

SUBUNIT: Το πεδίο εμφανίζεται στην περίπτωση που η αλληλουχία συμμετέχει στην δημιουργία τεταρτοταγούς δομής μιας πρωτεΐνης.

Πρέπει να σημειωθεί πως τα παραπάνω είναι μερικά από τα υπο-πεδία που μπορεί να περιέχονται στο πεδίο CC (Comments).

DR (Database cross-reference): Το πεδίο αυτό δίνει διασυνδέσεις σε άλλες βάσεις δεδομένων που σχετίζονται με την παρούσα εγγραφή όπως η PDB, η EMBL κ.α. με τους αντίστοιχους κωδικούς τους.

KW (Keyword): Το πεδίο αυτό περιέχει ειδικές λέξεις-κλειδιά για τον χαρακτηρισμό της αλληλουχίας όπως αυτές ταξινομούνται με βάση κριτήρια όπως η λειτουργία και η δομή τους.

FT (Feature Table): Το πεδίο αυτό περιέχει στοιχεία χαρακτηριστικά για την αλληλουχία αυτή καθεαυτή και αφορά συγκεκριμένα τμήματά της. Περιλαμβάνει πληροφορίες για:

α. Μεταμεταφραστικές τροποποιήσεις

β. Ποια τμήματα της αλληλουχίας είναι υπεύθυνα για την δέσμευση κάποιου μορίου (π.χ. Receptor-Ligand).

γ. Ποια τμήματα της αλληλουχίας συμμετέχουν για το σχηματισμό του ενεργού κέντρου αν πρόκειται για ένζυμο.

δ. Στοιχεία για τη δευτεροταγή δομή της αλληλουχίας.

ε. Επίσης μπορεί στο πεδίο αυτό μπορεί και να σημειώνονται και διαφορές στην αλληλουχία εάν έχουν προκύψει και αναφέρονται σε άλλες βιβλιογραφικές αναφορές.

SQ (Sequence): Το πεδίο αυτό περιέχει το μήκος της αλληλουχίας σε αμινοξικά κατάλοιπα (AA), το μοριακό βάρος (MW) σε Daltons.

Ακολουθεί η αναπαράσταση της αλληλουχίας ακολουθώντας τους παρακάτω κανόνες:

- Κάθε αμινοξικό κατάλοιπο απεικονίζεται με τον κώδικα του ενός γράμματος κατά IUPAC.

- Κάθε αλληλουχία αποτελείται από 60 αμινοξικά κατάλοιπα ανά γραμμή, σε ομάδες των δέκα αμινοξικών καταλοίπων, ξεκινώντας πάντα από την θέση 6 της γραμμής. Οι ομάδες των 10 καταλοίπων χωρίζονται μεταξύ τους με κενό διάστημα.

//: Τα σύμβολα αυτά υποδηλώνουν το τέλος της εγγραφής.

Π.χ.

```
SQ   SEQUENCE   346 AA;  37201 MW;  195147734CDF8B04 CRC64;
     MKKTAIAIAV ALAGFATVAQ AAPKDNTWYT GAKLGWSQYH DTGFINNNGP THENQLGAGA
     FGGYQVNPYV GFEMGYDWLG RMPYKGSVEN GAYKAQGVQL TAKLGYPITD DLDIYTRLGG
     MVWRADTKSN VYGKNHDTGV SPVFAGGVEY AITPEIATRL EYQWTNNIGD AHTIGTRPDN
     GMLSLGVSYR FGQGEAAPVV APAPAPAPEV QTKHFTLKSD VLFNFNKATL KPEGQAALDQ
     LYSQLSNLDP KDGSVVVLGY TDRIGSDAYN QGLSERRAQS VVDYLISKGI PADKISARGM
     GESNPVTGNT CDNVKQRAAL IDCLAPDRRV EIEVKGIKDV VTQPQA
//
```

3. Εγγραφή της PROSITE για την πρωτεϊνική αλληλουχία της Outer membrane protein A (ompA).

```
ID   OMPA_1; PATTERN.
AC   PS01068;
DT   NOV-1995 (CREATED); DEC-2004 (DATA UPDATE); FEB-2015 (INFO UPDATE).
DE   OmpA-like domain.
PA   [LIVMA]-x-[GT]-x-[TA]-[DAN]-x(2,3)-[DG]-[GSTPNKQ]-x(2)-[LFYDEPAVI]-[NQS]-
PA   x(2)-[LI]-[SG]-[QEA]-[KRQENAD]-R-A-x(2)-[LVAIT]-x(3)-[LIVMF]-x(4,5)-
PA   [LIVMF]-x(4)-[LIVM]-x(3)-[SGW]-x-G.
NR   /RELEASE=2015_04,548208;
NR   /TOTAL=55(55); /POSITIVE=55(55); /UNKNOWN=0(0); /FALSE_POS=0(0);
NR   /FALSE_NEG=10; /PARTIAL=2;
CC   /TAXO-RANGE=???P?; /MAX-REPEAT=1;
CC   /VERSION=1;
DR   P65594, ARFA_MYCBO , T; A1KH31, ARFA_MYCBP , T; P9WIU4, ARFA_MYCTO , T;
DR   P9WIU5, ARFA_MYCTU , T; Q9S3P9, MOTY_VIBAN , T; P46233, MOTY_VIBPA , T;
DR   Q8U9L5, OMP16_AGRT5, T; P0A3S9, OMP16_BRUAB, T; P0A3S7, OMP16_BRUME, T;
DR   P0A3S8, OMP16_BRUSU, T; Q98F85, OMP16_RHILO, T; Q926C3, OMP16_RHIME, T;
DR   P07050, OMP3_NEIGO , T; Q9S3R8, OMP40_PORGI, T; Q9S3R9, OMP41_PORGI, T;
DR   P0A0V2, OMP4_NEIMA , T; P0A0V3, OMP4_NEIMB , T; P43840, OMP51_HAEIN, T;
DR   P38368, OMP52_HAEIF, T; P45996, OMP53_HAEIF, T; Q05146, OMPA_BORAV , T;
DR   P57414, OMPA_BUCAI , T; Q8K9L4, OMPA_BUCAP , T; P24016, OMPA_CITFR , T;
DR   P0A911, OMPA_ECO57 , T; P0A910, OMPA_ECOLI , T; P09146, OMPA_ENTAE , T;
DR   B7LNW7, OMPA_ESCF3 , T; P0C8Z2, OMPA_ESCFE , T; P24754, OMPA_ESCHE , T;
DR   P24017, OMPA_KLEPN , T; Q8Z7S0, OMPA_SALTI , T; P02936, OMPA_SALTY , T;
DR   P04845, OMPA_SERMA , T; P24755, OMPA_SEROD , T; I2BAK7, OMPA_SHIBC , T;
DR   P0DJO6, OMPA_SHIBL , T; P02935, OMPA_SHIDY , T; Q8ZG77, OMPA_YERPE , T;
DR   P38399, OMPA_YERPS , T; Q89AJ5, PAL_BUCBP  , T; P0A913, PAL_ECO57  , T;
DR   P0A912, PAL_ECOLI  , T; P10324, PAL_HAEIN  , T; P26493, PAL_LEGPN  , T;
DR   Q51886, PAL_PASMU  , T; Q9I4Z4, PAL_PSEAE  , T; P0A138, PAL_PSEPK  , T;
DR   P0A139, PAL_PSEPU  , T; P0A914, PAL_SHIFL  , T; P13794, PORF_PSEAE , T;
DR   P37726, PORF_PSEFL , T; P22263, PORF_PSESY , T; P38369, TPN50_TREPA, T;
DR   P37665, YIAD_ECOLI , T;
DR   P85410, OMP5_HAEPR , P; P80444, OMPA_ACTLI , P;
DR   D3GSC3, LAFU_ECO44 , N; Q47154, LAFU_ECOLI , N; Q6RYW5, OMP38_ACIBA, N;
DR   A3M8K2, OMP38_ACIBT, N; P84838, OMPC_GLUDA , N; P07021, YFIB_ECOLI , N;
DR   P0C536, YN58_BRUAB , N; Q2YJ83, YP57_BRUA2 , N; Q8YDY8, YU36_BRUME , N;
DR   Q9RPX3, YU58_BRUSU , N;
3D   1OAP; 1R1M; 2AIZ; 2HQS; 2K1S; 2KGW; 2L26; 2LBT; 2LCA; 2W8B;
DO   PDOC00819;
//
```

Επεξηγήσεις των σημαντικότερων πεδίων μιας εγγραφής στην PROSITE

ID (Identification): Είναι της γενικής μορφής

ID ENTRY_NAME; ENTRY_TYPE

Το πρώτο τμήμα είναι η χαρακτηριστική ονομασία που εμφανίζει η εγγραφή χαρακτηριστική για τη βάση PROSITE, ενώ το δεύτερο τμήμα υποδηλώνει τον τύπο της εγγραφής.

AC (ACcession number): Πρόκειται για τον χαρακτηριστικό κωδικό που αποκτά μια νεοεισερχόμενη εγγραφή στην PROSITE και χρησιμεύει στην αναγνώριση της εγγραφής ανάμεσα στις διαφορετικές εκδόσεις της βάσης PROSITE.

DT (DaTe): Το πεδίο αυτό περιέχει τις ημερομηνίες δημιουργίας και τελευταίας ανανέωσης (σχολιασμός) της εγγραφής.

DE (DEscription): Περιέχει μια γενική περιγραφή για την συγκεκριμένη εγγραφή.

PA (PAttern): Στο πεδίο αυτό αναγράφεται το πρότυπο της αλληλουχίας (pattern) που ακολουθούν τα μέλη της συγκεκριμένης εγγραφής.

Οι συμβάσεις που ακολουθούμε για την αναπαράσταση του pattern είναι:

1. Τα αμινοξέα απεικονίζονται με τον κώδικα του ενός γράμματος κατά IUPAC.
2. Το σύμβολο **x** σημαίνει ότι στη θέση αυτή μπορεί να υπάρχει οποιοδήποτε αμινοξύ.
3. [...] Τα αμινοξέα που περιέχονται μέσα στις αγκύλες είναι τα επιτρεπτά για τη συγκεκριμένη θέση. Για παράδειγμα αν περιέχεται στις αγκύλες [ALT] σημαίνει ότι στη συγκεκριμένη θέση επιτρέπεται να βρίσκεται Αλανίνη ή Λευκίνη ή Θρεονίνη.
4. Τα άγκιστρα υποδηλώνουν ότι όσα αμινοξέα περιέχονται σε αυτά δεν επιτρέπεται να βρίσκονται στις συγκεκριμένες θέσεις.
5. Κάθε στοιχείο του μοτίβου χωρίζεται από το γειτονικό του με μια παύλα (-).
6. Αν ένα στοιχείο επαναλαμβάνεται μπορεί να αναπαρασταθεί με ένα αριθμητικό δείκτη σε παρενθέσεις που δηλώνει τον αριθμό των επαναλήψεων π.χ. x(3). Στην περίπτωση που εντός της παρενθέσεως περιέχονται δύο αριθμοί που χωρίζονται μεταξύ τους με κόμμα τούτο σημαίνει ότι ο αριθμός των επαναλήψεων μπορεί να παίρνει ένα εύρος τιμών που καθορίζεται από τις τιμές που περιέχονται στις παρενθέσεις π.χ. (2,4) Ο αριθμός των επαναλήψεων μπορεί να είναι 2 ή 3 ή 4.
7. Αν το μοτίβο περιορίζεται στο αμινοτελικό ή το καρβοξυτελικό άκρο η αναπαράσταση ξεκινά με τα σύμβολα '<' και '>' αντίστοιχα.
8. Η τελεία υποδηλώνει το τέλος του pattern.

NR (Numerical Results): Τα πεδία αυτά περιέχουν στοιχεία που προκύπτουν από την σάρωση (pattern scan) της βάσης SWISS-PROT με το pattern της PROSITE.

Πιο συγκεκριμένα περιλαμβάνουν:

/RELEASE: Η έκδοση της UNIPROT που έχει χρησιμοποιηθεί καθώς και ο αριθμός των εγγραφών που περιέχονται σε αυτή.

/TOTAL: Συνολικός αριθμός εγγραφών της UNIPROT όπου φαίνεται να συναντάται το μοτίβο.

/POSITIVE: Αριθμός των εγγραφών που είναι βέβαιο ότι συναντάται το pattern και ανήκουν σε οικογένεια της PROSITE.

/UNKNOWN: Αριθμός των εγγραφών που πιθανά ανήκει στην οικογένεια της PROSITE.

/FALSE_POS: Εγγραφές της UNIPROT όπου εμφανίζεται το pattern αλλά δεν σχετίζονται με την συγκεκριμένη οικογένεια.

/FALSE_NEG: Αριθμός εγγραφών της UNIPROT που ανήκουν στη συγκεκριμένη οικογένεια αλλά δεν βρέθηκαν κατά τη σάρωση μοτίβου.

/PARTIAL: Αριθμός αλληλουχιών της UNIPROT που δεν είναι πλήρεις (fragments), ανήκουν στην συγκεκριμένη οικογένεια της PROSITE, αλλά δεν ανιχνεύονται από το PROSITE λόγω έλλειψης τμημάτων της αλληλουχίας.

CC (Comments): Στα υπο-πεδία του Comments περιέχονται γενικά σχόλια που σχετίζονται με την PROSITE.

DR (Database Reference): Περιέχει όλες τις εγγραφές της UNIPROT που ακολουθούν το συγκεκριμένο μοτίβο.

3D (3D Structure): Περιέχει όλες τις εγγραφές της Protein Data Bank που περιέχει τις δομές βιομακρομορίων και ακολουθούν το συγκεκριμένο μοτίβο.

DO (Documentation): Σύνδεσμος για εγγραφή που αναλυτικά στοιχεία σχετικά με τη βιολογική λειτουργία των αλληλουχιών που περιέχουν το συγκεκριμένο μοτίβο καθώς και βιβλιογραφικές αναφορές.

//: Δηλώνει το τέλος της εγγραφής.

4. Εγγραφή της PDB για την δομή στο χώρο της Outer membrane protein A (ompA) από τον οργανισμό Escherichia coli.

```
HEADER    MEMBRANE PROTEIN                        03-OCT-98   1BXW
TITLE     OUTER MEMBRANE PROTEIN A (OMPA) TRANSMEMBRANE DOMAIN
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: PROTEIN (OUTER MEMBRANE PROTEIN A);
COMPND   3 CHAIN: A;
COMPND   4 FRAGMENT: TRANSMEMBRANE DOMAIN;
COMPND   5 ENGINEERED: YES;
COMPND   6 MUTATION: YES
SOURCE    MOL_ID: 1;
SOURCE   2 ORGANISM_SCIENTIFIC: ESCHERICHIA COLI BL21(DE3);
SOURCE   3 ORGANISM_TAXID: 469008;
SOURCE   4 STRAIN: BL21DE3;
SOURCE   5 GENE: OMPA;
SOURCE   6 EXPRESSION_SYSTEM: ESCHERICHIA COLI BL21(DE3);
SOURCE   7 EXPRESSION_SYSTEM_TAXID: 469008;
SOURCE   8 EXPRESSION_SYSTEM_STRAIN: BL21DE3;
SOURCE   9 EXPRESSION_SYSTEM_PLASMID: PET3B-171
KEYWDS    OUTER MEMBRANE, TRANSMEMBRANE PROTEIN
EXPDTA    X-RAY DIFFRACTION
AUTHOR    G.E.SCHULZ,A.PAUTSCH
REVDAT   3   24-FEB-09 1BXW    1       VERSN
REVDAT   2   22-DEC-99 1BXW    4       HEADER COMPND REMARK JRNL
REVDAT   2 2                    4       ATOM   SOURCE SEQRES
REVDAT   1   14-OCT-98 1BXW    0
JRNL        AUTH   A.PAUTSCH,G.E.SCHULZ
JRNL        TITL   STRUCTURE OF THE OUTER MEMBRANE PROTEIN A
JRNL        TITL 2 TRANSMEMBRANE DOMAIN.
JRNL        REF    NAT.STRUCT.BIOL.              V.   5  1013 1998
JRNL        REFN                   ISSN 1072-8368
JRNL        PMID   9808047
JRNL        DOI    10.1038/2983
REMARK   1
REMARK   2
REMARK   2 RESOLUTION.    2.50 ANGSTROMS.
REMARK   3
REMARK   3 REFINEMENT.
REMARK   3   PROGRAM     : REFMAC
REMARK   3   AUTHORS     : MURSHUDOV,VAGIN,DODSON
REMARK   3
REMARK   3  DATA USED IN REFINEMENT.
REMARK   3   RESOLUTION RANGE HIGH (ANGSTROMS) : 2.50
REMARK   3   RESOLUTION RANGE LOW  (ANGSTROMS) : 50.00
REMARK   3   DATA CUTOFF            (SIGMA(F)) : 0.000
REMARK   3   COMPLETENESS FOR RANGE        (%) : 89.0
REMARK   3   NUMBER OF REFLECTIONS             : 8328
REMARK   3
REMARK   3  FIT TO DATA USED IN REFINEMENT.
REMARK   3   CROSS-VALIDATION METHOD          : THROUGHOUT
REMARK   3   FREE R VALUE TEST SET SELECTION  : RANDOM
REMARK   3   R VALUE     (WORKING + TEST SET) : NULL
REMARK   3   R VALUE            (WORKING SET) : 0.189
REMARK   3   FREE R VALUE                     : 0.235
REMARK   3   FREE R VALUE TEST SET SIZE   (%) : 5.000
REMARK   3   FREE R VALUE TEST SET COUNT      : 404
REMARK   3
REMARK   3  NUMBER OF NON-HYDROGEN ATOMS USED IN REFINEMENT.
REMARK   3   PROTEIN ATOMS            : 1330
REMARK   3   NUCLEIC ACID ATOMS       : 0
```

```
REMARK   3   HETEROGEN ATOMS          : 21
REMARK   3   SOLVENT ATOMS            : 39
REMARK   3
REMARK   3  B VALUES.
REMARK   3   FROM WILSON PLOT           (A**2) : 49.20
REMARK   3   MEAN B VALUE      (OVERALL, A**2) : 60.40
REMARK   3   OVERALL ANISOTROPIC B VALUE.
REMARK   3    B11 (A**2) : NULL
REMARK   3    B22 (A**2) : NULL
REMARK   3    B33 (A**2) : NULL
REMARK   3    B12 (A**2) : NULL
REMARK   3    B13 (A**2) : NULL
REMARK   3    B23 (A**2) : NULL
REMARK   3
REMARK   3  ESTIMATED OVERALL COORDINATE ERROR.
REMARK   3   ESU BASED ON R VALUE                            (A): NULL
REMARK   3   ESU BASED ON FREE R VALUE                       (A): NULL
REMARK   3   ESU BASED ON MAXIMUM LIKELIHOOD                 (A): NULL
REMARK   3   ESU FOR B VALUES BASED ON MAXIMUM LIKELIHOOD (A**2): 3.640
REMARK   3
REMARK   3  RMS DEVIATIONS FROM IDEAL VALUES.
REMARK   3   DISTANCE RESTRAINTS.                     RMS    SIGMA
REMARK   3    BOND LENGTH                    (A) : 0.015 ; NULL
REMARK   3    ANGLE DISTANCE                 (A) : 0.030 ; NULL
REMARK   3    INTRAPLANAR 1-4 DISTANCE       (A) : NULL  ; NULL
REMARK   3    H-BOND OR METAL COORDINATION   (A) : NULL  ; NULL
REMARK   3
REMARK   3   PLANE RESTRAINT                 (A) : NULL  ; NULL
REMARK   3   CHIRAL-CENTER RESTRAINT      (A**3) : NULL  ; NULL
REMARK   3
REMARK   3   NON-BONDED CONTACT RESTRAINTS.
REMARK   3    SINGLE TORSION                 (A) : NULL  ; NULL
REMARK   3    MULTIPLE TORSION               (A) : NULL  ; NULL
REMARK   3    H-BOND (X...Y)                 (A) : NULL  ; NULL
REMARK   3    H-BOND (X-H...Y)               (A) : NULL  ; NULL
REMARK   3
REMARK   3   CONFORMATIONAL TORSION ANGLE RESTRAINTS.
REMARK   3    SPECIFIED                (DEGREES) : NULL  ; NULL
REMARK   3    PLANAR                   (DEGREES) : NULL  ; NULL
REMARK   3    STAGGERED                (DEGREES) : NULL  ; NULL
REMARK   3    TRANSVERSE               (DEGREES) : NULL  ; NULL
REMARK   3
REMARK   3  ISOTROPIC THERMAL FACTOR RESTRAINTS.    RMS    SIGMA
REMARK   3   MAIN-CHAIN BOND              (A**2) : NULL  ; NULL
REMARK   3   MAIN-CHAIN ANGLE             (A**2) : NULL  ; NULL
REMARK   3   SIDE-CHAIN BOND              (A**2) : NULL  ; NULL
REMARK   3   SIDE-CHAIN ANGLE             (A**2) : NULL  ; NULL
REMARK   3
REMARK   3  OTHER REFINEMENT REMARKS:  DISORDERED REGIONS ARE FROM GLY22-
REMARK   3  GLY28, GLY65-GLU68 AND ILE147-PRO147 WERE MODELED
REMARK   3  STEREOCHEMICALLY
REMARK   4
REMARK   4 1BXW COMPLIES WITH FORMAT V. 3.15, 01-DEC-08
REMARK 100
REMARK 100 THIS ENTRY HAS BEEN PROCESSED BY RCSB ON 19-AUG-99.
REMARK 100 THE RCSB ID CODE IS RCSB008140.
REMARK 200
REMARK 200 EXPERIMENTAL DETAILS
REMARK 200  EXPERIMENT TYPE                : X-RAY DIFFRACTION
REMARK 200  DATE OF DATA COLLECTION        : 15-JAN-98
REMARK 200  TEMPERATURE           (KELVIN) : 298
```

```
REMARK 200  PH                             : 5.0
REMARK 200  NUMBER OF CRYSTALS USED        : 1
REMARK 200
REMARK 200  SYNCHROTRON              (Y/N) : N
REMARK 200  RADIATION SOURCE               : ROTATING ANODE
REMARK 200  BEAMLINE                       : NULL
REMARK 200  X-RAY GENERATOR MODEL          : RIGAKU RU200
REMARK 200  MONOCHROMATIC OR LAUE    (M/L) : M
REMARK 200  WAVELENGTH OR RANGE        (A) : 1.5418
REMARK 200  MONOCHROMATOR                  : NI FILTER
REMARK 200  OPTICS                         : NULL
REMARK 200
REMARK 200  DETECTOR TYPE                  : AREA DETECTOR
REMARK 200  DETECTOR MANUFACTURER          : SIEMENS
REMARK 200  INTENSITY-INTEGRATION SOFTWARE : XDS
REMARK 200  DATA SCALING SOFTWARE          : CCP4 (SCALA)
REMARK 200
REMARK 200  NUMBER OF UNIQUE REFLECTIONS   : 8328
REMARK 200  RESOLUTION RANGE HIGH      (A) : 2.500
REMARK 200  RESOLUTION RANGE LOW       (A) : 50.000
REMARK 200  REJECTION CRITERIA  (SIGMA(I)) : NULL
REMARK 200
REMARK 200 OVERALL.
REMARK 200  COMPLETENESS FOR RANGE     (%) : 89.0
REMARK 200  DATA REDUNDANCY                : 2.100
REMARK 200  R MERGE                    (I) : NULL
REMARK 200  R SYM                      (I) : 0.02800
REMARK 200  <I/SIGMA(I)> FOR THE DATA SET  : 16.8000
REMARK 200
REMARK 200 IN THE HIGHEST RESOLUTION SHELL.
REMARK 200  HIGHEST RESOLUTION SHELL, RANGE HIGH (A) : 2.50
REMARK 200  HIGHEST RESOLUTION SHELL, RANGE LOW  (A) : 2.64
REMARK 200  COMPLETENESS FOR SHELL     (%) : 53.0
REMARK 200  DATA REDUNDANCY IN SHELL       : 1.20
REMARK 200  R MERGE FOR SHELL          (I) : NULL
REMARK 200  R SYM FOR SHELL            (I) : 0.11000
REMARK 200  <I/SIGMA(I)> FOR SHELL         : 6.600
REMARK 200
REMARK 200 DIFFRACTION PROTOCOL: SINGLE WAVELENGTH
REMARK 200 METHOD USED TO DETERMINE THE STRUCTURE: MIRAS
REMARK 200 SOFTWARE USED: SHARP
REMARK 200 STARTING MODEL: NULL
REMARK 200
REMARK 200 REMARK: NULL
REMARK 280
REMARK 280 CRYSTAL
REMARK 280 SOLVENT CONTENT, VS   (%): 66.70
REMARK 280 MATTHEWS COEFFICIENT, VM (ANGSTROMS**3/DA): 3.70
REMARK 280
REMARK 280 CRYSTALLIZATION CONDITIONS: 10 % PEG-8000 10 % MPD 0.05 M
REMARK 280  POTASSIUM PHOSPHATE PH 5.0
REMARK 290
REMARK 290 CRYSTALLOGRAPHIC SYMMETRY
REMARK 290 SYMMETRY OPERATORS FOR SPACE GROUP: C 1 2 1
REMARK 290
REMARK 290      SYMOP   SYMMETRY
REMARK 290     NNNMMM   OPERATOR
REMARK 290       1555   X,Y,Z
REMARK 290       2555   -X,Y,-Z
REMARK 290       3555   X+1/2,Y+1/2,Z
REMARK 290       4555   -X+1/2,Y+1/2,-Z
```

```
REMARK 290
REMARK 290     WHERE NNN -> OPERATOR NUMBER
REMARK 290           MMM -> TRANSLATION VECTOR
REMARK 290
REMARK 290 CRYSTALLOGRAPHIC SYMMETRY TRANSFORMATIONS
REMARK 290 THE FOLLOWING TRANSFORMATIONS OPERATE ON THE ATOM/HETATM
REMARK 290 RECORDS IN THIS ENTRY TO PRODUCE CRYSTALLOGRAPHICALLY
REMARK 290 RELATED MOLECULES.
REMARK 290   SMTRY1   1  1.000000  0.000000  0.000000        0.00000
REMARK 290   SMTRY2   1  0.000000  1.000000  0.000000        0.00000
REMARK 290   SMTRY3   1  0.000000  0.000000  1.000000        0.00000
REMARK 290   SMTRY1   2 -1.000000  0.000000  0.000000        0.00000
REMARK 290   SMTRY2   2  0.000000  1.000000  0.000000        0.00000
REMARK 290   SMTRY3   2  0.000000  0.000000 -1.000000        0.00000
REMARK 290   SMTRY1   3  1.000000  0.000000  0.000000       34.59000
REMARK 290   SMTRY2   3  0.000000  1.000000  0.000000       38.97500
REMARK 290   SMTRY3   3  0.000000  0.000000  1.000000        0.00000
REMARK 290   SMTRY1   4 -1.000000  0.000000  0.000000       34.59000
REMARK 290   SMTRY2   4  0.000000  1.000000  0.000000       38.97500
REMARK 290   SMTRY3   4  0.000000  0.000000 -1.000000        0.00000
REMARK 290
REMARK 290 REMARK: NULL
REMARK 300
REMARK 300 BIOMOLECULE: 1
REMARK 300 SEE REMARK 350 FOR THE AUTHOR PROVIDED AND/OR PROGRAM
REMARK 300 GENERATED ASSEMBLY INFORMATION FOR THE STRUCTURE IN
REMARK 300 THIS ENTRY. THE REMARK MAY ALSO PROVIDE INFORMATION ON
REMARK 300 BURIED SURFACE AREA.
REMARK 350
REMARK 350 COORDINATES FOR A COMPLETE MULTIMER REPRESENTING THE KNOWN
REMARK 350 BIOLOGICALLY SIGNIFICANT OLIGOMERIZATION STATE OF THE
REMARK 350 MOLECULE CAN BE GENERATED BY APPLYING BIOMT TRANSFORMATIONS
REMARK 350 GIVEN BELOW.  BOTH NON-CRYSTALLOGRAPHIC AND
REMARK 350 CRYSTALLOGRAPHIC OPERATIONS ARE GIVEN.
REMARK 350
REMARK 350 BIOMOLECULE: 1
REMARK 350 AUTHOR DETERMINED BIOLOGICAL UNIT: MONOMERIC
REMARK 350 APPLY THE FOLLOWING TO CHAINS: A
REMARK 350   BIOMT1   1  1.000000  0.000000  0.000000        0.00000
REMARK 350   BIOMT2   1  0.000000  1.000000  0.000000        0.00000
REMARK 350   BIOMT3   1  0.000000  0.000000  1.000000        0.00000
REMARK 470
REMARK 470 MISSING ATOM
REMARK 470 THE FOLLOWING RESIDUES HAVE MISSING ATOMS(M=MODEL NUMBER;
REMARK 470 RES=RESIDUE NAME; C=CHAIN IDENTIFIER; SSEQ=SEQUENCE NUMBER;
REMARK 470 I=INSERTION CODE):
REMARK 470   M RES CSSEQI  ATOMS
REMARK 470     HIS A  31    CG   ND1  CD2  CE1  NE2
REMARK 475
REMARK 475 ZERO OCCUPANCY RESIDUES
REMARK 475 THE FOLLOWING RESIDUES WERE MODELED WITH ZERO OCCUPANCY.
REMARK 475 THE LOCATION AND PROPERTIES OF THESE RESIDUES MAY NOT
REMARK 475 BE RELIABLE. (M=MODEL NUMBER; RES=RESIDUE NAME;
REMARK 475 C=CHAIN IDENTIFIER; SSEQ=SEQUENCE NUMBER; I=INSERTION CODE)
REMARK 475   M RES C SSEQI
REMARK 475     GLY A   22
REMARK 475     LEU A   23
REMARK 475     ILE A   24
REMARK 475     ASN A   25
REMARK 475     ASN A   26
REMARK 475     ASN A   27
```

```
REMARK 475     GLY A   28
REMARK 475     GLY A   65
REMARK 475     SER A   66
REMARK 475     VAL A   67
REMARK 475     GLU A   68
REMARK 475     ILE A  147
REMARK 475     GLY A  148
REMARK 475     ASP A  149
REMARK 475     ALA A  150
REMARK 475     HIS A  151
REMARK 475     THR A  152
REMARK 475     ILE A  153
REMARK 475     GLY A  154
REMARK 475     THR A  155
REMARK 475     ARG A  156
REMARK 475     PRO A  157
REMARK 480
REMARK 480 ZERO OCCUPANCY ATOM
REMARK 480 THE FOLLOWING RESIDUES HAVE ATOMS MODELED WITH ZERO
REMARK 480 OCCUPANCY. THE LOCATION AND PROPERTIES OF THESE ATOMS
REMARK 480 MAY NOT BE RELIABLE. (M=MODEL NUMBER; RES=RESIDUE NAME;
REMARK 480 C=CHAIN IDENTIFIER; SSEQ=SEQUENCE NUMBER; I=INSERTION CODE):
REMARK 480   M RES C SSEQI ATOMS
REMARK 480     LYS A   64   CB    CG    CD    CE    NZ
REMARK 500
REMARK 500 GEOMETRY AND STEREOCHEMISTRY
REMARK 500 SUBTOPIC: CLOSE CONTACTS
REMARK 500
REMARK 500 THE FOLLOWING ATOMS THAT ARE RELATED BY CRYSTALLOGRAPHIC
REMARK 500 SYMMETRY ARE IN CLOSE CONTACT.  AN ATOM LOCATED WITHIN 0.15
REMARK 500 ANGSTROMS OF A SYMMETRY RELATED ATOM IS ASSUMED TO BE ON A
REMARK 500 SPECIAL POSITION AND IS, THEREFORE, LISTED IN REMARK 375
REMARK 500 INSTEAD OF REMARK 500.  ATOMS WITH NON-BLANK ALTERNATE
REMARK 500 LOCATION INDICATORS ARE NOT INCLUDED IN THE CALCULATIONS.
REMARK 500
REMARK 500 DISTANCE CUTOFF:
REMARK 500 2.2 ANGSTROMS FOR CONTACTS NOT INVOLVING HYDROGEN ATOMS
REMARK 500 1.6 ANGSTROMS FOR CONTACTS INVOLVING HYDROGEN ATOMS
REMARK 500
REMARK 500  ATM1  RES C  SSEQI   ATM2  RES C  SSEQI  SSYMOP   DISTANCE
REMARK 500   OD1  ASN A    26     CA   PRO A    29     2556     1.44
REMARK 500   OD1  ASN A    26     C    PRO A    29     2556     1.68
REMARK 500   OD1  ASN A    26     N    PRO A    29     2556     1.72
REMARK 500   OD1  ASN A     5     CD1  ILE A   147     2657     2.03
REMARK 500   OD1  ASN A    26     O    PRO A    29     2556     2.08
REMARK 500   CG   ASN A    26     N    PRO A    29     2556     2.11
REMARK 500
REMARK 500 REMARK: NULL
REMARK 500
REMARK 500 GEOMETRY AND STEREOCHEMISTRY
REMARK 500 SUBTOPIC: COVALENT BOND LENGTHS
REMARK 500
REMARK 500 THE STEREOCHEMICAL PARAMETERS OF THE FOLLOWING RESIDUES
REMARK 500 HAVE VALUES WHICH DEVIATE FROM EXPECTED VALUES BY MORE
REMARK 500 THAN 6*RMSD (M=MODEL NUMBER; RES=RESIDUE NAME; C=CHAIN
REMARK 500 IDENTIFIER; SSEQ=SEQUENCE NUMBER; I=INSERTION CODE).
REMARK 500
REMARK 500 STANDARD TABLE:
REMARK 500 FORMAT: (10X,I3,1X,2(A3,1X,A1,I4,A1,1X,A4,3X),1X,F6.3)
REMARK 500
REMARK 500 EXPECTED VALUES PROTEIN: ENGH AND HUBER, 1999
```

```
REMARK 500 EXPECTED VALUES NUCLEIC ACID: CLOWNEY ET AL 1996
REMARK 500
REMARK 500  M RES CSSEQI ATM1   RES CSSEQI ATM2   DEVIATION
REMARK 500    GLY A  28   C     PRO A  29   N       0.125
REMARK 500    GLY A 148   N     GLY A 148   CA      0.090
REMARK 500    ARG A 156   CA    ARG A 156   C       0.206
REMARK 500    PRO A 157   N     PRO A 157   CA     -0.251
REMARK 500    PRO A 157   CD    PRO A 157   N      -0.368
REMARK 500    PRO A 157   CA    PRO A 157   C      -0.164
REMARK 500
REMARK 500 REMARK: NULL
REMARK 500
REMARK 500 GEOMETRY AND STEREOCHEMISTRY
REMARK 500 SUBTOPIC: COVALENT BOND ANGLES
REMARK 500
REMARK 500 THE STEREOCHEMICAL PARAMETERS OF THE FOLLOWING RESIDUES
REMARK 500 HAVE VALUES WHICH DEVIATE FROM EXPECTED VALUES BY MORE
REMARK 500 THAN 6*RMSD (M=MODEL NUMBER; RES=RESIDUE NAME; C=CHAIN
REMARK 500 IDENTIFIER; SSEQ=SEQUENCE NUMBER; I=INSERTION CODE).
REMARK 500
REMARK 500 STANDARD TABLE:
REMARK 500 FORMAT: (10X,I3,1X,A3,1X,A1,I4,A1,3(1X,A4,2X),12X,F5.1)
REMARK 500
REMARK 500 EXPECTED VALUES PROTEIN: ENGH AND HUBER, 1999
REMARK 500 EXPECTED VALUES NUCLEIC ACID: CLOWNEY ET AL 1996
REMARK 500
REMARK 500  M RES CSSEQI ATM1   ATM2   ATM3
REMARK 500    ASP A   4   CA  -  C   -  N   ANGL. DEV. =  18.3 DEGREES
REMARK 500    GLY A  22   O   -  C   -  N   ANGL. DEV. =  11.3 DEGREES
REMARK 500    ASN A  25   C   -  N   -  CA  ANGL. DEV. = -16.5 DEGREES
REMARK 500    ASN A  25   CA  -  C   -  N   ANGL. DEV. = -14.2 DEGREES
REMARK 500    ASN A  25   O   -  C   -  N   ANGL. DEV. =  14.8 DEGREES
REMARK 500    GLY A  28   O   -  C   -  N   ANGL. DEV. = -11.6 DEGREES
REMARK 500    ARG A  60   CD  -  NE  -  CZ  ANGL. DEV. =   9.8 DEGREES
REMARK 500    ARG A  60   NE  -  CZ  -  NH1 ANGL. DEV. =   3.7 DEGREES
REMARK 500    ARG A  60   NE  -  CZ  -  NH2 ANGL. DEV. =  -3.3 DEGREES
REMARK 500    GLU A  68   C   -  N   -  CA  ANGL. DEV. = -16.3 DEGREES
REMARK 500    GLN A  75   CB  -  CA  -  C   ANGL. DEV. =  13.1 DEGREES
REMARK 500    ASP A  90   CB  -  CG  -  OD1 ANGL. DEV. =   5.7 DEGREES
REMARK 500    SER A 120   N   -  CA  -  CB  ANGL. DEV. =  -9.2 DEGREES
REMARK 500    VAL A 122   CB  -  CA  -  C   ANGL. DEV. = -12.2 DEGREES
REMARK 500    ILE A 135   CA  -  CB  -  CG2 ANGL. DEV. =  15.6 DEGREES
REMARK 500    ARG A 138   CA  -  CB  -  CG  ANGL. DEV. =  14.8 DEGREES
REMARK 500    ARG A 138   CD  -  NE  -  CZ  ANGL. DEV. =  10.7 DEGREES
REMARK 500    ARG A 138   NE  -  CZ  -  NH2 ANGL. DEV. =  -3.3 DEGREES
REMARK 500    HIS A 151   CB  -  CA  -  C   ANGL. DEV. = -38.2 DEGREES
REMARK 500    ALA A 150   CA  -  C   -  N   ANGL. DEV. = -16.2 DEGREES
REMARK 500    ALA A 150   O   -  C   -  N   ANGL. DEV. =  17.6 DEGREES
REMARK 500    HIS A 151   CA  -  C   -  N   ANGL. DEV. = -38.2 DEGREES
REMARK 500    HIS A 151   O   -  C   -  N   ANGL. DEV. =  43.4 DEGREES
REMARK 500    THR A 152   C   -  N   -  CA  ANGL. DEV. =  30.0 DEGREES
REMARK 500    ARG A 156   CB  -  CA  -  C   ANGL. DEV. =  12.4 DEGREES
REMARK 500    ARG A 156   N   -  CA  -  CB  ANGL. DEV. = -15.9 DEGREES
REMARK 500    ARG A 156   NH1 -  CZ  -  NH2 ANGL. DEV. =  -6.6 DEGREES
REMARK 500    ARG A 156   NE  -  CZ  -  NH2 ANGL. DEV. =   3.7 DEGREES
REMARK 500    THR A 155   O   -  C   -  N   ANGL. DEV. = -14.6 DEGREES
REMARK 500    ARG A 156   C   -  N   -  CA  ANGL. DEV. = -16.4 DEGREES
REMARK 500    PRO A 157   CA  -  N   -  CD  ANGL. DEV. = -25.7 DEGREES
REMARK 500    PRO A 157   N   -  CA  -  CB  ANGL. DEV. = -25.4 DEGREES
REMARK 500    PRO A 157   CB  -  CG  -  CD  ANGL. DEV. = -24.6 DEGREES
REMARK 500    PRO A 157   N   -  CD  -  CG  ANGL. DEV. = -33.9 DEGREES
```

```
REMARK 500    PRO A 157   N   -  CA  -  C   ANGL. DEV. =  19.1 DEGREES
REMARK 500    PRO A 157   CA  -  C   -  O   ANGL. DEV. = -17.2 DEGREES
REMARK 500    PRO A 157   C   -  N   -  CA  ANGL. DEV. = -16.9 DEGREES
REMARK 500
REMARK 500 REMARK: NULL
REMARK 500
REMARK 500 GEOMETRY AND STEREOCHEMISTRY
REMARK 500 SUBTOPIC: TORSION ANGLES
REMARK 500
REMARK 500 TORSION ANGLES OUTSIDE THE EXPECTED RAMACHANDRAN REGIONS:
REMARK 500 (M=MODEL NUMBER; RES=RESIDUE NAME; C=CHAIN IDENTIFIER;
REMARK 500 SSEQ=SEQUENCE NUMBER; I=INSERTION CODE).
REMARK 500
REMARK 500 STANDARD TABLE:
REMARK 500 FORMAT:(10X,I3,1X,A3,1X,A1,I4,A1,4X,F7.2,3X,F7.2)
REMARK 500
REMARK 500 EXPECTED VALUES: GJ KLEYWEGT AND TA JONES (1996). PHI/PSI-
REMARK 500 CHOLOGY: RAMACHANDRAN REVISITED. STRUCTURE 4, 1395 - 1400
REMARK 500
REMARK 500  M RES CSSEQI        PSI       PHI
REMARK 500    ASN A   5       57.62   -113.00
REMARK 500    TYR A  18      120.18    166.90
REMARK 500    ASP A  20     -140.62   -156.38
REMARK 500    LEU A  23      150.39     68.74
REMARK 500    ASN A  25      -90.55     -9.26
REMARK 500    ASN A  26      121.49    -29.94
REMARK 500    HIS A  31      175.58    173.40
REMARK 500    TYR A  63      102.76   -169.58
REMARK 500    SER A  66       52.30    128.49
REMARK 500    VAL A  67       90.53     49.93
REMARK 500    VAL A 110      -72.06    -67.54
REMARK 500    ALA A 150     -164.72    173.09
REMARK 500    HIS A 151      -97.21     35.47
REMARK 500    THR A 152     -139.61   -128.76
REMARK 500    THR A 155     -137.22   -149.83
REMARK 500    ARG A 156     -162.43   -178.66
REMARK 500
REMARK 500 REMARK: NULL
REMARK 500
REMARK 500 GEOMETRY AND STEREOCHEMISTRY
REMARK 500 SUBTOPIC: NON-CIS, NON-TRANS
REMARK 500
REMARK 500 THE FOLLOWING PEPTIDE BONDS DEVIATE SIGNIFICANTLY FROM BOTH
REMARK 500 CIS AND TRANS CONFORMATION.  CIS BONDS, IF ANY, ARE LISTED
REMARK 500 ON CISPEP RECORDS.  TRANS IS DEFINED AS 180 +/- 30 AND
REMARK 500 CIS IS DEFINED AS 0 +/- 30 DEGREES.
REMARK 500                                    MODEL     OMEGA
REMARK 500 ARG A  156     PRO A  157                  -147.47
REMARK 500
REMARK 500 REMARK: NULL
REMARK 500
REMARK 500 GEOMETRY AND STEREOCHEMISTRY
REMARK 500 SUBTOPIC: MAIN CHAIN PLANARITY
REMARK 500
REMARK 500 THE FOLLOWING RESIDUES HAVE A PSEUDO PLANARITY
REMARK 500 TORSION, C(I) - CA(I) - N(I+1) - O(I), GREATER
REMARK 500 10.0 DEGREES. (M=MODEL NUMBER; RES=RESIDUE NAME;
REMARK 500 C=CHAIN IDENTIFIER; SSEQ=SEQUENCE NUMBER;
REMARK 500 I=INSERTION CODE).
REMARK 500
REMARK 500  M RES CSSEQI        ANGLE
```

```
REMARK 500    ARG A 156        -11.76
REMARK 500
REMARK 500 REMARK: NULL
REMARK 800
REMARK 800 SITE
REMARK 800 SITE_IDENTIFIER: AC1
REMARK 800 EVIDENCE_CODE: SOFTWARE
REMARK 800 SITE_DESCRIPTION: BINDING SITE FOR RESIDUE C8E A 172
DBREF  1BXW A    0   171  UNP    P0A910   OMPA_ECOLI      21    192
SEQADV 1BXW MET A    0  UNP  P0A910    ALA    21 SEE REMARK 999
SEQADV 1BXW LEU A   23  UNP  P0A910    PHE    44 MUTATION
SEQADV 1BXW LYS A   34  UNP  P0A910    GLN    55 MUTATION
SEQADV 1BXW TYR A  107  UNP  P0A910    LYS   128 MUTATION
SEQRES   1 A  172  MET ALA PRO LYS ASP ASN THR TRP TYR THR GLY ALA LYS
SEQRES   2 A  172  LEU GLY TRP SER GLN TYR HIS ASP THR GLY LEU ILE ASN
SEQRES   3 A  172  ASN ASN GLY PRO THR HIS GLU ASN LYS LEU GLY ALA GLY
SEQRES   4 A  172  ALA PHE GLY GLY TYR GLN VAL ASN PRO TYR VAL GLY PHE
SEQRES   5 A  172  GLU MET GLY TYR ASP TRP LEU GLY ARG MET PRO TYR LYS
SEQRES   6 A  172  GLY SER VAL GLU ASN GLY ALA TYR LYS ALA GLN GLY VAL
SEQRES   7 A  172  GLN LEU THR ALA LYS LEU GLY TYR PRO ILE THR ASP ASP
SEQRES   8 A  172  LEU ASP ILE TYR THR ARG LEU GLY GLY MET VAL TRP ARG
SEQRES   9 A  172  ALA ASP THR TYR SER ASN VAL TYR GLY LYS ASN HIS ASP
SEQRES  10 A  172  THR GLY VAL SER PRO VAL PHE ALA GLY GLY VAL GLU TYR
SEQRES  11 A  172  ALA ILE THR PRO GLU ILE ALA THR ARG LEU GLU TYR GLN
SEQRES  12 A  172  TRP THR ASN ASN ILE GLY ASP ALA HIS THR ILE GLY THR
SEQRES  13 A  172  ARG PRO ASP ASN GLY MET LEU SER LEU GLY VAL SER TYR
SEQRES  14 A  172  ARG PHE GLY
HET    C8E  A 172      21
HETNAM     C8E (HYDROXYETHYLOXY)TRI(ETHYLOXY)OCTANE
FORMUL   2  C8E    C16 H34 O5
FORMUL   3  HOH   *39(H2 O)
SHEET    1  S1 1 THR A   6  SER A  16  0
SHEET    1  S2 1 LYS A  34  VAL A  45  0
SHEET    1  S3 1 VAL A  49  ARG A  60  0
SHEET    1  S4 1 TYR A  72  PRO A  86  0
SHEET    1  S5 1 LEU A  91  THR A 106  0
SHEET    1  S6 1 ASN A 114  ALA A 130  0
SHEET    1  S7 1 ILE A 135  TRP A 143  0
SHEET    1  S8 1 MET A 161  PHE A 170  0
LINK         OD2 ASP A 149                 C17 C8E A 172     2657   1555  1.24
LINK         CB  ASP A 149                 C17 C8E A 172     2657   1555  1.88
LINK         OD1 ASP A 149                 C17 C8E A 172     2657   1555  1.59
LINK         OD2 ASP A 149                 O18 C8E A 172     2657   1555  1.96
LINK         CA  ASP A 149                 O18 C8E A 172     2657   1555  1.92
LINK         CB  ASP A 149                 O18 C8E A 172     2657   1555  1.18
LINK         OD1 ASP A 149                 O18 C8E A 172     2657   1555  1.38
LINK         N   ASP A 149                 C19 C8E A 172     2657   1555  1.68
LINK         C   ASP A 149                 C19 C8E A 172     2657   1555  1.87
LINK         CA  ASP A 149                 C20 C8E A 172     2657   1555  1.45
LINK         C   ASP A 149                 C20 C8E A 172     2657   1555  1.22
LINK         CB  ASP A 149                 C20 C8E A 172     2657   1555  2.02
LINK         N   ALA A 150                 O21 C8E A 172     2657   1555  1.70
LINK         N   ALA A 150                 C20 C8E A 172     2657   1555  1.34
SITE     1 AC1  4 TYR A  43  PHE A  51  LEU A  79  GLY A  99
CRYST1   69.180   77.950   50.930  90.00  91.52  90.00 C 1 2 1       4
ORIGX1      1.000000  0.000000  0.000000        0.00000
ORIGX2      0.000000  1.000000  0.000000        0.00000
ORIGX3      0.000000  0.000000  1.000000        0.00000
SCALE1      0.014455  0.000000  0.000383        0.00000
SCALE2      0.000000  0.012829  0.000000        0.00000
SCALE3      0.000000  0.000000  0.019642        0.00000
```

Επεξήγηση πεδίων μιας εγγραφής PDB

HEADER: *Περιέχει ένα τετραψήφιο κωδικό για την αναγνώριση της εγγραφής στην PDB, μια γενική ταξινόμηση του μακρομορίου καθώς και την ημερομηνία κατάθεσης της δομής στην Protein Data Bank.*

TITLE: *Τίτλος που περιλαμβάνει συνήθως τα περιεχόμενα της εγγραφής, τι είδους πειραματική διαδικασία χρησιμοποιήθηκε, ύπαρξη μεταλλάξεων. Επιτρέπει στον ερεύνητη που κατέθεσε τη δομή να καταδείξει τη σημαντικότητα της εργασίας αυτής.*

COMPOUND: *Το πεδίο compound περιέχει πληροφορίες για το μακρομόριο που αναφέρεται στη δομή καθώς και τα άλλα μόρια (μικρές οργανικές ενώσεις, μέταλλα) με τα οποία έχει τυχόν συμπλοκοποιηθεί.*

SOURCE: *Βιολογική προέλευση του μακρομορίου που αναφέρεται στην εγγραφή.*

KEYWDS: *Χαρακτηριστικές λέξεις-κλειδιά για τον χαρακτηρισμό της εγγραφής.*

EXPDTA: *Πειραματική τεχνική για τον προσδιορισμό της δομής (X-Ray Crystallography/NMR/Theoretical Model).*

AUTHOR: *Λίστα με τα ονόματα των ερευνητών που συμμετείχαν στον προσδιορισμό της δομής.*

JRNL: *Πρωταρχική βιβλιογραφική αναφορά η οποία αναφερέται στον προσδιορισμό της δομής που αναφέρεται στην συγκεκριμένη εγγραφή.*

REMARK: *Το πεδίο REMARK περιλαμβάνει μια σειρά από πληροφορίες σχετικές με την κατατεθειμένη δομή.*

Καταρχήν περιέχει βιβλιογραφικές αναφορές που σχετίζονται άμεσα με το προς μελέτη μακρομόριο.

Στο πεδίο REMARK περιλαμβάνονται και στοιχεία σχετικά με την πειραματική διαδικασία που ακολουθήθηκε για την λύση της δομής όπως είναι τα προγράμματα που χρησιμοποιήθηκαν, οι τιμές διαφόρων δεικτών, γενικά πληροφορίες που αποδεικνύουν την ορθότητα της δομής.

SEQRES: *Περιέχει την αλληλουχία του προς μελέτη μακρομορίου. Για τις πρωτεΐνες ακολουθείται ο κώδικας των 3 γραμμάτων.*

HET: *Αναφέρεται στα μόρια (ετεροάτομα) που δεν είναι αμινοξέα ή νουκλεοτίδια. Αυτά μπορεί να είναι προσθετικές ομάδες και ιόντα για τα οποία έχουν προσδιοριστεί οι συντεταγμένες τους. Τα στοιχεία που δίνονται για αυτά είναι ένας κωδικός για να διευκρινίζονται σε σχέση με τα άλλα κατάλοιπα της εγγραφής, η αρίθμηση που έχουν μέσα στο αρχείο των συντεταγμένων και τέλος ο αριθμός των ατόμων από τα οποία αποτελούνται.*

HETNAM: *Ονοματολογία των καταλοίπων που περιέχονται στο πεδίο HET.*

FORMUL: *Μοριακός τύπος των καταλοίπων που αναφέρονται στο πεδίο HET.*

HELIX: *Τμήματα της αλληλουχίας που έχουν ελικοειδή δομή.*

SHEET: *Τμήματα της αλληλουχίας που έχουν εκτεταμένη δομή.*

CRYST1: *Περιέχει τις παραμέτρους μοναδιαίας κυψελλίδας και την ομάδα συμμετρίας χώρου.*

ORIGXn(n=1..3): *Πίνακας Μετατροπής από σύστημα ορθογωνίων συντεταγμένων στις συντεταγμένες που κατατέθηκαν αρχικά στην PDB.*

SCALEn: *Πίνακας Μετατροπής από σύστημα ορθογωνίων συντεταγμένων στις κρυσταλλογραφικές συντεταγμένες.*

ATOM: *Περιέχει τις συντεταγμένες των ατόμων στους άξονες X, Y, Z. Περιλαμβάνει επίσης και άλλα στοιχεία όπως τα άτομα για τα οποία αναφέρονται οι συντεταγμένες και σε ποια κατάλοιπα ανήκουν. Πρέπει να σημειωθεί ότι κάθε είδους*

δεδομένο που περιέχεται στο πεδίο ATOM είναι τοποθετημένο σε καθορισμένες θέσεις (στήλες) της εγγραφής όπως αυτές παρουσιάζονται παρακάτω:

ΣΤΗΛΕΣ Περιεχόμενα κάθε στήλης

---

1 - 6 *"ATOM " δηλώνει ότι πρόκειται για το πεδίο ATOM.*

7 - 11 *Αύξων αριθμός του ατόμου.*

13 - 16 *Τύπος ατόμου.*

18 - 20 *Όνομα καταλοίπου. Για τα αμινοξέα ακολουθείται ο κώδικας των 3 γραμμάτων.*

22 *(chainID) Χαρακτήρας που ταυτοποιεί την αλυσίδα, αν περιέχονται περισσότερες από μια στην εγγραφή.*

23 - 26 *Αρίθμηση του καταλοίπου στην αλυσίδα*

31 - 38 *x Συντεταγμένες ατόμου (σε Angstroms) στον άξονα X σε τρισορθογώνιο σύστημα αξόνων.*

39 - 46 *y Συντεταγμένες ατόμου (σε Angstroms) στον άξονα Y σε τρισορθογώνιο σύστημα αξόνων.*

47 - 54 *z Συντεταγμένες ατόμου (σε Angstroms) στον άξονα Z σε τρισορθογώνιο σύστημα αξόνων.*

55 - 60 *Συντελεστής κατάληψης(occupancy)*

61 - 66 *Παράγοντας θερμοκρασίας(Temperature factor)*

77 - 78 *Σύμβολο του ατόμου.*

79 - 80 *Φορτίο του ατόμου (Αν υπάρχει).*

TER: *Το πεδίο TER δηλώνει το τέλος της παράθεσης των ατόμων που απαρτίζουν μια αλυσίδα.*

HETATM: *Συντεταγμένες των ετεροατόμων. Η μορφοποίηση παρουσίασης τους ακολουθεί του ίδιους κανόνες με το πεδίο ATOM.*

CONECT: *Το πεδίο CONECT καθορίζει τα άτομα τα οποία συμμετέχουν στον σχηματισμό δεσμών. Κάθε άτομο συμβολίζεται με την αρίθμηση του όπως είναι καθορισμένη στα πεδία ATOM.*

MASTER: *Αποτελεί ένα πεδίο που χρησιμοποιείται για μια απλή οργάνωση της εγγραφής. Πρόκειται για μια σειρά από αριθμούς που δεν είναι τίποτε άλλο από το άθροισμα των γραμμών για συγκεκριμένα πεδία της εγγραφής.*

END: *Υποδηλώνει τη λήξη της εγγραφής.*

Βιβλιογραφία

Akiva, E., Brown, S., Almonacid, D. E., Barber, A. E., 2nd, Custer, A. F., Hicks, M. A., . . . Babbitt, P. C. (2014). The Structure-Function Linkage Database. *Nucleic acids research, 42*(Database issue), D521-530.

Akondi, K. B., Muttenthaler, M., Dutertre, S., Kaas, Q., Craik, D. J., Lewis, R. J., & Alewood, P. F. (2014). Discovery, synthesis, and structure-activity relationships of conotoxins. *Chemical reviews, 114*(11), 5815-5847.

Alexander, S. P., Benson, H. E., Faccenda, E., Pawson, A. J., Sharman, J. L., McGrath, J. C., . . . Zimmermann, M. (2013). The Concise Guide to PHARMACOLOGY 2013/14: overview. *British journal of pharmacology, 170*(8), 1449-1458.

Alexander, S. P., Mathie, A., & Peters, J. A. (2011). Guide to Receptors and Channels (GRAC), 5th edition. *British journal of pharmacology, 164 Suppl 1*, S1-324.

Almonacid, D. E., & Babbitt, P. C. (2011). Toward mechanistic classification of enzyme functions. *Current opinion in chemical biology, 15*(3), 435-442.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research, 25*(17), 3389-3402.

Ananiadou, S., Kell, D. B., & Tsujii, J. (2006). Text mining and its potential applications in systems biology. *Trends Biotechnol, 24*(12), 571-579.

Andreeva, A., Howorth, D., Brenner, S. E., Hubbard, T. J., Chothia, C., & Murzin, A. G. (2004). SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res, 32*(Database issue), D226-229.

Atkinson, H. J., Morris, J. H., Ferrin, T. E., & Babbitt, P. C. (2009). Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. *PloS one, 4*(2), e4345.

Babbitt, P. C., & Gerlt, J. A. (1997). Understanding enzyme superfamilies. Chemistry As the fundamental determinant in the evolution of new catalytic activities. *The Journal of biological chemistry, 272*(49), 30591-30594.

Babbitt, P. C., Hasson, M. S., Wedekind, J. E., Palmer, D. R., Barrett, W. C., Reed, G. H., . . . Gerlt, J. A. (1996). The enolase superfamily: a general strategy for enzyme-catalyzed abstraction of the alpha-protons of carboxylic acids. *Biochemistry, 35*(51), 16489-16501.

Bairoch, A. (1999). The ENZYME data bank in 1999. *Nucleic acids research, 27*(1), 310-311.

Barrett, T., & Edgar, R. (2006). Mining microarray data at NCBI's Gene Expression Omnibus (GEO)*. *Methods Mol Biol, 338*, 175-190.

Baxevanis, A. D., Arents, G., Moudrianakis, E. N., & Landsman, D. (1995). A variety of DNA-binding and multimeric proteins contain the histone fold motif. *Nucleic acids research, 23*(14), 2685-2691.

Baxevanis, A. D., & Landsman, D. (1996). Histone Sequence Database: a compilation of highly-conserved nucleoprotein sequences. *Nucleic acids research, 24*(1), 245-247.

Becker, K. G., Barnes, K. C., Bright, T. J., & Wang, S. A. (2004). The genetic association database. *Nat Genet, 36*(5), 431-432.

Benson, D. A., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2014). GenBank. *Nucleic acids research*.

Berman, H., Henrick, K., Nakamura, H., & Markley, J. L. (2007). The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic acids research, 35*(Database issue), D301-303.

Bertram, L., McQueen, M. B., Mullin, K., Blacker, D., & Tanzi, R. E. (2007). Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. *Nat Genet, 39*(1), 17-23.

Bhaskara, R. M., Mehrotra, P., Rakshambikai, R., Gnanavel, M., Martin, J., & Srinivasan, N. (2014). The relationship between classification of multi-domain proteins using an alignment-free approach and their functions: a case study with immunoglobulins. *Molecular BioSystems, 10*(5), 1082-1093.

Biggs, J. S., Watkins, M., Puillandre, N., Ownby, J. P., Lopez-Vera, E., Christensen, S., . . . Olivera, B. M. (2010). Evolution of Conus peptide toxins: analysis of Conus californicus Reeve, 1844. *Molecular phylogenetics and evolution, 56*(1), 1-12.

Bingham, J., Plowman, G. D., & Sudarsanam, S. (2000). Informatics issues in large-scale sequence analysis: elucidating the protein kinases of C. elegans. *J Cell Biochem, 80*(2), 181-186.

Bockaert, J., & Pin, J. P. (1999). Molecular tinkering of G protein-coupled receptors: an evolutionary success. *Embo J, 18*(7), 1723-1729.

Bonner, T. I. (2014). Should pharmacologists care about alternative splicing? IUPHAR Review 4. *British journal of pharmacology, 171*(5), 1231-1240.

Bradham, C. A., Foltz, K. R., Beane, W. S., Arnone, M. I., Rizzo, F., Coffman, J. A., . . . Manning, G. (2006). The sea urchin kinome: a first look. *Dev Biol, 300*(1), 180-193.

Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Vilo, J., Abeygunawardena, N., . . . Sansone, S. A. (2003). ArrayExpress--a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res, 31*(1), 68-71.

Caenepeel, S., Charydczak, G., Sudarsanam, S., Hunter, T., & Manning, G. (2004). The mouse kinome: discovery and comparative genomics of all mouse protein kinases. *Proc Natl Acad Sci U S A, 101*(32), 11707-11712.

Campbell, J. A., Davies, G. J., Bulone, V., & Henrissat, B. (1997). A classification of nucleotide-diphospho-sugar glycosyltransferases based on amino acid sequence similarities. *The Biochemical journal, 326 (Pt 3)*, 929-939.

Cantarel, B. L., Coutinho, P. M., Rancurel, C., Bernard, T., Lombard, V., & Henrissat, B. (2009). The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic acids research, 37*(Database issue), D233-238.

Chang, D., & Duda, T. F., Jr. (2012). Extensive and continuous duplication facilitates rapid evolution and diversification of gene families. *Molecular biology and evolution, 29*(8), 2019-2029.

Chatonnet, A., Cousin, X., & Robinson, A. (2001). Links between kinetic data and sequences in the alpha/beta-hydrolases fold database. *Briefings in bioinformatics, 2*(1), 30-37.

Chibucos, M. C., Mungall, C. J., Balakrishnan, R., Christie, K. R., Huntley, R. P., White, O., . . . Giglio, M. (2014). Standardized description of scientific evidence using the Evidence Ontology (ECO). *Database (Oxford), 2014*.

Christopoulos, A., Changeux, J. P., Catterall, W. A., Fabbro, D., Burris, T. P., Cidlowski, J. A., . . . Langmead, C. J. (2014). International union of basic and clinical pharmacology. XC. multisite pharmacology: recommendations for the nomenclature of receptor allosterism and allosteric ligands. *Pharmacological reviews, 66*(4), 918-947.

Cousin, X., Hotelier, T., Lievin, P., Toutant, J. P., & Chatonnet, A. (1996). A cholinesterase genes server (ESTHER): a database of cholinesterase-related sequences for multiple alignments, phylogenetic relationships, mutations and structural data retrieval. *Nucleic acids research, 24*(1), 132-136.

Craik, D. J. (2006). Chemistry. Seamless proteins tie up their loose ends. *science, 311*(5767), 1563-1564.

Cuff, A. L., Sillitoe, I., Lewis, T., Clegg, A. B., Rentzsch, R., Furnham, N., . . . Orengo, C. A. (2011). Extending CATH: increasing coverage of the protein structure universe and linking structure with function. *Nucleic acids research, 39*(Database issue), D420-426.

Davis, J., Jones, A., & Lewis, R. J. (2009). Remarkable inter- and intra-species complexity of conotoxins revealed by LC/MS. *Peptides, 30*(7), 1222-1227.

Demeter, J., Beauheim, C., Gollub, J., Hernandez-Boussard, T., Jin, H., Maier, D., . . . Ball, C. A. (2007). The Stanford Microarray Database: implementation of new analysis tools and open source release of software. *Nucleic Acids Res, 35*(Database issue), D766-770.

Deshmukh, K., Anamika, K., & Srinivasan, N. (2010). Evolution of domain combinations in protein kinases and its implications for functional diversity. *Progress in biophysics and molecular biology, 102*(1), 1-15.

Dreos, R., Ambrosini, G., Cavin Perier, R., & Bucher, P. (2013). EPD and EPDnew, high-quality promoter resources in the next-generation sequencing era. *Nucleic Acids Res, 41*(Database issue), D157-164.

Duda, T. F., Jr., Chang, D., Lewis, B. D., & Lee, T. (2009). Geographic variation in venom allelic composition and diets of the widespread predatory marine gastropod Conus ebraeus. *PloS one, 4*(7), e6245.

Duda, T. F., Jr., & Lee, T. (2009). Ecological release and venom evolution of a predatory marine snail at Easter Island. *PloS one, 4*(5), e5558.

Eddy, S. R. (2009). A new generation of homology search tools based on probabilistic inference. *Genome informatics. International Conference on Genome Informatics, 23*(1), 205-211.

Eisen, J. A., Coyne, R. S., Wu, M., Wu, D., Thiagarajan, M., Wortman, J. R., . . . Orias, E. (2006). Macronuclear genome sequence of the ciliate Tetrahymena thermophila, a model eukaryote. *PLoS Biol, 4*(9), e286.

Fernández-Suárez, X. M., Rigden, D. J., & Galperin, M. Y. (2014). The 2014 nucleic acids research database issue and an updated NAR online molecular biology database collection. *Nucleic acids research, 42*(D1), D1-D6.

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., . . . Punta, M. (2014). Pfam: the protein families database. *Nucleic acids research, 42*(Database issue), D222-D230.

Fleischmann, A., Darsow, M., Degtyarenko, K., Fleischmann, W., Boyce, S., Axelsen, K. B., . . . Apweiler, R. (2004). IntEnz, the integrated relational enzyme database. *Nucleic acids research, 32*(Database issue), D434-437.

Gandhimathi, A., Nair, A. G., & Sowdhamini, R. (2012). PASS2 version 4: an update to the database of structure-based sequence alignments of structural domain superfamilies. *Nucleic acids research, 40*(Database issue), D531-534.

Garland, S. L. (2013). Are GPCRs Still a Source of New Targets? *Journal of Biomolecular Screening, 18*(9), 947-966.

Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., . . . Overington, J. P. (2012). ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic acids research, 40*(Database issue), D1100-1107.

Gerlt, J. A., & Babbitt, P. C. (2001). Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies. *Annual review of biochemistry, 70*, 209-246.

Gerlt, J. A., Babbitt, P. C., & Rayment, I. (2005). Divergent evolution in the enolase superfamily: the interplay of mechanism and specificity. *Archives of biochemistry and biophysics, 433*(1), 59-70.

Gnanavel, M., Mehrotra, P., Rakshambikai, R., Martin, J., Srinivasan, N., & Bhaskara, R. M. (2014). CLAP: a web-server for automatic classification of proteins with special reference to multi-domain proteins. *BMC bioinformatics, 15*, 343.

The dictyostelium kinome--analysis of the protein kinases from a simple model organism, 3, 2 Cong. Rec. e38 (2006).

Gowri, V. S., Krishnadev, O., Swamy, C. S., & Srinivasan, N. (2006). MulPSSM: a database of multiple position-specific scoring matrices of protein domain families. *Nucleic acids research, 34*(Database issue), D243-246.

Griffiths-Jones, S., Grocock, R. J., van Dongen, S., Bateman, A., & Enright, A. J. (2006). miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res, 34*(Database issue), D140-144.

Grosjean, J., Soualmia, L., Bouarech, K., Jonquet, C., & Darmoni, S. (2014). *An Approach to Compare Bio-Ontologies Portals.* Paper presented at the MIE'2014: 26th International Conference of the European Federation for Medical Informatics.

Hanks, S. K., & Hunter, T. (1995). Protein kinases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology, 9*(8), 576-596.

HapMap. (2003). The International HapMap Project. *Nature, 426*(6968), 789-796.

Harmar, A. J., Hills, R. A., Rosser, E. M., Jones, M., Buneman, O. P., Dunbar, D. R., . . . Spedding, M. (2009). IUPHAR-DB: the IUPHAR database of G protein-coupled receptors and ion channels. *Nucleic acids research, 37*(Database issue), D680-685.

Henrissat, B. (1991). A classification of glycosyl hydrolases based on amino acid sequence similarities. *The Biochemical journal, 280 (Pt 2)*, 309-316.

Holliday, G. L., Almonacid, D. E., Bartlett, G. J., O'Boyle, N. M., Torrance, J. W., Murray-Rust, P., . . . Thornton, J. M. (2007). MACiE (Mechanism, Annotation and Classification in Enzymes): novel tools for searching catalytic mechanisms. *Nucleic acids research, 35*(Database issue), D515-520.

Holliday, G. L., Andreini, C., Fischer, J. D., Rahman, S. A., Almonacid, D. E., Williams, S. T., & Pearson, W. R. (2012). MACiE: exploring the diversity of biochemical reactions. *Nucleic acids research, 40*(Database issue), D783-789.

Holliday, G. L., Bairoch, A., Bagos, P. G., Chatonnet, A., Craik, D. J., Flinn, R. D., . . . Bateman, A. (2015). Key challenges for the creation and maintenance of specialist protein resources. *Proteins*.

Horn, F., Bettler, E., Oliveira, L., Campagne, F., Cohen, F. E., & Vriend, G. (2003). GPCRDB information system for G protein-coupled receptors. *Nucleic Acids Res, 31*(1), 294-297.
Horn, F., Weare, J., Beukers, M. W., Horsch, S., Bairoch, A., Chen, W., . . . Vriend, G. (1998). GPCRDB: an information system for G protein-coupled receptors. *Nucleic Acids Res, 26*(1), 275-279.
Hsu, S. D., Lin, F. M., Wu, W. Y., Liang, C., Huang, W. C., Chan, W. L., . . . Huang, H. D. (2011). miRTarBase: a database curates experimentally validated microRNA-target interactions. *Nucleic Acids Res, 39*(Database issue), D163-169.
Hunter, T., & Plowman, G. D. (1997). The protein kinases of budding yeast: six score and more. *Trends Biochem Sci, 22*(1), 18-22.
Isberg, V., Vroling, B., van der Kant, R., Li, K., Vriend, G., & Gloriam, D. (2014). GPCRDB: an information system for G protein-coupled receptors. *Nucleic Acids Res, 42*(1), D422-425.
Jamison, D. C. (2003). Structured Query Language (SQL) fundamentals. *Curr Protoc Bioinformatics, Chapter 9*, Unit9 2.
Joosten, R. P., Long, F., Murshudov, G. N., & Perrakis, A. (2014). The PDB_REDO server for macromolecular structure model optimization. *IUCrJ, 1*(4), 0-0.
Kaas, Q., Westermann, J. C., & Craik, D. J. (2010). Conopeptide characterization and classifications: an analysis using ConoServer. *Toxicon : official journal of the International Society on Toxinology, 55*(8), 1491-1509.
Kaas, Q., Yu, R., Jin, A. H., Dutertre, S., & Craik, D. J. (2012). ConoServer: updated content, knowledge, and discovery tools in the conopeptide database. *Nucleic acids research, 40*(Database issue), D325-330.
Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., & Tanabe, M. (2014). Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic acids research, 42*(D1), D199-D205.
Karp, P. D., Riley, M., Saier, M., Paulsen, I. T., Collado-Vides, J., Paley, S. M., . . . Gama-Castro, S. (2002). The EcoCyc Database. *Nucleic Acids Res, 30*(1), 56-58.
Katritch, V., Cherezov, V., & Stevens, R. C. (2013). Structure-function of the G protein-coupled receptor superfamily. *Annu Rev Pharmacol Toxicol, 53*, 531-556.
Kedarisetti, P., Mizianty, M. J., Kaas, Q., Craik, D. J., & Kurgan, L. (2014). Prediction and characterization of cyclic proteins from sequences in three domains of life. *Biochimica et biophysica acta, 1844*(1 Pt B), 181-190.
Kirby, A. J. (2001). The lysozyme mechanism sorted — after 50 years. *Nature Structural Biology, 8*, 737-739.
Knudsen, M., & Wiuf, C. (2010). The CATH database. *Hum Genomics, 4*(3), 207-212.
Kouranov, A., Xie, L., de la Cruz, J., Chen, L., Westbrook, J., Bourne, P. E., & Berman, H. M. (2006). The RCSB PDB information portal for structural genomics. *Nucleic Acids Res, 34*(Database issue), D302-305.
Kozma, D., Simon, I., & Tusnady, G. E. (2013). PDBTM: Protein Data Bank of transmembrane proteins after 8 years. *Nucleic Acids Res, 41*(Database issue), D524-529.
Krupa, A., Abhinandan, K., & Srinivasan, N. (2004). KinG: a database of protein kinases in genomes. *Nucleic acids research, 32*(suppl 1), D153-D155.
Krupp, M., Marquardt, J. U., Sahin, U., Galle, P. R., Castle, J., & Teufel, A. (2012). RNA-Seq Atlas--a reference database for gene expression profiling in normal tissue by next-generation sequencing. *Bioinformatics, 28*(8), 1184-1185.
Lagerstrom, M. C., & Schioth, H. B. (2008). Structural diversity of G protein-coupled receptors and significance for drug discovery. *Nat Rev Drug Discov, 7*(4), 339-357.
Lane, L., Argoud-Puy, G., Britan, A., Cusin, I., Duek, P. D., Evalet, O., . . . Bairoch, A. (2012). neXtProt: a knowledge platform for human proteins. *Nucleic acids research, 40*(Database issue), D76-83.
Lenfant, N., Hotelier, T., Bourne, Y., Marchot, P., & Chatonnet, A. (2013). Proteins with an alpha/beta hydrolase fold: Relationships between subfamilies in an ever-growing superfamily. *Chemico-biological interactions, 203*(1), 266-268.
Lenfant, N., Hotelier, T., Bourne, Y., Marchot, P., & Chatonnet, A. (2014). Tracking the origin and divergence of cholinesterases and neuroligins: the evolution of synaptic proteins. *Journal of molecular neuroscience : MN, 53*(3), 362-369.
Lenfant, N., Hotelier, T., Velluet, E., Bourne, Y., Marchot, P., & Chatonnet, A. (2013). ESTHER, the database of the alpha/beta-hydrolase fold superfamily of proteins: tools to explore diversity of functions. *Nucleic acids research, 41*(Database issue), D423-429.

Lombard, V., Ramulu, H. G., Drula, E., Coutinho, P. M., & Henrissat, B. (2014). The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic acids research, 42*(D1), D490-D495.
Lu, C. T., Huang, K. Y., Su, M. G., Lee, T. Y., Bretana, N. A., Chang, W. C., . . . Huang, H. D. (2013). DbPTM 3.0: an informative resource for investigating substrate site specificity and functional association of protein post-translational modifications. *Nucleic Acids Res, 41*(Database issue), D295-305.
Manning, G. (2005). Genomic overview of protein kinases. *WormBook*, 1-19.
Evolution of protein kinase signaling from yeast to man, 10, 27 Cong. Rec. 514-520 (2002).
The protein kinase complement of the human genome, 5600, 298 Cong. Rec. 1912-1934 (2002).
Marchot, P., & Chatonnet, A. (2012). Enzymatic activity and protein interactions in alpha/beta hydrolase fold proteins: moonlighting versus promiscuity. *Protein and peptide letters, 19*(2), 132-143.
Marino-Ramirez, L., Levine, K. M., Morales, M., Zhang, S., Moreland, R. T., Baxevanis, A. D., & Landsman, D. (2011). The Histone Database: an integrated resource for histones and histone fold-containing proteins. *Database (Oxford), 2011*, bar048.
Martin, A. C., Orengo, C. A., Hutchinson, E. G., Jones, S., Karmirantzou, M., Laskowski, R. A., . . . Thornton, J. M. (1998). Protein folds and functions. *Structure, 6*(7), 875-884.
Martin, J., Anamika, K., & Srinivasan, N. (2010). Classification of protein kinases on the basis of both kinase and non-kinase regions. *PloS one, 5*(9), e12460.
McDonald, A. G., Boyce, S., & Tipton, K. F. (2009). ExplorEnz: the primary source of the IUBMB enzyme list. *Nucleic acids research, 37*(Database issue), D593-597.
Moszer, I., Jones, L. M., Moreira, S., Fabry, C., & Danchin, A. (2002). SubtiList: the reference database for the Bacillus subtilis genome. *Nucleic Acids Res, 30*(1), 62-65.
Murzin, A. G., Brenner, S. E., Hubbard, T., & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of molecular biology, 247*(4), 536-540.
Nagano, N. (2005). EzCatDB: the Enzyme Catalytic-mechanism Database. *Nucleic acids research, 33*(Database issue), D407-412.
Nagano, N., Nakayama, N., Ikeda, K., Fukuie, M., Yokota, K., Doi, T., . . . Tomii, K. (2014). EzCatDB: the enzyme reaction database, 2015 update. *Nucleic acids research*.
Nagy, A., Hegyi, H., Farkas, K., Tordai, H., Kozma, E., Banyai, L., & Patthy, L. (2008). Identification and correction of abnormal, incomplete and mispredicted proteins in public databases. *BMC bioinformatics, 9*, 353.
Overington, J. P., Al-Lazikani, B., & Hopkins, A. L. (2006). How many drug targets are there? *Nat Rev Drug Discov, 5*(12), 993-996.
Pawson, A. J., Sharman, J. L., Benson, H. E., Faccenda, E., Alexander, S. P., Buneman, O. P., . . . Harmar, A. J. (2014). The IUPHAR/BPS Guide to PHARMACOLOGY: an expert-driven knowledgebase of drug targets and their ligands. *Nucleic acids research, 42*(Database issue), D1098-1106.
Pegg, S. C., Brown, S. D., Ojha, S., Seffernick, J., Meng, E. C., Morris, J. H., . . . Babbitt, P. C. (2006). Leveraging enzyme structure-function relationships for functional inference and experimental design: the structure-function linkage database. *Biochemistry, 45*(8), 2545-2555.
Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., & Ferrin, T. E. (2004). UCSF Chimera--a visualization system for exploratory research and analysis. *Journal of computational chemistry, 25*(13), 1605-1612.
Pettifer, S., Ison, J., Kalas, M., Thorne, D., McDermott, P., Jonassen, I., . . . Vriend, G. (2010). The EMBRACE web service collection. *Nucleic Acids Res, 38*(Web Server issue), W683-688.
Plowman, G. D., Sudarsanam, S., Bingham, J., Whyte, D., & Hunter, T. (1999). The protein kinases of Caenorhabditis elegans: a model for signal transduction in multicellular organisms. *Proc Natl Acad Sci U S A, 96*(24), 13603-13610.
Poth, A. G., Chan, L. Y., & Craik, D. J. (2013). Cyclotides as grafting frameworks for protein engineering and drug design applications. *Biopolymers, 100*(5), 480-491.
Puillandre, N., Koua, D., Favreau, P., Olivera, B. M., & Stocklin, R. (2012). Molecular phylogeny, classification and evolution of conopeptides. *Journal of molecular evolution, 74*(5-6), 297-309.
Rakshambikai, R., Gnanavel, M., & Srinivasan, N. (2014). Hybrid and rogue kinases encoded in the genomes of model eukaryotes. *PloS one, 9*(9), e107956.

Rawlings, N. D., Waller, M., Barrett, A. J., & Bateman, A. (2014). MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. *Nucleic acids research, 42*(Database issue), D503-D509.
Reddy, T. B., Thomas, A. D., Stamatis, D., Bertsch, J., Isbandi, M., Jansson, J., . . . Kyrpides, N. C. (2015). The Genomes OnLine Database (GOLD) v.5: a metadata management system based on a four level (meta)genome project classification. *Nucleic Acids Res, 43*(Database issue), D1099-1106.
Rhodes, D. R., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., . . . Chinnaiyan, A. M. (2004). ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia, 6*(1), 1-6.
Rose, P. W., Bi, C., Bluhm, W. F., Christie, C. H., Dimitropoulos, D., Dutta, S., . . . Bourne, P. E. (2013). The RCSB Protein Data Bank: new resources for research and education. *Nucleic acids research, 41*(Database issue), D475-482.
Rose, P. W., Prlic, A., Bi, C., Bluhm, W. F., Christie, C. H., Dutta, S., . . . Burley, S. K. (2014). The RCSB Protein Data Bank: views of structural biology for basic and applied research and education. *Nucleic acids research*.
Saier, M. H., Jr. (2000). A functional-phylogenetic classification system for transmembrane solute transporters. *Microbiol Mol Biol Rev, 64*(2), 354-411.
Saier, M. H., Reddy, V. S., Tamang, D. G., & Västermark, Å. (2014). The Transporter Classification Database. . *Nucleic acids research, 42*(Database issue), D251-D258.
Scherf, M., Epple, A., & Werner, T. (2005). The next generation of literature analysis: integration of genomic analysis into text mining. *Brief Bioinform, 6*(3), 287-297.
Schnoes, A. M., Brown, S. D., Dodevski, I., & Babbitt, P. C. (2009). Annotation Error in Public Databases: Misannotation of Molecular Function in Enzyme Superfamilies. *PLoS computational biology, 5*(12), e1000605.
Schully, S. D., Yu, W., McCallum, V., Benedicto, C. B., Dong, L. M., Wulf, A., . . . Khoury, M. J. (2011). Cancer GAMAdb: database of cancer genetic associations from meta-analyses and genome-wide association studies. *Eur J Hum Genet, 19*(8), 928-930.
Sethupathy, P., Corda, B., & Hatzigeorgiou, A. G. (2006). TarBase: A comprehensive database of experimentally supported animal microRNA targets. *RNA, 12*(2), 192-197.
Shepelev, V., & Fedorov, A. (2006). Advances in the Exon-Intron Database (EID). *Brief Bioinform, 7*(2), 178-185.
Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res, 29*(1), 308-311.
Sigrist, C. J., Cerutti, L., de Castro, E., Langendijk-Genevaux, P. S., Bulliard, V., Bairoch, A., & Hulo, N. (2010). PROSITE, a protein domain database for functional characterization and annotation. *Nucleic Acids Res, 38*(Database issue), D161-166.
Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., . . . Lewis, S. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology, 25*(11).
Sowdhamini, R., Burke, D. F., Huang, J. F., Mizuguchi, K., Nagarajaram, H. A., Srinivasan, N., . . . Blundell, T. L. (1998). CAMPASS: a database of structurally aligned protein superfamilies. *Structure, 6*(9), 1087-1094.
Spedding, M. (2011). Resolution of controversies in drug/receptor interactions by protein structure. Limitations and pharmacological solutions. *Neuropharmacology, 60*(1), 3-6.
Srivastava, M., Simakov, O., Chapman, J., Fahey, B., Gauthier, M. E., Mitros, T., . . . Rokhsar, D. S. (2010). The Amphimedon queenslandica genome and the evolution of animal complexity. *Nature, 466*(7307), 720-726.
Stajich, J. E., Wilke, S. K., Ahren, D., Au, C. H., Birren, B. W., Borodovsky, M., . . . Pukkila, P. J. (2010). Insights into evolution of multicellular fungi from the assembled chromosomes of the mushroom Coprinopsis cinerea (Coprinus cinereus). *Proc Natl Acad Sci U S A, 107*(26), 11889-11894.
Stark, C., Breitkreutz, B. J., Reguly, T., Boucher, L., Breitkreutz, A., & Tyers, M. (2006). BioGRID: a general repository for interaction datasets. *Nucleic Acids Res, 34*(Database issue), D535-539.
Stein, L. (2013). Creating databases for biological information: an introduction. *Curr Protoc Bioinformatics, Chapter 9*, Unit9 1.

Sun, H., Palaniswamy, S. K., Pohar, T. T., Jin, V. X., Huang, T. H., & Davuluri, R. V. (2006). MPromDb: an integrated resource for annotation and visualization of mammalian gene promoters and ChIP-chip experimental data. *Nucleic Acids Res, 34*(Database issue), D98-103.

Tan, N. C., & Berkovic, S. F. (2010). The Epilepsy Genetic Association Database (epiGAD): analysis of 165 genetic association studies, 1996-2008. *Epilepsia, 51*(4), 686-689.

Terlau, H., & Olivera, B. M. (2004). Conus venoms: a rich source of novel ion channel-targeted peptides. *Physiological reviews, 84*(1), 41-68.

Theodoropoulou, M. C., Bagos, P. G., Spyropoulos, I. C., & Hamodrakas, S. J. (2008). gpDB: a database of GPCRs, G-proteins, effectors and their interactions. *Bioinformatics, 24*(12), 1471-1472.

Tipton, K. F. (1994). Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB). Enzyme nomenclature. Recommendations 1992. Supplement: corrections and additions. *European journal of biochemistry / FEBS, 223*(1), 1-5.

Tough, D. F., Lewis, H. D., Rioja, I., Lindon, M. J., & Prinjha, R. K. (2014). Epigenetic pathway targets for the treatment of disease: accelerating progress in the development of pharmacological tools: IUPHAR Review 11. *British journal of pharmacology, 171*(22), 4981-5010.

Trabi, M., & Craik, D. J. (2002). Circular proteins--no end in sight. *Trends Biochem Sci, 27*(3), 132-138.

Tsaousis, G. N., Tsirigos, K. D., Andrianou, X. D., Liakopoulos, T. D., Bagos, P. G., & Hamodrakas, S. J. (2010). ExTopoDB: a database of experimentally derived topological models of transmembrane proteins. *Bioinformatics, 26*(19), 2490-2492.

Tsirigos, K. D., Bagos, P. G., & Hamodrakas, S. J. (2011). OMPdb: a database of {beta}-barrel outer membrane proteins from Gram-negative bacteria. *Nucleic acids research, 39*(Database issue), D324-331.

Umemura, M., Nagano, N., Koike, H., Kawano, J., Ishii, T., Miyamura, Y., . . . Machida, M. (2014). Characterization of the biosynthetic gene cluster for the ribosomally synthesized cyclic peptide ustiloxin B in Aspergillus flavus. *Fungal Genet Biol, 68*, 23-30.

UniProt. (2014). Activities at the Universal Protein Resource (UniProt). *Nucleic acids research, 42*(Database issue), D191-198.

Vroling, B., Sanders, M., Baakman, C., Borrmann, A., Verhoeven, S., Klomp, J., . . . Vriend, G. (2011). GPCRDB: information system for G protein-coupled receptors. *Nucleic acids research, 39*(Database issue), D309-D319.

Wang, C. K., Kaas, Q., Chiche, L., & Craik, D. J. (2008). CyBase: a database of cyclic protein sequences and structures, with applications in protein discovery and engineering. *Nucleic acids research, 36*(suppl 1), D206-D210.

Wong, W. C., Maurer-Stroh, S., & Eisenhaber, F. (2010). More than 1,001 problems with protein domain databases: transmembrane regions, signal peptides and the issue of sequence homology. *PLoS Comput Biol, 6*(7), e1000867.

Wren, J. D. (2008). URL decay in MEDLINE—a 4-year follow-up study. *Bioinformatics, 24*(11), 1381-1385.

Xenarios, I., Salwinski, L., Duan, X. J., Higney, P., Kim, S. M., & Eisenberg, D. (2002). DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res, 30*(1), 303-305.

Xia, J., Wang, Q., Jia, P., Wang, B., Pao, W., & Zhao, Z. (2012). NGS catalog: A database of next generation sequencing studies in humans. *Hum Mutat, 33*(6), E2341-2355.