

Διδακτορική Διατριβή

Βιοπληροφορική ανάλυση πρωτεϊνωματικών δεδομένων και σημάτων στόχευσης στο βακτήριο *E.coli*.

Ορφανουδάκη Γεωργία

Τριμελής Επιτροπή

Οικονόμου Αναστάσιος (Επιβλέπων Καθηγητής)

Τσαμαρδίνος Ιωάννης (Συνεπιβλέπων καθηγητής)

Ποϊράζη Παναγιώτα

2010-2015







<b>Περιεχόμενα</b>		
<b>Συντομογραφίες</b>	<b>15</b>	
<b>Περίληψη</b>	<b>17</b>	
<b>ΚΕΦΑΛΑΙΟ 1</b>	<b>Εισαγωγή</b>	<b>21</b>
1.1	Δομή του βακτηριακού κυττάρου	21
1.2	Κυτταρικός φάκελος και τα βιολογικά μέρη από τα οποία αποτελείται	22
1.2.1.	Η πλασματική μεμβράνη και οι πρωτεΐνες που σχετίζονται με αυτήν	22
1.2.2.	Το κυτταρικό τοίχωμα της πεπτιδογλυκάνης	23
1.2.3.	Η εξωτερική μεμβράνη και οι πρωτεΐνες που εμπεριέχονται σε αυτήν	24
1.2.4.	Σύνθεση λιποσακχαριτών και μηχανισμός έκκρισης τους	25
1.2.5.	Σύνθεση λιποπρωτεϊνών	26
1.3	Συστήματα έκκρισης πρωτεϊνών	27
1.4	Η μεταθετάση Sec	28
1.5	Χαρακτηριστικά εκκρινόμενων πρωτεϊνών	32
1.1.1.	Σηματοδοτικό πεπτιδίο	32
1.1.2.	Ωριμο τμήμα	34
<b>ΚΕΦΑΛΑΙΟ 2</b>	<b>Αποτελέσματα – Μελέτη της κατανομής των πρωτεϊνών του βακτηρίου <i>E.coli</i> στα διάφορα υποκυτταρικά διαμερίσματα</b>	<b>37</b>
2.1	Υποκυτταρικά διαμερίσματα βακτηρίου <i>E.coli</i>	38
2.2	Βασικό πρωτεϊνωμα του <i>E.coli</i>	39
2.3	Πλήρης ταξινόμηση των πρωτεϊνών του βακτηρίου <i>E.coli</i> σε υποκυτταρικά διαμερίσματα	42
2.4	Ασυνέπεια μεταξύ παλαιότερων ταξινομήσεων	43

K-12.	2.5	Συνδυαστική ανάλυση για την εμπειριστατωμένη <i>de novo</i> ταξινόμηση του πρωτεϊνώματος του <i>E.coli</i>	47
	2.6	Ταξινόμηση «άγνωστων» πρωτεϊνών	47
	2.7	Διαλεύκανση αντιφάσεων στις υπάρχουσες ταξινομήσεις	51
	2.8	Βιβλιογραφική αναζήτηση πρωτεϊνών	53
	2.9	Πρωτεΐνες με πολλαπλές υποκυτταρικές τοποθεσίες.	57
	2.10	Περιφερικές πρωτεΐνες	58
	2.11	Σύγκριση και αντιστοίχιση δύο πρωτεϊνωμάτων <i>E.coli</i>	58
	2.12	Οργάνωση των εκκριτικών μηχανισμών στο <i>E.coli</i> και προσδιορισμός των πρωτεϊνών που συμμετέχουν σε αυτά.	59
	2.12.1.	Εκκριτικά Μονοπάτια στο <i>E.coli</i> K-12	59
	2.12.2.	Εκκριτικές πρωτεΐνες	60
	2.13	Η βάση δεδομένων STEPdb (Sub-cellular Topologies of <i>E.coli</i> Polypeptides database)	64
	2.13.1.	Βιοπληροφορικά εργαλεία που συμπεριλαμβάνονται στο STEPdb	65
	2.13.2.	Δομικές και φυσικοχημικές ιδιότητες των πρωτεϊνών	65
	2.13.3.	Απεικόνιση πρωτεϊνικών συμπλόκων	66
	2.14	Συζήτηση	68

### **ΚΕΦΑΛΑΙΟ 3      Αποτελέσματα – Μελέτη των υποστρωμάτων του συστήματος έκκρισης Sec με στόχο τον χαρακτηρισμό νέων σημάτων έκκρισης στην περιοχή του ώριμου τμήματος.      71**

	3.1	Επιλογή δεδομένων	74
	3.2	Συντηρημένα αμινοξέα στις αλληλουχίες των εκκρινόμενων πρωτεϊνών, σύγκριση με κυτταροπλασματικές.	74
	3.3	Επιλογή της ελάχιστης αλληλουχίας προς ανάλυση	77

3.4	Εκπαίδευση μοντέλων διαχωρισμού χρησιμοποιώντας το GEMS σαν εργαλείο μηχανικής μάθησης	78
3.5	Μεταβλητές Εκπαίδευσης	78
3.5.1.	Μοντέλα πρόδρομης και ώριμης μορφής	79
3.5.2.	Μοντέλα με άλλες μεταβλητές	84
3.6	Μελέτη της ικανότητας αναδίπλωσης των εκκρινόμενων πρωτεϊνών	84
3.7	Εκπαίδευση μοντέλων διαχωρισμού χρησιμοποιώντας την πληροφορία της πιθανότητας αταξίας	89
3.8	Μελέτη υδρόφοβων περιοχών στα ώριμα τμήματα	97
3.9	Μελέτη των περιοχών που προάγουν τον σχηματισμό συσσωματώσεων	99
3.10	<i>In silico</i> συνδυασμοί σηματοδοτικών πεπτιδίων και ώριμων τμημάτων	100
3.11	Πειραματική επιβεβαίωση των μοντέλων διαχωρισμού	103
3.12	Σύγκριση απόδοσης με άλλα βιοπληροφορικά εργαλεία	104
3.13	Πρόβλεψη εκκριτικών πρωτεϊνών σε άλλα Gram <sup>-</sup> και Gram <sup>+</sup> βακτήρια	105
3.14	Συζήτηση	107
<b>ΚΕΦΑΛΑΙΟ 4 Αποτελέσματα – Μελέτη των περιφερικών πρωτεϊνών του <i>E.coli</i> με στόχο την χαρτογράφηση των πρωτεϊνικών αλληλεπιδράσεων και των κυτταρικών λειτουργιών στις οποίες συμμετέχουν.</b>		<b>113</b>
4.1	Μελέτη της κυτταρικής λειτουργίας των περιφερικών πρωτεϊνών	115
4.2	Συζήτηση	121
<b>ΚΕΦΑΛΑΙΟ 5 Αποτελέσματα – Μελέτη των φυσικοχημικών ιδιοτήτων των πρωτεϊνών του ΚΦ με στόχο την βέλτιστη ανίχνευση με τεχνικές φασματομετρίας μάζας</b>		<b>127</b>
5.1	Υποκυτταρική ταξινόμηση του θεωρητικού πρωτεϊνώματος του <i>E.coli</i> BL21-DE3	129
5.2	Εκτίμηση του υποσύνολου του πρωτεϊνώματος που εκφράζεται όταν το βακτήριο αναπτύσσεται σε πλούσιο θρεπτικό μέσο.	131

5.3	Μελέτη των φυσικοχημικών χαρακτηριστικών των πρωτεϊνών του ΚΦ και των πεπτιδίων που παράγονται από αυτές στο στάδιο της πρωτεόλυσης	134
5.4	Πειραματική ανίχνευση των πρωτεϊνών του κυτταρικού φακέλου	138
5.5	Σύγκριση μεταξύ των πειραματικών μεθόδων	143
5.6	Μελέτη των φυσικοχημικών χαρακτηριστικών των μεμβρανικών πρωτεϊνών που δεν ανιχνεύτηκαν με καμία μέθοδο	144
5.7	Σχετική ποσοτικοποίηση πρωτεϊνών χωρίς χημική σήμανση με ισότοπα (Label-free)	147
5.8	Συζήτηση	151
<b>ΚΕΦΑΛΑΙΟ 6</b>	<b>Υλικά και μέθοδοι</b>	<b>157</b>
6.1	Υποκυτταρική ταξινόμηση πρωτεϊνώματος <i>E.coli</i>	157
6.1.1.	Το πρωτεϊνωμα αναφοράς του <i>E.coli</i> K-12	157
6.1.2.	Βιοπληροφορικά εργαλεία και καθορισμός παραμέτρων.	157
6.1.3.	Εκτίμηση αξιοπιστίας πειραματικών δεδομένων	158
6.1.4.	Σύγκριση ανάμεσα στα στελέχη K-12 και BL21-DE3 του <i>E.coli</i>	160
6.1.5.	Ανάπτυξη της βάσης δεδομένων STEPdb	162
6.2	Μοντέλα διαχωρισμού εκκρινόμενων από κυττροπλασματικές πρωτεΐνες	162
6.2.1.	Πρόβλεψη σημείου εκκίνησης ώριμου τμήματος και καθορισμός του ελάχιστου δυνατού μήκους προς ανάλυση	162
6.2.2.	Αναπαράσταση των αμινοξικών αλληλουχιών	164
6.2.3.	Συναρτήσεις πυρήνα (kernels)	167
6.2.4.	Εκτίμηση της απόδοσης των μοντέλων	168
6.2.5.	Επιλογή Χαρακτηριστικών	168
6.2.6.	Σύνολα εκπαίδευσης και αξιολόγησης (test and train sets)	168



6.2.7.	Βάρη των επιλεγμένων χαρακτηριστικών	169
6.2.8.	Ψεύδο-αμινοξική σύσταση (Pseudo amino-acid Composition)	170
6.2.9.	Παράγωγα εκκρινόμενων πρωτεϊνών, εκτίμηση πειραματικών δεδομένων	172
6.2.10.	Σύγκριση με άλλα βιοπληροφορικά εργαλεία	177
6.2.11.	Εκτίμηση των υδρόφοβων περιοχών	177
6.2.12.	Υπολογισμός προδιάθεσης σχηματισμού συσσωματώσεων	178
6.2.13.	Υπολογισμός μέσου προφίλ αταξίας	178
6.2.14.	Εκπαίδευση μοντέλων με την πληροφορία της ενέργειας αλληλεπίδρασης των αμινοξέων	179
6.2.15.	<i>In silico</i> συνδυασμοί σηματοδοτικών πεπτιδίων με αντίστοιχα ώριμα τμήματα	179
6.2.16.	Πρόβλεψη εκκρινόμενων πρωτεϊνών σε άλλα βακτήρια	179
6.3	Μελέτη των περιορισμών ανίχνευσης των πρωτεϊνών του ΚΦ με μεθόδους πρωτεομικής	181
6.3.1.	Προσδιορισμός του ανιχνεύσιμου πρωτεϊνώματος σε συνθήκες πλούσιου θρεπτικού μέσου (LB) 181	
6.3.2.	Υπολογισμός της κατανομή GRAVY σε συνάρτηση με το μήκος και δισδιάστατης κατανομή των μεμβρανικών πρωτεϊνών	186
6.3.3.	Προβλεπόμενο διαμεμβρανικό κομμάτι των πεπτιδίων	186
6.3.4.	Ανάλυση των μεταγραφικών μονάδων (TUs) που εκφράζονται	186
6.3.5.	Κριτήρια για το καθορισμό των πρωτεϊνών που ανιχνεύονται συστηματικά	191
6.3.6.	Μοντελοποίηση και προσδιορισμός πεπτιδίων που ανιχνεύονται με την πειραματική διαδικασία της πρωτεόλυσης επιφάνειας AMK	191
6.3.7.	Πρωτεόλυση επιφάνειας – υπολογισμός ανιχνεύσιμου μήκους πρωτεΐνης ( $L_{SP}$ )	193
6.3.8.	Διόρθωση τιμών σχετικής ποσότητας NSAF ( $NSAF_{SP}$ )	193
6.3.9.	Διόρθωση τιμών σχετικής ποσότητας emPAI ( $emPAI_{SP}$ )	194

6.3.10. Στατιστική ανάλυση για τον καθορισμό των πρωτεϊνών που η ποσότητα τους έχει μεταβληθεί σημαντικά ανάμεσα στα αγρίου τύπου και SecYEG κύτταρα 194

## **ΠΑΡΑΡΤΗΜΑΤΑ 211**

## Ευρετήριο Εικόνων

Εικόνα 1.1 - Απεικόνιση της κυτταρικής δομής των κατά Gram θετικών (Gram <sup>+</sup> ) και αρνητικών (Gram <sup>-</sup> ) βακτηρίων .....	21
Εικόνα 1.2 – Περιφερικές πρωτεΐνες .....	23
Εικόνα 1.3 - Η δομή της πρωτεΐνης LamB .....	24
Εικόνα 1.4 - Λιποπρωτεΐνες στα βακτήρια και το κανονικό μονοπάτι βιοσύνθεσης (Nakayama et al, 2012) .....	27
Εικόνα 1.5 - Σηματοδοτικό πεπτιδίο τύπου I .....	33
Εικόνα 2.1 – Κατηγορίες υποκυτταρικού εντοπισμού .....	42
Εικόνα 2.2 – Εμπεριστατωμένη υποκυτταρική ταξινόμηση του συνολικού πρωτεϊνώματος του <i>E.coli</i> .....	46
Εικόνα 2.3 – Διάγραμμα ροής της διαδικασίας υποκυτταρικής ταξινόμησης του πρωτεϊνώματος <i>E.coli</i> .....	50
Εικόνα 2.4 – Σύνοψη της υποκυτταρικής ταξινόμησης των πρωτεϊνών και των πειραματικά επαληθευμένων πρωτεϊνών. ....	57
Εικόνα 2.5 - Τα βασικά εκκριτικά μονοπάτια στο βακτήριο <i>E.coli</i> .....	63
Εικόνα 2.6 – Προσανατολισμός μεμβρανικών πρωτεϊνών .....	64
Εικόνα 2.7 – Η ιστοσελίδα της βάσης δεδομένων STEPdb .....	67
Εικόνα 3.1 – Σχηματική απεικόνιση της τομής ενός Gram <sup>-</sup> κυτάρου – Sec σύστημα έκκρισης. ....	72
Εικόνα 3.2 - Συντηρημένα αμινοξικά χαρακτηριστικά των εκκρινόμενων και κυτταροπλασματικών πρωτεϊνών .....	76
Εικόνα 3.3 – Συντηρημένα μοτίβα που αναγνωρίζονται από τις πεπτιδάσες τύπου I και II .....	77
Εικόνα 3.4- Απεικόνιση μοντέλων πρόδρομης και ώριμης μορφής .....	83
Εικόνα 3.5 – Πίνακας πιθανότητας ενέργειας αλληλεπίδρασης .....	86
Εικόνα 3.6 – Βασικοί άξονες ενέργειας αναδίπλωσης πρωτεϊνών. ....	87
Εικόνα 3.7 – Μέσο προφίλ δομικής αταξίας (disorder) .....	88
Εικόνα 3.8 – Μοντέλο δομικής αταξίας – επιλεγμένα χαρακτηριστικά .....	90
Εικόνα 3.9 – Σύνοψη των χαρακτηριστικών των υδρόφοβων περιοχών. ....	98
Εικόνα 3.10 – Σύνοψη περιοχών που προάγουν τον σχηματισμό συσσωματώσεων .....	100

Εικόνα 3.11 – <i>In silico</i> συνδυασμοί ΣΠ με ΩΤ και κυτταροπλασματικές αλληλουχίες .....	101
Εικόνα 3.12 Προφίλ αταξίας των βέλτιστων και χειρότερων συνδυασμών.....	102
Εικόνα 3.13 – Πειραματικά δεδομένα, μέση πιθανότητα αταξίας .....	103
Εικόνα 4.1 – Πρωτεΐνες που σχετίζονται με την πλασματική μεμβράνη .....	114
Εικόνα 4.2 – «Πανοραμική» εικόνα των περιφερικών πρωτεϊνών του <i>E.coli</i> .....	119
Εικόνα 4.3 – Ταξινόμηση των περιφερικών πρωτεϊνών με βάση τις κυτταρικές λειτουργίες στις οποίες συμμετέχουν.....	120
Εικόνα 5.1 - Εμπεριστατωμένη ταξινόμηση του πρωτεϊνώματος του <i>E.coli</i> BL21-DE3 και πειραματική διαδικασία ανίχνευσης με μεθόδους πρωτεομικής ανάλυσης.....	130
Εικόνα 5.2 – Στάδια προς την ταυτοποίηση πρωτεϊνών με μεθόδους πρωτεομικής ανάλυσης φασματομετρίας μάζας .....	131
Εικόνα 5.3 Διάγραμμα ροής των κριτηρίων για την επιλογή των ανιχνεύσιμων πεπτιδίων με φασματομετρία μάζας ('MS-detectable' peptides).....	135
Εικόνα 5.4 – Φυσικοχημικές ιδιότητες των κυτταροπλασματικών (CYTO), μεμβρανικών (IM) πρωτεϊνών και των πεπτιδίων τους.....	137
Εικόνα 5.5 – Ανίχνευση μεμβρανικών πρωτεϊνών με διαφορετικές μεθόδους προετοιμασίας δείγματος, σύγκριση φυσικοχημικών ιδιοτήτων των αντίστοιχων πεπτιδίων .....	139
Εικόνα 5.6 – Υποκυτταρικός εντοπισμός των πρωτεϊνών που ανιχνεύτηκαν με διαφορετικές μεθόδους προετοιμασίας δείγματος σε συνδυασμό με φασματομετρία μάζας.....	142
Εικόνα 5.7 – Ανάλυση των μεμβρανικών πρωτεϊνών που δεν ταυτοποιήθηκαν .....	146
Εικόνα 5.8 – Ποσοτική ανάλυση των πρωτεϊνών του ΚΦ.....	149
Εικόνα 5.9 – Σύνοψη των πρωτεϊνών του ΚΦ που έχουν ταυτοποιηθεί μέχρι σήμερα από πρωτεομικές αναλύσεις .....	153
Εικόνα 6.1- Ανάλυση για τον ορισμό των πεπτιδίων που ανιχνεύονται με την μέθοδο της πρωτεόλυσης της επιφάνειας ΑΜΚ (MS-detectable'SP).....	192

## Ευρετήριο Πινάκων

Πίνακας 2.1 – Κινητά στοιχεία των στελεχών <i>E.coli</i> K-12 και BL21-DE3 .....	40
Πίνακας 2.2 - Ανάλυση των ψευδογονιδίων στο στέλεχος <i>E.coli</i> K-12 .....	41
Πίνακας 2.3 - Αντιστοιχία ανάμεσα στην ονοματολογία του STEPdb και των όρων οντολογίας (gene ontology) .....	43
Πίνακας 2.4 – Κριτήρια απόφασης για την ταξινόμηση των πρωτεϊνών με βάση τις προβλέψεις των βιοπληροφορικών εργαλείων (BE).....	53
Πίνακας 2.5 – Παραδείγματα πρωτεϊνών με αντικρουόμενες προτεινόμενες υποκυτταρικές ταξινομήσεις οι οποίες επανεξετάστηκαν και επιλύθηκαν.....	54
Πίνακας 2.6 - Σύνοψη της ήδη υπάρχουσας ταξινόμησης και σύγκριση με την ταξινόμηση στη βάση δεδομένων STEPdb.....	56
Πίνακας 3.1 – Σύγκριση συμπαγούς και χαλαρής ομαδοποίησης αμινοξέων σε μοντέλα διαχωρισμού ώριμων τμημάτων από κυτταροπλασματικές πρωτεΐνες.....	81
Πίνακας 3.2 – Σύγκριση απόδοσης μοντέλων .....	92
Πίνακας 3.3 – Μοντέλα διαχωρισμού των ώριμων μορφών των εκκριτικών πρωτεϊνών από τις κυτταροπλασματικές πρωτεΐνες.....	94
Πίνακας 3.4 – Σύγκριση των μοντέλων διαχωρισμού με άλλα βιοπληροφορικά εργαλεία .....	104
Πίνακας 3.5 – Πρόβλεψη σε άλλα βακτήρια .....	105
Πίνακας 5.1 – Ανιχνεύσιμο πρωτεΐνωμα του <i>E. coli</i> BL21-DE3 .....	132
Πίνακας 6.1 - Ευρείας κλίμακας πρωτεομικές, γονιδιωματικές και βιοχημικές αναλύσεις στις οποίες βασίστηκε η υποκυτταρική ταξινόμηση του <i>E.coli</i> .....	159
Πίνακας 6.2 - Κατάλογος με τα 43 στελέχη <i>E.coli</i> με βάση τα οποία προσδιορίστηκε το πρωτεΐνωμα πυρήνα.....	161
Πίνακας 6.3 – Κατάλογος πρωτεϊνών με τα μικρότερα σε μήκος ώριμα τμήματα .....	163
Πίνακας 6.4 – Απλή αναπαράσταση των αμινοξέων .....	165
Πίνακας 6.5 – Συμπαγής αναπαράσταση των αμινοξέων.....	166
Πίνακας 6.6 – Χαλαρή αναπαράσταση των αμινοξέων .....	167
Πίνακας 6.7 – Σύνοψη των συνόλου εκπαίδευσης και αξιολόγησης .....	169
Πίνακας 6.8 – Παράγωγα εκκρινόμενων πρωτεϊνών .....	173

Πίνακας 6.9 – Κατάλογος Gram <sup>+</sup> και Gram <sup>-</sup> βακτηρίων και σύνοψη του αριθμού των πιθανών εκκρινόμενων πρωτεϊνών τύπου Sec .....	180
Πίνακας 6.10 – Ταυτοποίηση πρωτεϊνών του ΚΦ από πρωτεομικές αναλύσεις.....	183
Πίνακας 6.11 – Μεταγραφικές μονάδες που αναμένεται να εκφράζονται σε επίπεδο πρωτεΐνης.	187

## Συντομογραφίες

AUC	Area Under the Curve
Bam	Beta-barrel assembly machine
CU	Chaperone-usher pathway
DNA	Deoxyribonucleic acid
Fla	Flagellum
IM	Inner Membrane
IMP	Inner Membrane Protein
CE	Cell Envelope
CEP	Cell Envelope
CYTO	Cytoplasmic
LB	Luria-Bertani
LOL	Lipoprotein outer-membrane localization
mRNA	Messenger Ribonucleic acid
OM	Outer Membrane
PhoA	Alkaline phosphatase A
SDS PAGE	Sodium Dodecyl Sulfate Polyacrylamide Gel Electrophoresis
SP	Signal Peptide
SRP	Signal Recognition Particle
SVM	Support Vector Machines
T2S	Type II secretion
T3S	Type III secretion
T4S	Type IV secretion
T5S	Type V secretion
TM	Transmembrane
AUC	Area Under the Curve
FASP	Filter Aided Sample Preparation
GRAVY index	Grand average of hydropathicity index
AMK	Ανεστραμμένα μεμβρανικά κυστίδια
BE	Βιοπληροφορικά Εργαλεία
ΔΠ	Διαμεμβρανική Περιοχή
EM	Εξωτερική Μεμβράνη
ΚΦ	Κυτταρικός Φάκελος
ΜΠ	Μεμβρανικό πρωτεΐνωμα
ΜΚΕΜ	Μεμβρανικά Κυστίδια της Εξωτερικής Μεμβράνης
ΠΚΦ	Πρωτεΐνωμα Κυτταρικού Φακέλου
ΠΜ	Πλασματική Μεμβράνη
ΠΠ	Πλέγμα της πεπτιδογλυκάνης
ΣΠ	Σηματοδοτικό πεπτίδιο
ΦΜ	Φασματομετρία μάζας
ΩΤ	Ωριμο Τμήμα





## Περίληψη

Η ύπαρξη διακριτών περιοχών στο κύτταρο εξυπηρετεί την απομόνωση αλλά και την εξειδίκευση των διαφορετικών κυτταρικών λειτουργιών. Για το λόγο αυτό τα κύτταρα έχουν αναπτύξει διαφορετικούς μηχανισμούς έκκρισης και καθοδήγησης των πρωτεϊνών στα σημεία όπου πρόκειται να πραγματοποιήσουν τις λειτουργίες τους. Στα βακτήρια έχουν αναγνωριστεί τουλάχιστον 16 τέτοιοι μηχανισμοί εκ των οποίων ο μηχανισμός Sec είναι κοινός και αναγκαίος για την βιωσιμότητα του κυττάρου.

Εκτιμάται ότι περίπου το ένα τρίτο των πρωτεϊνών μετά την σύνθεση τους στο κυτταρόπλασμα οδηγείται σε άλλα υποκυτταρικά διαμερίσματα ή στο εξωκυττάριο χώρο. Η γνώση της κατανομής των πρωτεϊνών μέσα στο κύτταρο αλλά και των αλληλεπιδράσεων μεταξύ τους αποτελεί το πρώτο βήμα για την κατανόηση του κυττάρου ως ολότητα. Συνιστά επίσης, απαραίτητη προεργασία για οποιαδήποτε πειραματική μελέτη ευρείας κλίμακας (π.χ. πρωτεομικές αναλύσεις) με μελλοντικό στόχο την *in silico* μοντελοποίηση και κατανόηση των κυττάρων αλλά και για την ανάπτυξη αξιόπιστων βιοπληροφορικών εργαλείων.

Στην παρούσα διατριβή αρχικά θα περιγράψουμε την συνδυαστική ανάλυση που ακολουθήσαμε για την εμπειριστατωμένη ταξινόμηση του πρωτεϊνώματος του *E.coli* σε 13 υποκυτταρικά διαμερίσματα και η οποία βασίστηκε κυρίως σε επισταμένη βιβλιογραφική έρευνα (Κεφάλαιο 2). Στη συνέχεια θα περιγράψουμε πως με εφελτήριο και ως βασικό εργαλείο την εκτεταμένη υποκυτταρική ταξινόμηση των πρωτεϊνών του *E.coli* οδηγήσαμε: στην ανάπτυξη νέων βιοπληροφορικών εργαλείων, στην ανάδειξη και τον χαρακτηρισμό άγνωστων μέχρι σήμερα σημάτων έκκρισης, στην βελτίωση πειραματικών μεθοδολογιών μέσω της διερεύνησης των φυσικοχημικών ιδιοτήτων των πρωτεϊνών αλλά και σε μια πρώτη χαρτογράφηση των πρωτεϊνικών αλληλεπιδράσεων και κυτταρικών λειτουργιών (Κεφάλαια 3-5).

Στο Κεφάλαιο 3 χρησιμοποιώντας μεθόδους μηχανικής μάθησης αλλά και άλλες βιοπληροφορικές αναλύσεις θα μελετήσουμε τα χαρακτηριστικά των πρωτεϊνών που εκκρίνονται μέσω της Sec μεταθετάσης. Οι πρωτεΐνες αυτές είναι γνωστό ότι εμπεριέχουν σήματα στόχευσης στο αμινοτελικό τους άκρο (σηματοδοτικό πεπτιδίο). Βασιζόμενοι στην πρόσφατη πειραματική απόδειξη ότι και τα ώριμα τμήματα των εκκρινόμενων πρωτεϊνών εμπεριέχουν σήματα στόχευσης θα οδηγηθούμε στην ανάδειξη νέων πιθανών σημάτων έκκρισης και δομικών χαρακτηριστικών.

Στο Κεφάλαιο 4 θα εστιάσουμε σε μία κατηγορία πρωτεϊνών με ιδιαίτερο βιολογικό ρόλο, τις περιφερικές πρωτεΐνες, ένα υποσύνολο που μέχρι πρόσφατα ήταν ελλιπώς χαρακτηρισμένο. Βασιζόμενοι σε καλά χαρακτηρισμένα πρωτεϊνικά σύμπλοκα και αλληλεπιδράσεις πρωτεϊνών (Intact, EcoCyc) θα προχωρήσουμε για πρώτη φορά στην χαρτογράφηση του δικτύου των κυτταρικών διεργασιών στις οποίες συμμετέχουν οι περιφερικές πρωτεΐνες.

Τέλος στο Κεφάλαιο 5 θα πραγματοποιήσουμε μια πλήρη συγκριτική ανάλυση των ιδιοτήτων των πρωτεϊνών του κυτταρικού φακέλου. Η ανάλυση αυτή θα μας βοηθήσει να προσδιορίσουμε τους περιορισμούς που υπεισέρχονται στην στάδια της ανίχνευσης των πρωτεϊνών σε μεθόδους πρωτεομικής ανάλυσης αλλά και να μοντελοποιήσουμε το εύρος ανίχνευσης διαφορετικών πειραματικών προσεγγίσεων.

Η συγκεκριμένη διατριβή καταλήγει σε ενδιαφέροντα συμπεράσματα για την κατανομή των πρωτεϊνών στο κύτταρο και παρουσιάζει μια νέα εικόνα του κυττάρου *E.coli* εμπλουτίζοντας την κατηγορία των περιφερικών πρωτεϊνών. Τέλος περιγράφει για πρώτη φορά δομικά χαρακτηριστικά των Sec εκκρινόμενων πρωτεϊνών που οδηγούν σε ενδιαφέροντα συμπεράσματα για τον μηχανισμό πρόσδεσης των πρωτεϊνών στην Sec μεταθετάση. Τα μοντέλα αυτά ανοίγουν το δρόμο για μια νέα σειρά πειραμάτων που αναμένεται να επιβεβαιώσουν τα αποτελέσματα της θεωρητικής ανάλυσης.



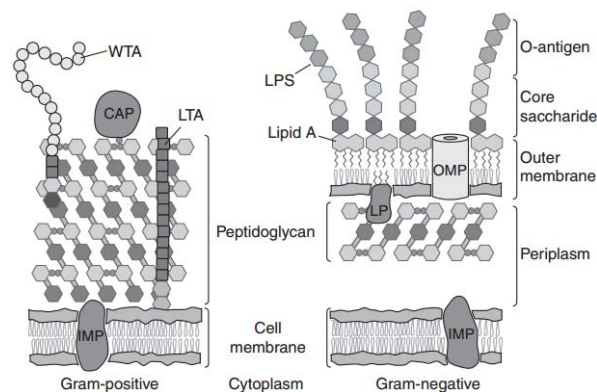


## **ΚΕΦΑΛΑΙΟ 1 Εισαγωγή**

### **1.1 Δομή του βακτηριακού κυττάρου**

Όλα τα κύτταρα έχουν εξελίξει υποκυτταρικά διαμερίσματα τα οποία οριοθετούνται με λιπιδικές μεμβράνες. Ο διαχωρισμός του κυττάρου σε διακριτούς χώρους βοηθάει στην απομόνωση και εξειδίκευση των κυτταρικών λειτουργιών, στην κατανομή των διεργασιών, στον έλεγχο της ροής και της συγκέντρωσης οργανικών και ανόργανων μορίων αλλά και στην βελτίωση της απόδοσης χημικών αντιδράσεων (Silhavy et al, 2010).

Τα βακτήρια χωρίζονται σε δύο γενικές κατηγορίες τα (Gram<sup>+</sup>) και (Gram<sup>-</sup>) με βάση την μέθοδο χρώσης κυττάρων που ανέπτυξε ο Cristian Gram το 1884. Όσα βακτήρια αποκτούν ερυθρό χρώμα λόγω της χρώσης ονομάζονται Gram θετικά (Gram<sup>+</sup>) στην αντίθετη περίπτωση ονομάζονται Gram αρνητικά (Gram<sup>-</sup>). Οι διαφορές στην ικανότητα χρώσης μαρτυρά δομικές διαφορές ανάμεσα στο κυτταρικό τοίχωμα των βακτηρίων.



**Εικόνα 1.1 - Απεικόνιση της κυτταρικής δομής των κατά Gram θετικών (Gram<sup>+</sup>) και αρνητικών (Gram<sup>-</sup>) βακτηρίων**

*CAP*: ομοιοπολικά δεμένη πρωτεΐνη, *IMP*: πρωτεΐνη ενσωματωμένη στην ΠΜ. *LP*: λιποπρωτεΐνη, *LPS*: λιποσακχαρίτης, *LTA*, τειχοϊκό οξύ; *OMP*: πρωτεΐνη της εξωτερικής μεμβράνης; *WTA*, τειχοϊκό οξύ τοίχου (Silhavy et al, 2010)

Όλα τα βακτήρια αποτελούνται από ένα υδατικό όγκο, το κυτταρόπλασμα, που περιβάλλεται από μονή (Gram<sup>+</sup>) ή διπλή (Gram<sup>-</sup>) διπλοστοιβάδα λιπιδίων. Στα Gram<sup>-</sup> βακτήρια το κυτταρόπλασμα περικλείεται από μια πολύ-επίπεδη δομή που ονομάζεται κυτταρικός φάκελος (ΚΦ) (Silhavy et al, 2010). Ο ΚΦ αποτελείται από την εσωτερική ή αλλιώς πλασματική μεμβράνη (ΠΜ), ένα λεπτό πλέγμα μορίων πεπτιδογλυκάνης (ΠΠ) και από μια επιπλέον εξωτερική

διπλοστοιβάδα λιπιδίων, την εξωτερική μεμβράνη (EM). Στην επιφάνεια της EM βρίσκονται αγκυροβολημένα μόρια λιποσακχαρίτη (LPS: lipopolysaccharide). Ανάμεσα στις δύο μεμβράνες (PM και EM) βρίσκεται το περίπλασμα, ένας πολυπληθής υδατικός όγκος που περιέχει πρωτεϊνικά μόρια (π.χ. λιποπρωτεΐνες) και άλλες μακρομοριακές δομές (π.χ. πλέγμα πεπτιδογλυκάνης) (Silhavy et al, 2010) (Εικόνα 1.1).

## 1.2 Κυτταρικός φάκελος και τα βιολογικά μόρια από τα οποία αποτελείται

Ο ΚΦ των Gram- βακτηρίων προφυλάσσει το εσωτερικού του κυτάρου και ελέγχει την ροή ανόργανων και οργανικών μορίων. Αποτελείται από διαφορετικού τύπου βιολογικά μόρια (πρωτεΐνες, λιπίδια, πεπτιδογλυκάνες, λιποσακχαρίτες) τα οποία παράγονται κυρίως στο κυτταρόπλασμα και σε ορισμένες περιπτώσεις πάνω στις μεμβράνες (Silhavy et al, 2010).

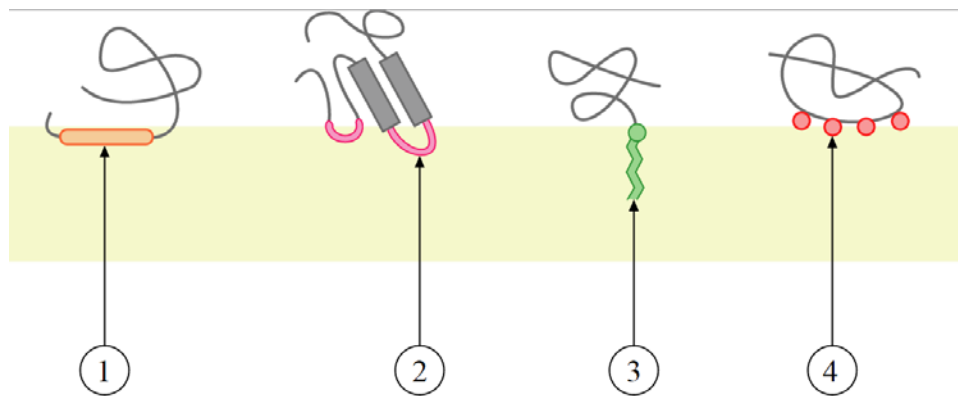
Μια πρωτεΐνη του ΚΦ μπορεί να: βρίσκεται ενσωματωμένη σε κάποια από τις μεμβράνες (διαμεμβρανικές μεμβρανικές πρωτεΐνες), να εκκρίθει στο περίπλασμα (περίπλασμικές πρωτεΐνες), να αγκυροβοληθεί στις επιφάνειες των μεμβρανών (λιποπρωτεΐνες, περιφερικές πρωτεΐνες) ή τέλος να αποβληθεί εντελώς από κύτταρο (εξωκυττάρια πρωτεΐνες).

### 1.2.1 Η πλασματική μεμβράνη και οι πρωτεΐνες που σχετίζονται με αυτήν

Η PM μεμβράνη σχηματίζεται από μία διπλή στοιβάδα λιπιδίων η οποία αποτελείται κυρίως από φωσφολιπίδια (φωσφατιδικής αιθανολαμίνη και φωσφατιδικής γλυκερόλη). Πολλές ζωτικές λειτουργίες του κυτάρου λαμβάνουν χώρα πάνω σε αυτήν όπως: παραγωγή ενέργειας, σύνθεση και έκκριση πρωτεϊνών, κυτταρική διαίρεση, βιοσύνθεση λιπιδίων, πρωτεόλυση και αποδόμηση του RNA (Papanastasiou et al, 2013). Οι πρωτεΐνες σχετίζονται με την PM με διάφορους τρόπους: είτε την διασχίζουν πλήρως (μεμβρανικές πρωτεΐνες), είτε αγκυροβολούν πάνω της μέσω λιπιδίων που προστίθενται έπειτα από χημική τροποποίηση του αμινοτελικού τους άκρου (λιποπρωτεΐνες, (Nakayama et al, 2012)) είτε τέλος με επιφανειακές αλληλεπιδράσεις (περιφερικές πρωτεΐνες, (Singer et al, 1972)).

Οι περιφερικές πρωτεΐνες αποτελούν μια ιδιαίτερη κατηγορία μορίων που ενώ είναι διαλυτά μπορούν και προσκολλώνται στις μεμβράνες. Η αλληλεπίδραση είναι επιφανειακή σε αντίθεση με τις διαμεμβρανικές πρωτεΐνες οι οποίες εισέρχονται στον υδρόφοβο πυρήνα της διπλοστοιβάδας λιπιδίων. Αγκυροβολούν στην μεμβράνη είτε μέσω αλληλεπιδράσεων με μεμβρανικές πρωτεΐνες (έμμεση αλληλεπίδραση) είτε εισχωρώντας μερικώς στην διπλοστοιβάδα

των λιπιδίων (άμεση αλληλεπίδραση) (Εικόνα 1.2). Οι τύποι άμεσης αλληλεπίδρασης με την ΠΜ είναι: α) με ηλεκτροστατικές αλληλεπιδράσεις, αφορά πρωτεΐνες που αλληλεπιδρούν με τις αρνητικά φορτισμένες κεφαλές των λιπιδίων μέσω θετικά φορτισμένων επιφανειών, β) με αμφίφιλες έλικες (Lu et al, 2013; Shih et al, 2011; Villegas et al, 2011), γ) μέσω λιπιδίων που προστίθεται με χημική τροποποίηση στο αμινοτελικό άκρο των πρωτεϊνών (Okuda et al, 2011; Paetzel et al, 2002) και δ) μέσω της έκθεσης μη πολικών βρόγχων (Headlam et al, 2003; Lomize et al, 2012) (Εικόνα 1.2).



**Εικόνα 1.2 – Περιφερικές πρωτεΐνες**

Σχηματική απεικόνιση των διαφορετικών τύπων αλληλεπίδρασης μιας περιφερικής πρωτεΐνης με την ΠΜ: 1. μέσω αμφίφιλης α-έλικας παράλληλη στο επίπεδο που ορίζει η ΠΜ (in-plane membrane helix) 2. μέσω υδρόφοβου βρόγχου (hydrophobic loop) 3. μέσω χημικά προσδεμένου λιπιδίου (covalently bound membrane lipid) (lipidation) 4. μέσω ηλεκτροστατικών και ιονικών αλληλεπιδράσεων με τα λιπίδια της ΠΜ (electrostatic or ionic interactions).

Δύο ενδιαφέροντα παραδείγματα είναι: οι περιφερικές μονάδες καναλιών που διοχετεύουν ιόντα ή βιολογικά μόρια (Lee et al, 2007a; Wissenbach et al, 1995) και ορισμένοι μεταγραφικοί παράγοντες που παραμένουν προσωρινά δεσμευμένοι στην μεμβράνη μέχρις ότου κάποιο εξωτερικό ερέθισμα τις απελευθερώσει στο κυτταρόπλασμα (Papanastasiou et al, 2013).

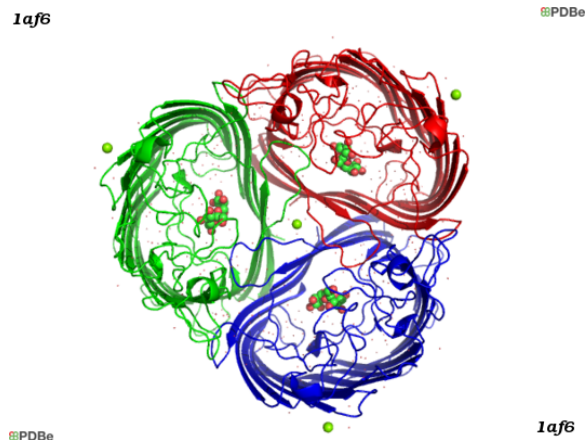
### 1.2.2. Το κυτταρικό τοίχωμα της πεπτιδογλυκάνης

Το τοίχωμα της πεπτιδογλυκάνης (ή αλλιώς μουρεΐνης) αποτελείται από επαναλαμβανόμενες μονάδες δισακχαρίτη που συνδέονται μεταξύ τους μέσω πενταπεπτιδικών αλυσίδων (Vollmer, 2008). Είναι μια πολύ σταθερή δομή η οποία καθορίζει το σχήμα του βακτηρίου (Shiomi et al, 2008b), όπως για παράδειγμα το ραβδοειδές σχήμα των εντεροβακτηρίων.

Η EM μεμβράνη είναι «συρραμμένη» πάνω στο τοίχωμα της πεπτιδογλυκάνης μέσω της λιποπρωτεΐνης Lpp (ή αλλιώς λιποπρωτεΐνης Braun) (Braun, 1975). Η Lpp είναι από τις πιο πολυπληθείς πρωτεΐνες στο κύτταρο (500.000 μόρια) ενώ εκτιμάται ότι αποτελεί το 40% του βάρους του τοιχώματος της πεπτιδογλυκάνης (Hirashima et al, 1974). Εισχωρεί στην EM μεμβράνη μέσω ενός λιπιδίου που προστίθεται στο αμινοτελικό της άκρο και δημιουργεί ομοιοπολικούς δεσμούς με το τοίχωμα της πεπτιδογλυκάνης μέσω του καρβοξυτελικού άκρου της (Inouye et al, 1974).

### 1.2.3. Η εξωτερική μεμβράνη και οι πρωτεΐνες που εμπεριέχονται σε αυτήν

Η EM αποτελεί μια ασύμμετρη διπλή στοιβάδα λιπιδίων, δηλαδή η σύσταση κάθε στοιβάδας είναι διαφορετική. Η εσωτερική στοιβάδα αποτελείται από φωσφολιπίδια ενώ η εξωτερική από γλυκολιπίδια και λιποπολυσακχαρίτες (Kamio et al, 1976). Περίπου το 50% της EM αποτελείται από πρωτεϊνικά μόρια όπως λιποπρωτεΐνες (Nakayama et al, 2012) και πορίνες (Bishop, 2008).



**Εικόνα 1.3 - Η δομή της πρωτεΐνης LamB**

Η πρωτεΐνη LamB αποτελείται από τρία αντιπαράλληλα β-βαρέλια των 18 β-κλώνων και περιέχει τρία ανεξάρτητα κανάλια.

Δύο τύποι πρωτεϊνών εισχωρούν στην EM: α) λιποπρωτεΐνες που βρίσκονται αγκυροβολημένες στην επιφάνεια της EM μέσω ενός λιπιδίου που προστίθεται στο αμινοτελικό τους άκρο και β) πρωτεΐνες που ενσωματώνονται στην EM (πορίνες, β-βαρέλια). Στην δεύτερη κατηγορία ανήκουν κυρίως πορίνες που εξυπηρετούν την διέλευση διάφορων μορίων (Koebnik et



al, 2000) ενώ ελάχιστες φαίνεται να έχουν κάποιο ενζυμικό ρόλο (Bishop, 2008; Bishop et al, 2000; Sugimura et al, 1988).

Σχεδόν αποκλειστικά οι πρωτεΐνες που ενσωματώνονται στην EM αποτελούνται από β-πυκνωτές σε αντίθεση με τις πρωτεΐνες που ενσωματώνονται στην ΠΜ που αποτελούνται μόνο από α-έλικες. Οι β-πυκνωτές επιφάνειες τυλίγονται σχηματίζοντας κυλίνδρους ή αλλιώς κλειστά βαρέλια (Koebnik et al, 2000). Για το λόγο αυτό οι πρωτεΐνες αυτές ονομάζονται και β-βαρέλια (b-barrels). Είναι γνωστό ότι κάποιες από αυτές τις πρωτεΐνες λειτουργούν ως παθητικά κανάλια (πορίνες) διέλευσης μορίων όπως μονοσακχαρίτες και αμινοξέα. Μερικά παραδείγματα είναι η πορίνη της μαλτόζης, LamB (Εικόνα 1.3; (Wang et al, 1997)) και τα συστήματα μετάδοσης ενέργειας (TonB-depended energy transducing systems, (Letain et al, 1997)).

#### 1.2.4. Σύνθεση λιποσακχαριτών και μηχανισμός έκκρισης τους

Οι λιποπολυσακχαρίτες (LPS: liposaccharides) είναι ακετυλιωμένα σακχαρολιπίδια που αποτελούν κύριο συστατικό της EM των Gram<sup>-</sup> βακτηρίων (Kamio et al, 1976). Βρίσκονται αγκυροβολημένα στην κυτταρική επιφάνεια (Εικόνα 1.1) όπου και σχηματίζουν ενός είδους κυτταρικού «φράγματος» με ρόλο την παρεμπόδιση της παθητικής διέλευσης υδρόφοβων μορίων μέσα στο κύτταρο (π.χ. αντιβιοτικά (Tamaki et al, 1971) και απορρυπαντικά (Zhang et al, 2013)).

Ένα μόριο λιποπολυσακχαρίτη αποτελείται από τρία τμήματα: α) το Α-λιπίδιο (lipid A), β) τον πυρήνα ολιγοσακχαρίτη και γ) το αντιγόνο-Ο. Σε κάποια βακτήρια το τμήμα του αντιγόνου-Ο απουσιάζει και ο λιποπολυσακχαρίτης αποτελείται μόνο από το Α-λιπίδιο και τον πυρήνα ολιγοσακχαρίτη. Η βιοσύνθεση των λιποσακχαριτών και η μεταφορά τους στην επιφάνεια του κυττάρου απαιτεί πολλά στάδια και λαμβάνει χώρα σε τρία διαφορετικά υποκυτταρικά διαμερίσματα. Το Α-λιπίδιο και το αντιγόνο-Ο συντίθενται στο κυτταρόπλασμα ενώ η προσθήκη μορίων γλυκόζης και ο σχηματισμός του ολιγοσακχαρίτη του Α-λιπιδίου καταλύεται στη ΠΜ (Osborn et al, 1972). Το πρόδρομο μόριο μεταφέρεται στην εξωτερική επιφάνεια της ΠΜ μέσω αντίστοιχου διαύλου (πρωτεΐνη MsbA). Τέλος τα μόρια του λιποπολυσακχαρίτη μεταφέρονται στην επιφάνεια του κυττάρου μέσω του μονοπατιού Lpt (Liechti et al, 2012; Polissi et al, 2014) όπου και σχηματίζουν ένα πλέγμα που εμποδίζει την διέλευση υδρόφοβων μορίων, αλληλεπιδρώντας μεταξύ τους μέσω δισθενών κατιόντων.

Πρόσφατα πειράματα έδειξαν ότι η βιοσύνθεση των οι λιποσακχαριτών δεν είναι απαραίτητη για την βιωσιμότητα του κυττάρου (Moffatt et al, 2010) έχει όμως σημαντικές επιπτώσεις στην συναρμολόγηση άλλων στοιχείων του ΚΦ όπως κάποιες πρωτεΐνες της EM (π.χ. πρωτεΐνε OmpA, OmpC και OmpF (Ried et al, 1990)).

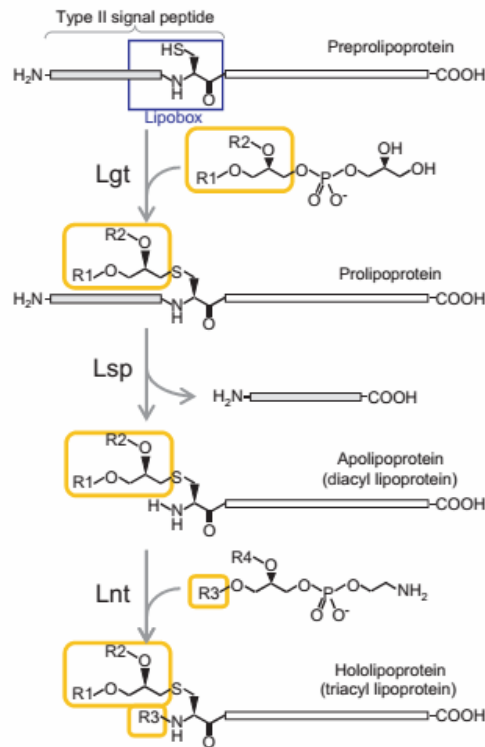
Εικάζεται ότι οι λιποπολυσακχαρίτες διευκολύνουν την συναρμολόγηση πρωτεϊνών της EM (πορίνες) και ότι λειτουργούν ως μοριακοί οδηγοί αυτών (Zhang et al, 2013). Δύο γνωστά παραδείγματα είναι η πορίνη OmpF η οποία αποκτά την τριμερή δομή της παρουσία λιποπολυσακχαριτών (de Cock et al, 2001) και η πορίνη PhoE (Hagge et al, 2002) της οποίας η λειτουργία ως διάυλος επηρεάζεται από την δομή που αποκτούν οι λιποπολυσακχαρίτες (Hagge et al, 2002). Επίσης η πρωτεάση OmpT η οποία γίνεται ενζυμικά ενεργή μόνο παρουσία λιποπολυσακχαριτών (Kramer et al, 2002). Η αναστολή της βιοσύνθεσης των λιποπολυσακχαριτών οδηγεί σε συσσώρευση μη σωστά αναδιπλωμένων πρωτεϊνών της EM (Collinet et al, 2000).

#### 1.2.5. Σύνθεση λιποπρωτεϊνών

Η πρώτη λιποπρωτεΐνη ανακαλύφθηκε το 1973 από τους Hantke και Braun (Hantke et al, 1973), όπου έδειξαν για πρώτη φορά με βιοχημικά δεδομένα ότι η ώριμη πρωτεΐνη Lpp περιέχει ένα κατάλοιπο τροποποιημένης κυστεΐνης (S-diacylglyceryl cysteine) με τρία μόρια λιπαρών οξέων. Προς τιμή της πρωτοποριακής αυτής δουλειάς η πρωτεΐνη Lpp ονομάστηκε και Braun's lipoprotein. Βιοχημικές αναλύσεις που ακολούθησαν ανέδειξαν περισσότερες πρωτεΐνες που τροποποιούνται χημικά με λιπίδια οι οποίες ονομάστηκαν λιποπρωτεΐνες. Οι λιποπρωτεΐνες φέρουν ΣΠ τύπου II, το οποίο χαρακτηρίζεται από ένα συντηρημένο κατάλοιπο κυστεΐνης στο αμινοτελικό άκρο. Στο συγκεκριμένο κατάλοιπο προστίθεται το μόριο λιπιδίου, έπειτα από χημική τροποποίηση, το οποίο λειτουργεί ως «άγκυρα» των πρωτεϊνών στις μεμβράνες.

Στην Εικόνα 1.4 αναπαριστάται το κανονικό μονοπάτι βιοσύνθεσης των λιποπρωτεϊνών στα βακτήρια. Τα τρία διαδοχικά βήματα βιοσύνθεσης των λιποπρωτεϊνών στα βακτήρια καταλύονται από τρία ένζυμα: την πρωτεΐνη Lgt (prelipoprotein diacylglycerol transferase), την Lsp (prelipoprotein signal peptidase) και την Lnt (apolipoprotein N-acyltransferase) (Nakayama et al, 2012). Το ένζυμο Lgt βρίσκεται ενσωματωμένο στην πλασματική μεμβράνη και η δράση του πάνω στις λιποπρωτεΐνες βασίζεται στην ομάδαθειόλης του συντηρημένου κατάλοιπου κυστεΐνης που βρίσκεται στην θέση +1 του συντηρημένου μοτίβου πάνω στο ΣΠ τύπου II (lipobox). Το

μοτίβο αυτό είναι της μορφής [LVI]<sub>3</sub>[ASTVI]<sub>2</sub>[GAS]<sub>1</sub>[C]<sub>+1</sub>. Η πρωτεάση Lsp ή αλλιώς πεπτιδάση του ΣΠ τύπου II, καταλύει την διάσπαση του πεπτιδικού δεσμού ανάμεσα στο +1 κατάλοιπο της κυστεΐνης και στο ΣΠ τύπου II. Βρίσκεται ενσωματωμένη στην ΠΜ και δέχεται ως υποστρώματα τις τροποποιημένες από την Lgt λιποπρωτεΐνες (Hussain et al, 1980).



**Εικόνα 1.4 - Λιποπρωτεΐνες στα βακτήρια και το κανονικό μονοπάτι βιοσύνθεσης (Nakayama et al, 2012)**

Η προπρωλιποπρωτεΐνη (preprolipoprotein) μεταφέρεται στο εσωτερικό της ΠΜ μέσω των εκκριτικών συστημάτων Sec (ή TAT σε κάποια Gram<sup>+</sup> βακτήρια). Στην συνέχεια το ένζυμο Lgt (preprolipoprotein diacylglyceryl transferase) μεταφέρει ένα τμήμα (diacylglyceryl) του φωσφολιπιδίου στην ομάδα της θειόλης (-SH) του +1 κατάλοιπου κυστεΐνης του συντηρημένου μοτίβου (lipobox) μέσα στο ΣΠ τύπου II, σχηματίζοντας ένα θειοεστερικό δεσμό. Έπειτα η πρωτεΐνη Lsp (prolipoprotein signal peptidase) κόβει το ΣΠ στο αμινοτελικό άκρο της τροποποιημένης κυστεΐνης (S-diacylglyceryl cysteine) της προλιποπρωτεΐνης (prolipoprotein). Η τροποποιημένη πρωτεΐνη που προκύπτει ονομάζεται δις-ακυλιωμένη λιποπρωτεΐνη ή λιποπρωτεΐνη-α. Τέλος το ένζυμο Lnt (apolipoproteinN-acyltransferase) προσθέτει ένα επιπλέον ομάδα ακυλίου (acyl group) από ένα άλλο φωσφολιπίδιο στην α-αμινοξική ομάδα της

τροποποιημένης κυστεΐνης (S-diacylglyceryl cysteine) της λιποπρωτεΐνη-α προκύπτοντας η τρις-ακυλιωμένη λιποπρωτεΐνη

### 1.3 Συστήματα έκκρισης πρωτεϊνών

Πάνω από το ένα τρίτο των πρωτεϊνών που παράγονται στο κυτταρόπλασμα μεταφέρεται στις μεμβράνες και πέρα από αυτές (Chatzi et al, 2013). Στα βακτήρια οι πρωτεΐνες που παράγονται στο κυτταρόπλασμα καταλήγουν σε διαφορετικές τοποθεσίες μέσα στο κύτταρο, όπως για παράδειγμα στο περίπλασμα στην εξωτερική μεμβράνη (Gram<sup>-</sup> βακτήρια), στην εξωκυττάρια επιφάνεια και στον εξωκυττάριο χώρο.

Τα βακτήρια έχουν αναπτύξει τουλάχιστον 16 διαφορετικούς μηχανισμούς έκκρισης και στόχευσης των πρωτεϊνών (Paranikou et al, 2007). Μόνο τα συστήματα έκκρισης Sec και Tat είναι πανταχού παρόντα, δηλαδή τα συναντάμε και στις τρεις επικράτειες της ζωής: βακτήρια,

αρχαία, ευκαρυώτες). Την ανάγκη ύπαρξης διαφορετικών συστημάτων έκκρισης υπαγορεύουν η ταχύτητα αναδίπλωσης των πρωτεϊνών αλλά και ο τελικός προορισμός τους. Τα περισσότερα συστήματα έκκρισης είναι ολιγομερικά (5-8 πρωτεΐνες, π.χ. Sec και TAT; (Chatzi et al, 2014)) ενώ ελάχιστα αποτελούν μεγάλους πρωτεϊνικούς σχηματισμούς (>20 πρωτεΐνες, π.χ μαστίγια και ινίδια) (Mota et al, 2005; Van Gerven et al, 2011).

Παρακάτω αναφέρουμε περιληπτικά τα 12 πιο καλά χαρακτηρισμένα συστήματα έκκρισης στα βακτήρια. Τα εκκριτικά συστήματα μπορούν να χωριστούν σε δύο βασικές κατηγορίες σε αυτά που εξαρτώνται από το σύστημα Sec και σε όσα λειτουργούν ανεξάρτητα (Orfanoudaki et al, 2014).

#### 1.4 Η μεταθετάση Sec

Η μεταθετάση Sec σχηματίζεται είτε από δύο αντίγραφα του μεμβρανικού συμπλόκου SecYEG (ομοδιμερές SecYEG) είτε από ένα σύμπλοκο SecYEG και τις βοηθητικές υπομονάδες SecDF–YajC–YidC ή μόνο την YidC (Schulze et al, 2014). Η μεταθετάση Sec εκκρίνει μόνο μη εγγενώς αναδιπλωμένες πρωτεΐνες με συνμεταφραστικό ή μετα-μεταφραστικό τρόπο. Η μετα-μεταφραστική μετατόπιση καθοδηγείται από την υπομονάδα SecA, την πρωτεΐνη κινητήρα του συστήματος (Chatzi et al, 2014). Η συνμεταφραστική μετατόπιση απαιτεί την πρωτεΐνη Srp (signal recognition particle) που μαζί με το μικρό ριβοσωμικό RNA 4.5S αποτελούν την πρωτεΐνη Ffh. Η πρωτεΐνη Srp προσκολλάται στην ΠΜ μεμβράνη μέσω του υποδοχέα FtsY που στο *E.coli* αποτελεί μια περιφερική πρωτεΐνη που βρίσκεται σε ισορροπία ανάμεσα στην διαλυτή και μεμβρανική της κατάσταση (Grudnik et al, 2009; von Loeffelholz et al, 2013). Το μονοπάτι SRP χρησιμοποιείται κυρίως για την ενσωμάτωση μεμβρανικών πρωτεϊνών στην ΠΜ (Neumann-Haefelin et al, 2000). Σε ορισμένες περιπτώσεις φαίνεται ότι το μεμβρανικό κανάλι SecYEG συνεργάζεται με την πολυτοπική πρωτεΐνη YidC για την εισαγωγή μεμβρανικών πρωτεϊνών (Kudva et al, 2013). Επίσης η YidC λειτουργεί και αυτόνομα (Dalbey et al, 2014; Kudva et al, 2013). Σε αντίθεση με τη μεταθετάση Sec το εκκριτικό σύστημα TAT, το οποίο αποτελείται από τις υπομονάδες TatABC, εκκρίνει πλήρως αναδιπλωμένες πρωτεΐνες chaperones (Lee et al, 2006). Στο σύστημα Tat υπάρχουν αντίστοιχες πρωτεΐνες με βοηθητικό ρόλο ως μοριακές οδηγίες (Lee et al, 2006).

## 1.1 Sec εξαρτώμενα συστήματα έκκρισης

Ο ρόλος των Sec εξαρτώμενων συστημάτων έκκρισης αφορά κατά κανόνα την στόχευση πρωτεϊνών στη EM (στα Gram<sup>-</sup> βακτήρια) και η έκκριση τους πέρα από αυτή, αφού το Sec σύστημα έχει ολοκληρώσει την μεταφορά τους έξω από την ΠΜ. Το σύστημα Sec στα βακτήρια αναλαμβάνει επίσης την ενσωμάτωση πρωτεϊνών στην ΠΜ μέσω μιας δεύτερης πλευρικής πύλης (Chatzi et al, 2013). Η πλειονότητα των πρωτεϊνών διασχίζει την ΠΜ μέσω αυτού του συστήματος, πολύ λίγες πρωτεΐνες χρησιμοποιούν το σύστημα TAT ενώ ελάχιστες τα υπόλοιπα συστήματα έκκρισης.

Οι λιποπρωτεΐνες διοχετεύονται πέρα από την ΠΜ μέσω του συστήματος Sec. Στη συνέχεια το σύστημα Lol καταλύει την διαλογή των λιποπρωτεϊνών στην EM ή ΠΜ. Το σύμπλοκο LolCDE βρίσκεται στην ΠΜ και καταλύει την απελευθέρωση των λιποπρωτεϊνών από αυτήν. Η περιπλασμική πρωτεΐνη LolA «συλλαμβάνει» τις λιποπρωτεΐνες που απελευθερώνονται από την ΠΜ στο περίπλασμα και τις μεταφέρει στην EM (Miyamoto et al, 2007). Η πρωτεΐνη LolB αποτελεί τον υποδοχέα του συμπλόκου LolA-λιποπρωτεΐνης στην EM. Βοηθητικές μονάδες που συμμετέχουν στην ωρίμανση των λιποπρωτεϊνών είναι: α) η πεπτιδάση του ΣΠ τύπου II LspA (SPase II) η οποία καταλύει την αποκοπή του ΣΠ (Yu et al, 1984), β) η πρωτεΐνη Lgt (preprolipoprotein diacylglycerol transferase) η οποία τροποποιεί χημικά το αμινοτελικό κατάλοιπο της Κυστεΐνης (Sankaran et al, 1994) και γ) η πρωτεΐνη Lnt (N-acyl transferase) η οποία μεταφέρει στις λιποπρωτεΐνες μια επιπλέον αλυσίδα ακυλίου (acyl chain) (Gupta et al, 1991).

Οι ενσωματωμένες στην EM πρωτεΐνες, αποτελούνται σχεδόν αποκλειστικά από β-πυχωτές και σχηματίζουν β-βαρέλια. Οι πρωτεΐνες εξαρτώνται από την μεταθετάση Sec για τη έκκριση τους στο περίπλασμα και στην συνέχεια αναδιπλώνονται και ενσωματώνονται στην μεμβράνη μέσω του συμπλόκου BAM (Solon'eva et al, 2012). Η σωστή αναδίπλωση των πρωτεϊνών που σχηματίζουν β-βαρέλια απαιτεί την παρουσία περιπλασμικών βοηθών όπως η SurA (peptidyl prolyl cis-trans isomerases), η PpiD (Stymest et al, 2008) ή πρωτεΐνες που μεσολαβούν για τον σχηματισμό δισουλφιδικών δεσμών π.χ. DsbA και DsbB (Goemans et al, 2013). Για παράδειγμα η περιπλασμική μοριακή οδηγός Skp, αλλά και τα μόρια του λιποσακχαρίτη (LPS) χρειάζονται για την αποτελεσματική εισαγωγή της πρωτεΐνης OmpA στη διπλοστοιβάδα λιπιδίων όπως αποδεικνύεται σε *in vitro* πειράματα.

Ένα υποσύνολο πρωτεϊνών της εξωτερικής μεμβράνης φαίνεται να μπορεί να εισάγει τον εαυτό του στην EM (AT: autotransporters)(Henderson et al, 2000; Ieva et al, 2008). Τον τελευταίο καιρό υπάρχουν αυξανόμενες ενδείξεις ότι κάποιες από αυτές χρησιμοποιούν εξειδικευμένα συστήματα όπως ο μηχανισμός TAM (Selkrig et al, 2012).

Τα μονοπάτια τύπου chaperone–usher (CU) είναι γνωστό ότι συναρμολογούν και εκκρίνουν υπομονάδες από οργανίδια που σχηματίζονται στην επιφάνεια του κυττάρου, γνωστά και ως τριχίδια (pili, fimbriae), τα οποία επιτελούν ρόλους προσκόλλησης σε άλλες επιφάνειες. Υπάρχουν διαφορετικές παραλλαγές εκκριτικών μονοπατιών τύπου CU. Κάθε μονοπάτι αποτελείται από μια ομάδα πρωτεϊνών που κωδικοποιούνται στο ίδιο οπερόνιο και συμπεριλαμβάνουν: μία τουλάχιστον πρωτεΐνη συνοδό (usher), ένα μοριακό οδηγό (chaperone) και την αντίστοιχη υπομονάδα του τριχιδίου (Busch et al, 2012).

Ένα άλλο είδος εξωκυττάρων τριχιδίων είναι τα curli. Τα curli είναι αμυλοειδής ίνες και σχετίζονται με την προσκόλληση σε επιφάνειες και το σχηματισμό αλυσίδων από βακτηριακά κύτταρα γνωστά και ως βιοφιλμ (Barnhart et al, 2006). Οι πρωτεΐνες CsgA και CsgB αποτελούν τις δομικές μονάδες αυτών των αμυλοειδών σχηματισμών (Nenninger et al, 2009). Κριτικής σημασίας για την έναρξη του πολυμερισμού των CsgA και CsgB αποτελεί η υπομονάδα CsgF η οποία επιτελεί ρόλο μοριακής οδηγού. Η CsgF εκκρίνεται στην επιφάνεια του κυττάρου και βοηθάει στην σωστή τοποθέτηση της CsgB (Nenninger et al, 2009). Οι τρεις αυτές υπομονάδες χρειάζονται την βοήθεια της μοριακής οδηγού CsgE αλλά και της μεμβρανικής υπομονάδας CsgG για την έκκριση τους (Nenninger et al, 2009).

Οι διαλυτές πρωτεΐνες ενδέχεται να απελευθερώνονται στον εξωκυττάριο χώρο μέσω μεμβρανικών κυστιδίων της EM (MKEM) (OMVs : outer membrane vesicles) (Beveridge, 1999). Τα MKEM μεσολαβούν για την ενδοκυτταρική επικοινωνία (Schertzer et al, 2013). Είναι σφαιρικά σωματίδια με διπλή στοιβάδα πρωτεολιπιδίων και περιέχουν πρωτεΐνες της EM. Πρωτεΐνες όπως οι OmpA, OmpC, OmpF, και OmpW έχουν αναγνωριστεί πειραματικά ως στοιχεία των εγγενών κυστιδίων της εξωτερικής μεμβράνης των κυττάρων DH5a (Lee et al, 2007b). Μάλιστα οι πρωτεΐνες OmpF, OmpC, και OmpX μέσω ενός άγνωστου μηχανισμού, διευκολύνουν την έκκριση της πρωτεΐνης YebF (Prehna et al, 2012). Πρόσφατα αποδείχθηκε ότι η FliC και η πρωτεΐνη των μαστιγίων FlgK προάγουν την παραγωγή των MKEM (Manabe et al, 2013).

Τέλος το σύστημα έκκρισης τύπου II (T2SS) είναι μια νανομηχανή η οποία διασχίζει και τις δύο μεμβράνες και αποτελείται από ~14 πρωτεΐνες (Francetic et al, 2000) (Orfanoudaki et al, 2014). Όταν ενεργοποιείται, εκκρίνει αναδιπλωμένες πρωτεΐνες από το περίπλασμα στον εξωκυττάριο χώρο (Korotkov et al, 2012). Η βασική εκκριτική υπομονάδα είναι η GspG (pseudopilin) η οποία εκκρίνεται από το κυτταρόπλασμα στο περίπλασμα μέσω του μονοπατιού Sec-SRP (Korotkov et al, 2012). Οι υπόλοιπες υπομονάδες του συστήματος είναι οι πρωτεΐνες GspH, GspI, GspJ, και GspK οι οποίες πιθανά να ακολουθούν επίσης το εκκριτικό μονοπάτι Sec-SRP.

## 1.2 Μη εξαρτώμενα από το Sec συστήματα έκκρισης

Το βακτηριακό μαστίγιο αποτελεί μια ανεξάρτητη μηχανή έκκρισης που αποτελείται από ~26 πρωτεΐνες και επίσης διαπερνά και τις δύο κυτταρικές μεμβράνες (Van Gerven et al, 2011). Τα μαστίγια εξελικτικά και λειτουργικά σχετίζονται με το σύστημα έκκρισης τύπου III (T3S system) που συναντάται στα παθογόνα στελέχη του *E. coli* (π.χ. το εντεροπαθογόνο *E. coli*, EPEC (Chen et al, 2013)) αλλά και σε άλλα Gram- βακτήρια (Cornelis, 2006). Το T3SS «χτίζει» ένα διαμεμβρανικό μεγαλομοριακό μηχανισμό ("injectisome") που το βακτήριο τον χρησιμοποιεί για να «αγκυροβολήσει» σε κύτταρα ξενιστές και να εμβολιάσει παθογόνους παράγοντες κατευθείαν στο εσωτερικό των κυττάρων ξενιστών (Blocker et al, 2003). Το μαστίγιο σχηματίζεται αντίστοιχα από διάφορες υπομονάδες οι οποίες κατά κύριο λόγο χρησιμοποιούν το σύστημα Sec για να εκκριθούν και να χτίσουν τον διαμεμβρανικό μηχανισμό. Εξωκυττάρια υπομονάδες του συστήματος όπως αυτές που σχηματίζουν το «αγκίστρι», εκκρίνονται διαμέσου ενός πόρου που σχηματίζει ο ίδιος ο μηχανισμός, ενώ εξειδικευμένες πρωτεΐνες βοηθούν στην στόχευση τους στην μεμβράνη (Badea et al, 2009).

Εκκριτικό σύστημα τύπου TAT (the twin-arginine translocation pathway) το οποίο είναι υπεύθυνο για την έκκριση αναδιπλωμένων πρωτεϊνών πέραν της ΠΜ. Αποτελείται από τις μεμβρανικές πρωτεΐνες TatA, TatB, και TatC ενώ δεν έχει διευκρινιστεί πλήρως η αλληλουχία των γεγονότων που οδηγούν στη μετατόπιση των πρωτεϊνών αλλά και ο μηχανισμός που απαγορεύει την διέλευση μη αναδιπλωμένων πρωτεϊνών (Lee et al, 2006). 29 πιθανά υποστρώματα του συστήματος Tat έχουν επιβεβαιωθεί και πειραματικά (Tullman-Ercek et al, 2007).

## 1.5 Χαρακτηριστικά εκκρινόμενων πρωτεϊνών

### 1.1.1. Σηματοδοτικό πεπτιδίο

Το 1971 οι συνεργάτες Günter Blobel και David Sabatini διατύπωσαν για πρώτη φορά ότι πληροφορία η οποία βρίσκεται στο αμινοτελικό άκρο των εκκρινόμενων πρωτεϊνών είναι υπεύθυνη για την στόχευση τους στην μεμβράνη (Sabatini et al, 1971). Για την ανακάλυψη αυτή ο Günter Blobel έλαβε το 1999 το νόμπελ Φυσιολογίας και Ιατρικής. Λίγο αργότερα επιβεβαιώθηκε ότι και στα βακτήρια υπάρχουν αντίστοιχες αλληλουχίες στόχευσης (Sekizawa et al, 1977) για την έκκριση πέρα από την πλασματική μεμβράνη, οι οποίες ονομάστηκαν σηματοδοτικά πεπτιδία (ΣΠ). Οι πρωτεΐνες που συνθέτονται έχοντας ΣΠ ονομάζονται προ-πρωτεΐνες (ή πρόδρομες μορφές).

Με την αποκοπή του σηματοδοτικού πεπτιδίου προκύπτουν οι ώριμες πρωτεΐνες, ή ώριμα τμήματα (ΩΤ). Το στάδιο αυτό είναι απαραίτητο πριν απελευθερωθεί η πρωτεΐνη στο περιπλασμα και γίνεται μέσω κατάλληλων ενζύμων (πρωτεάσες) οι οποίες ονομάζονται πεπτιδάσες. Τρεις γνωστές πεπτιδάσες υπάρχουν στα βακτήρια: η **πεπτιδάση τύπου I** (SPase I), η **πεπτιδάση τύπου II** (SPase II) και η τύπου IV (SPase IV). Οι μη λιποπρωτεΐνες που εκκρίνονται μέσω του συστήματος Sec ονομάζονται **εκκρινόμενες πρωτεΐνες τύπου I** καθώς αναγνωρίζονται από την SPase I ενώ οι λιποπρωτεΐνες ονομάζονται **εκκρινόμενες πρωτεΐνες τύπου II** διότι το σηματοδοτικό τους πεπτιδίο κόβεται αντίστοιχα από την SPase II (Paetzel et al, 2002)

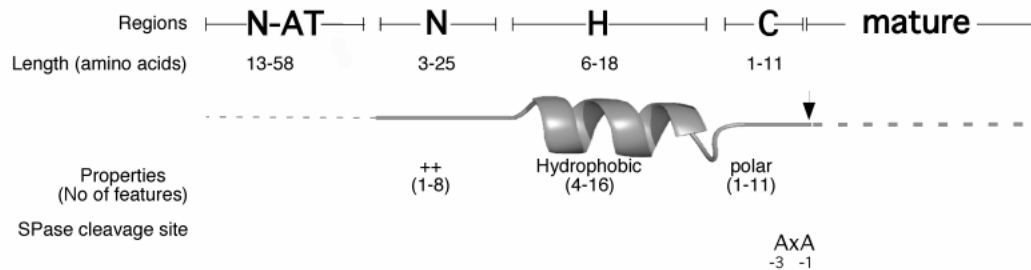
### 1.2.1. Σηματοδοτικά πεπτιδία τύπου I

Τα ΣΠ τύπου I αναγνωρίζονται από την SPase I στις θέσεις -1 και -3 όπου υπάρχει συντηρημένο το μοτίβο [AGSCT]<sub>-1</sub>[X]<sub>-2</sub>[AGSCIVL]<sub>-3</sub> (αλλιώς κανόνας -3, -1) (von Heijne, 1983). Τα σημεία αυτά έχουν επιβεβαιωθεί με σημειακές μεταλλάξεις (Collier et al, 1989; Karamyshev et al, 1998). Οι θέσεις -2 και +1 δεν είναι κριτικής σημασίας καθώς όλα τα αμινοξέα είναι ανεκτά εκτός από Προλίνη στην θέση +1 (Nilsson et al, 1992). Έχει προταθεί σαν μοντέλο ότι στα σημεία -3 και -1 του ΣΠ σχηματίζεται μια στροφή της έλικας η οποία εισχωρεί σε συγκεκριμένη κοιλότητα της πεπτιδάσης (Choo et al, 2008b).

Παρόλο που τα ΣΠ έχουν κοινό ρόλο, οι αμινοξικές αλληλουχίες τους έχουν χαμηλή ομολογία, δηλαδή δεν υπάρχουν συντηρημένα αμινοξέα σε συγκεκριμένες θέσεις (Nesmeyanova et al, 1997). Παρουσιάζουν όμως συντηρημένες φυσικοχημικές ιδιότητες σε συγκεκριμένες



περιοχές και μπορούν να χωριστούν σε τρία τμήματα: την αμινοτελική περιοχή (**N-domain**) που περιέχει κυρίως θετικά φορτισμένα αμινοξέα, την υδρόφοβη περιοχή (**H-domain**) και την καρβοξυτελική περιοχή (**C-domain**) που περιέχει τις θέσεις αναγνώρισης και αποκοπής από τις πεπτιδάσες (Εικόνα 1.5).



**Εικόνα 1.5 - Σηματοδοτικό πεπτιδίο τύπου I**

Στην παραπάνω εικόνα συνοψίζονται τα χαρακτηριστικά των ΣΠ. 425 εκκρινόμενες προ-πρωτεΐνες συλλέχθηκαν από τις βάσεις δεδομένων EchoLOCATION και Uniprot (Dimmer et al, 2012; Horler et al, 2009) και αναλύθηκαν. Η εκτίμηση των περιοχών N, H και C αλλά και του σημείου αποκοπής των από την SPase I έγινε με το λογισμικό SignalP 3 (Bendtsen et al, 2004). Μια υποκατηγορία πρωτεϊνών της εξωτερικής μεμβράνης που αποκαλούνται αυτομεταφορείς διαθέτουν επιμηκυμένες N περιοχές (N-AT) (Hiss et al, 2009). Μία τυπική περιοχή N διαθέτει αμινοξέα με θετικό φορτίο και ακολουθείται από έναν υδρόφοβο πυρήνα (H) και τελικά από μια καρβοξυτελική περιοχή με πολικά αμινοξέα. Τέλος η περιοχή που αναγνωρίζεται από την πρωτεάση των πεπτιδίων σημάτων διαθέτει καλά συντηρημένα αμινοξέα, συνήθως μια Αλανίνη ή Σερίνη. (G. Orfanoudaki, M. Papanastasiou and A. Economidou, μη δημοσιευμένα δεδομένα). Στην συνέχεια ακολουθεί το ώριμο τμήμα της προ-πρωτεΐνης.

Οι σημασία των τριών αυτών περιοχών στην στόχευση αλλά και έκκριση των πρωτεϊνών έχει εξεταστεί χρησιμοποιώντας σαν μοντέλο κυρίως την πρωτεΐνη της αλκαλικής φωσφατάσης (PhoA) (Kajava et al, 2000; Kim et al, 2000; Laforet et al, 1989; Nesmeyanova et al, 1997) και σπάνια άλλες πρωτεΐνες όπως η MalE (maltose binding protein) (Bedouelle et al, 1980) και η MBP (maltose binding protein) (Bankaitis et al, 1984). Η αλκαλική φωσφατάση είναι ένα ένζυμο το οποίο εκκρίνεται στο περίπλασμα όπου και καταλύει την υδρόλυση και την φωσφορυλίωση μεγάλης ποικιλίας φωσφορικών μονο-εστέρων. Η έκκριση της είναι εύκολο να διαπιστωθεί μέσω της ενζυμικής της δραστηριότητας στο περίπλασμα.

Όταν φορτισμένα κατάλοιπα προστέθηκαν στην υδρόφοβη περιοχή H του ΣΠ της πρωτεΐνης MBP οδήγησαν σε προβληματική έκκριση. Αντίστοιχος φαινότυπος παρατηρήθηκε όταν υδρόφοβα κατάλοιπα αφαιρέθηκαν από την ίδια περιοχή (Puziss et al, 1989). Τα θετικά φορτισμένα αμινοξέα στο περιοχή N των ΣΠ φαίνεται να επηρεάζουν την έκκριση σε συνδυασμό

με μεταλλάξεις στην υδρόφοβη περιοχή (Puziss et al, 1989) ή και από μόνα τους (Nesmeyanova et al, 1997). Η αντικατάσταση των θετικά φορτισμένων αμινοξέων στην περιοχή N φαίνεται μειώνουν το ποσό της εκκρινόμενης πρωτεΐνης (Nesmeyanova et al, 1997).

Τέλος το ΣΠ πεππίδιο δεν είναι ικανό από μόνο του να οδηγήσει σε έκκριση οποιαδήποτε πρωτεΐνη καθώς το παράγωγο που προκύπτει από την σύντηξη μιας κυτταροπλασματικής πρωτεΐνης με ΣΠ συχνά δεν εκκρίνεται (Boyd et al, 1990; Kadonaga et al, 1984; Moreno et al, 1980). Επίσης παρά την ομοιότητα που παρουσιάζουν τα ΣΠ των βακτηρίων και των ευκαρυωτικών οργανισμών, δεν μπορούν όλες οι πρωτεΐνες που φυσιολογικά εκκρίνονται στους ευκαρυώτες να οδηγηθούν σε έκκριση σε ένα βακτήριο ακόμα και με προκαρυωτικό ΣΠ. Όλα τα παραπάνω συνηγορούν στο γεγονός ότι χαρακτηριστικά που βρίσκονται στο ΩΤ των πρωτεϊνών συνεισφέρουν ή παρακωλύουν την εκκριτική διαδικασία.

### 1.1.2. Ωριμο τμήμα

Πρόσφατα βιοχημικά δεδομένα συνηγορούν στο συμπέρασμα ότι το ΩΤ των εκκρινόμενων πρωτεϊνών περιέχει σήματα στόχευσης των πρωτεϊνών στη μεταθετάση Sec (Gouridis et al, 2009). Τα χαρακτηριστικά αυτά δεν έχουν ακόμα πλήρως διευκρινιστεί. Παλαιότερη στατιστική ανάλυση έχει δείξει ότι τα ΩΤ των Sec εκκρινόμενων πρωτεϊνών παρουσιάζουν στατιστικά σημαντική απόκλιση στο καθαρό συνολικό φορτίο και συγκεκριμένα ότι στα Gram<sup>-</sup> το φορτίο αυτό είναι αρνητικό (Kajava et al, 2000). Μεταλλάξεις στο ΩΤ του περιπλασματικού ενζύμου της αλκαλικής φωσφατάσης (PhoA) όπου βασικά κατάλοιπα (Αργινίνη και Λυσίνη) προστέθηκαν στις θέσεις +2 μέχρι +30 του ΩΤ, έδειξαν ότι τα πρώτα 1-14 αμινοξέα του ΩΤ δεν επιδέχονται θετικά φορτία ενώ προβληματική αποδείχτηκε η εισαγωγή Αργινίνης ακόμα και έπειτα από την θέση +30 (Kajava et al, 2000).

Κυτταροπλασματικές πρωτεΐνες με ρόλο μοριακών οδηγών γενικά πιστεύεται ότι αλληλεπιδρούν με το ΩΤ των εκκρινόμενων πρωτεϊνών για να εμποδίσουν την αναδίπλωση των πρωτεϊνών και για να στοχεύσουν τις πρωτεΐνες στις μεμβράνες και το εκκριτικό κανάλι (Khisty et al, 1995; Topping et al, 1994). Για ορισμένες εκκρινόμενες πρωτεΐνες δεν απαιτείται η παρουσία μοριακών οδηγών για να εκκριθούν, όπως στην περίπτωση της αλκαλικής φωσφατάσης (Kim et al, 2000). Θετικά φορτισμένα ή υδρόφοβα αμινοξέα στο πρώιμο ΩΤ τμήμα οδηγούν σε προβληματική έκκριση η οποία διορθώνεται με την παρουσία της μοριακής οδηγού SecB (Kim et al, 2000). Παρά τα βιοχημικά πειράματα που έχουν γίνει σε εκκριτικές πρωτεΐνες δεν έχει βρεθεί

μέχρι σήμερα κάποιο συντηριμένο μοτίβο στα ΩΤ που να έχει σχετιστεί και πειραματικά με μηχανισμούς στόχευσης στην μεμβράνη.



## **ΚΕΦΑΛΑΙΟ 2 Αποτελέσματα – Μελέτη της κατανομής των πρωτεϊνών του βακτηρίου *E.coli* στα διάφορα υποκυτταρικά διαμερίσματα**

Όλα τα κύτταρα αποτελούνται από διακριτές περιοχές που οριοθετούνται μέσω βιολογικών μεμβρανών. Η διαμερισματοποίηση του κυττάρου εξυπηρετεί πολλαπλούς ρόλους όπως την απομόνωση και εξειδίκευση κυτταρικών διεργασιών, την κατανομή φόρτου εργασίας, τον έλεγχο διέλευσης μορίων και τον έλεγχο μακρομοριακών συγκεντρώσεων με στόχο τη ρύθμιση της απόδοσης χημικών αντιδράσεων.

Εκτιμάται ότι το ένα τρίτο των πρωτεϊνών που παράγονται στο εσωτερικό ενός κυττάρου οδηγείται σε άλλα υποκυτταρικά διαμερίσματα. Οι πρωτεΐνες ενός κυττάρου συμμετέχουν σε μεγάλη ποικιλία κυτταρικών διεργασιών όπως η βιογένεση βιολογικών μεμβρανών, η διατήρηση της κυτταρικής δομής, η εισροή και εκροή βιολογικών μορίων και η διακυτταρική επικοινωνία μέσω της μεταγωγής σημάτων. Επιπλέον, πρωτεΐνες που βρίσκονται ή δρουν στην επιφάνεια του κυττάρου συχνά αποτελούν καθοριστικούς παράγοντες παθογένειας. Συνεπώς ο καθορισμός της κατανομής των πρωτεϊνών στα διάφορα υποκυτταρικά διαμερίσματα αλλά και του δικτύου πρωτεϊνικών αλληλεπιδράσεων αποτελεί μια απαραίτητη προεργασία για την κατανόηση και μελέτη του κυττάρου ως ολότητα.

Η πειραματική ανίχνευση πρωτεϊνών και η παρακολούθηση του τρόπου/τόπου που αυτές μετακινούνται μέσα στο κύτταρο καθίσταται δυνατή μέσω τεχνικών όπως η μικροσκοπία επισημασμένων πρωτεϊνών. Η τεχνική αυτή δυστυχώς έχει εφαρμοστεί για μεμονωμένες περιπτώσεις πρωτεϊνών και όχι για το σύνολο του πρωτεϊνώματος του *E.coli* (Daley et al, 2005; Reyes-Lamothe, 2012). Μειονεκτήματα της μεθόδου είναι η παρακώλυση της επισημασμένης πρωτεΐνης από την φυσιολογική της λειτουργία/μετακίνησης.

Μια άλλη τεχνική ανίχνευσης πρωτεϊνών ανά υποκυτταρικό διαμέρισμά είναι η *in vitro* κλασματοποίηση του κυττάρου (Ishihama et al, 2008; Masuda et al, 2009; Papanastasiou et al, 2013). Η τεχνική αυτή έχει αντίστοιχα τους δικούς της περιορισμούς όπως ότι πάσχει από το ενδεχόμενο επιμόλυνσης μεταξύ των διαφορετικών κυτταρικών διαμερισμάτων (Chandramouli et al, 2009).

Για συγκεκριμένες κατηγορίες πρωτεϊνών η τελική τους θέση μέσα στο κύτταρο μπορεί να προβλεφθεί με τη χρήση βιοπληροφορικών εργαλείων (Bendtsen et al, 2005b; Gardy et al, 2003;

Goldberg et al, 2014; Imai et al, 2008; Juncker et al, 2003; Kall et al, 2007; Krogh et al, 2001; Paramasivam et al, 2011; Petersen et al, 2011). Συγκεκριμένα στοιχεία όπως το σηματοδοτικό πεπτιδίο (ΣΠ) αποτελούν ασφαλή αναγνωριστικά των εκκρινόμενων πρωτεϊνών (Bagos et al, 2010; Bendtsen et al, 2005b; Juncker et al, 2003; Kall et al, 2007; Petersen et al, 2011) που χρησιμοποιούν το κύριο εκκριτικό σύστημα Sec (Chatzi et al, 2013) ή το έλασσον εκκριτικό σύστημα TAT (Patel et al, 2014). Εντούτοις τέτοια στοιχεία μπορεί να μην είναι εύκολα αναγνωρίσιμα (Bendtsen et al, 2005a) ή να απουσιάζουν από πρωτεΐνες για τις οποίες έχει διαπιστωθεί ότι εκκρίνονται (Bilous et al, 1988). Στοιχεία δευτεροταγούς δομής συνιστούν επίσης αναγνωριστικά στοιχεία κυτταρική θέσης όπως οι διαμεμβρανικές α-έλικες (Kall et al, 2004; Kall et al, 2007) και τα β-βαρέλια (Koronakis et al, 2000; Wimley, 2003).

Η ολοκληρωμένη ταξινόμηση των πρωτεϊνών δεν είναι διαθέσιμη για το βακτήριο *E.coli*. Οι υπάρχουσες βάσεις δεδομένων όπως το UniProt (Dimmer et al, 2012) αλλά και το EchoLOCATION (Horler et al, 2009) ταξινομούν είτε μέρος του πρωτεϊνώματος είτε το μεγαλύτερο μέρος της ταξινόμησης βασίζεται σε προβλέψεις.

Στις ενότητες που ακολουθούν θα περιγράψουμε την συνδυαστική ανάλυση που ακολουθήσαμε έχοντας ως στόχο να προσδιορίσουμε τον υποκυτταρικό εντοπισμό του συνόλου του πρωτεϊνώματος του βακτηρίου *E.coli*. Η ανάλυση αυτή συνδυάζει την χρήση βιοπληροφορικών εργαλείων, υπάρχουσες ταξινομήσεις σε βάσεις δεδομένων, συλλογή και σύγκριση γονιδιακών, πρωτεοματικών αλλά και βιοχημικών δεδομένων καθώς επίσης και εκτεταμένη βιβλιογραφική αναζήτηση. Η ανάλυση αυτή οδήγησε στην εμπεριστατωμένη αναθεώρηση της υποκυτταρικής ταξινόμησης του πρωτεϊνώματος του *E.coli* η οποία βασίστηκε στη χρήση μοντέρνων εργαλείων πρόβλεψης αλλά κυρίως στην συλλογή πειραματικών δεδομένων και στην εξαντλητική βιβλιογραφική έρευνα.

## 2.1 Υποκυτταρικά διαμερίσματα βακτηρίου *E.coli*

Όλα τα κύτταρα αποτελούνται από διακριτές περιοχές που οριοθετούνται μέσω βιολογικών μεμβρανών. Στα βακτήρια το κυτταρόπλασμα περικλείεται από μονή (Gram<sup>+</sup>) ή διπλή (Gram<sup>-</sup>) διπλοστοιβάδα λιπιδίων. Στα Gram<sup>-</sup> βακτήρια το κυτταρόπλασμα περικλείεται από μία πολύεπίπεδη δομή που ονομάζεται κυτταρικός φάκελος (ΚΦ).

---

Ο ΚΦ αποτελείται από την πλασματική/εσωτερική μεμβράνη (ΠΜ) και μια δεύτερη εξωτερική διπλοστοιβάδα λιπιδίων την εξωτερική μεμβράνη (ΕΜ). Πάνω στην ΕΜ βρίσκονται προσδεμένα μόρια λιποσακχαρίτη (Silhavy et al, 2010). Ο υδάτινος όγκος ανάμεσα στις δύο μεμβράνες ονομάζεται περίπλασμα και περιέχει πρωτεΐνες αλλά και ένα δίκτυο από μόρια πεπτιδογλυκάνης.

## 2.2 Βασικό πρωτεΐνωμα του *E.coli*

Τα διαφορετικά στελέχη του *E.coli* επιβιώνουν σε ποικίλες περιβαλλοντολογικές συνθήκες. Η ευελιξία αυτή προκύπτει από ένα σύνολο χαρακτηριστικών που έχουν συσσωρεύσει στο γονιδίωμα τους, στην πορεία της εξέλιξης. Για παράδειγμα, γονίδια τα οποία τους προσδίδουν ανθεκτικότητα σε αφιλόξενα περιβάλλοντα, στοιχεία που τους βοηθάνε να ανασυνδυάσουν συγκεκριμένες γονιδιακές περιοχές (transposases) καθώς επίσης και γονίδια τα οποία δεν εκφράζονται (silent genes) (Choudhury et al, 2012). Προγενέστερο βήμα της ανάλυσης της υποκυτταρική ταξινόμησης στο *E.coli* ήταν να ορίσουμε το πρωτεΐνωμα στο οποίο δεν περιλαμβάνονται τέτοιου είδους γονίδια, το οποίο ονομάσαμε *βασικό πρωτεΐνωμα*. Το *βασικό πρωτεΐνωμα* δεν περιέχει πρωτεΐνες που πιθανόν δεν συντίθενται και/ή κωδικοποιούνται σε περιοχές στο γονιδίωμα όπου είναι πλούσιες σε: 1) εισαγωγές (genomic insertions), 2) γονίδια προφάγων (defective prophages), 3) γονίδια μεταθέτες (transposons), 4) ψευδογονίδια ή 5) κινητά στοιχεία (mobile elements) (Πίνακας 2.1, Πίνακας 2.2). Προσδιορίσαμε ότι το *βασικό πρωτεΐνωμα* αποτελείται από 3897 πρωτεΐνες (Πίνακας 2.1) εκ των οποίων 3849 εκφράζονται σε επίπεδο mRNA (Patten et al, 2004b; Taniguchi et al, 2010; Wang et al, 2005a; Yoon et al, 2012) και 3178 σε επίπεδο πρωτεΐνης (Iwasaki et al, 2010; Pan et al, 2010).

### Πίνακας 2.1 – Κινητά στοιχεία των στελεχών *E.coli* K-12 και BL21-DE3

Κατηγορίες κινητών στοιχείων: α. **γονίδια προφάγων** (prophage integrases): πρωτεΐνες προφάγων που καταλύουν την εισαγωγή και ενσωμάτωση του DNA του προφάγου στο γονιδίωμα του βακτηρίου β. **Εισαγωγικά στοιχεία** (insertion elements / sequences): κινητά στοιχεία στο DNA που συνήθως κωδικοποιούν πρωτεΐνες που χρειάζονται για γονιδιακή μετάθεση για παράδειγμα μεταθέτες, γ. **μεταθέτες** (transposases): πρωτεΐνες που είναι απαραίτητες για χρωμοσωμικές μεταθέσεις, δ. **στοιχεία Rhs**: πρωτεΐνες με υψηλό περιεχόμενο σε αμινοξικές επαναλήψεις, οι επαναλήψεις αυτές συχνά έχουν σχετιστεί με ρόλους πρόσδεσης.

\* Το κοινό πρωτεϊνωμα ανάμεσα στα στελέχη *E.coli* K-12 και BL21-DE3 (δες ενότητες 2.11 και 6.1.4). Οι πρωτεΐνες YBEQ\_ECOLI και FDHF\_ECOLI χωρίστηκαν στις C6EK49/C6EK48 και C6ED45/C6ED46 αντίστοιχα στο *E.coli* BL21-DE3 στέλεχος; \*\* Κάθε πρωτεΐνη εδώ μετράει μία φορά παρόλο που μπορεί να έχει παραπάνω από μία ομόλογες πρωτεΐνες; <sup>1</sup> Το πρωτεϊνωμα αναφοράς του στελέχους *E.coli* K-12 (UniProt: Αύγουστος 2013); <sup>2</sup> Το πρωτεϊνωμα αναφοράς του στελέχους *E.coli* BL21-DE3 (UniProt: Αύγουστος 2013); <sup>4</sup> Αριθμός πιθανών ψευδογονιδίων, όπως προέκυψε από τον συνδυασμό της υπάρχουσας πληροφορίας στις βάσεων δεδομένων UniProt, EcoGene και της γονιδιακής ανάλυσης των Ochman 2006 et al (Πίνακας 2.2). Εδώ αναφέρεται ο αριθμός των ψευδογονιδίων και κινητών στοιχείων.

	<i>E.coli</i> K-12 (MG1655)	Κοινό πρωτεϊνωμα των στελεχών K-12 και BL21-DE3 **	<i>E.coli</i> BL21-DE3
<b>Συνολικό πρωτεϊνωμα</b>	4303 <sup>1</sup>		4842 <sup>2</sup>
<b>Ψευδογονίδια</b> (πρόβλεψη)	147	106 / 100	100
<b>Κινητά στοιχεία</b> (γονίδια προφάγων / εισαγωγικά στοιχεία / στοιχεία Rhs / μεταθέτες)	259	166 / 149	190
<b>Συνολικά Ψευδογονίδια /Κινητά στοιχεία</b>	406	272 / 249	290
<b>Βασικό πρωτεϊνωμα</b>	3897		4552
Κοινό πρωτεϊνωμα * (ποσοστό ταυτοποίησης >40%)	4037		4483
Μοναδικές πρωτεΐνες	266		359



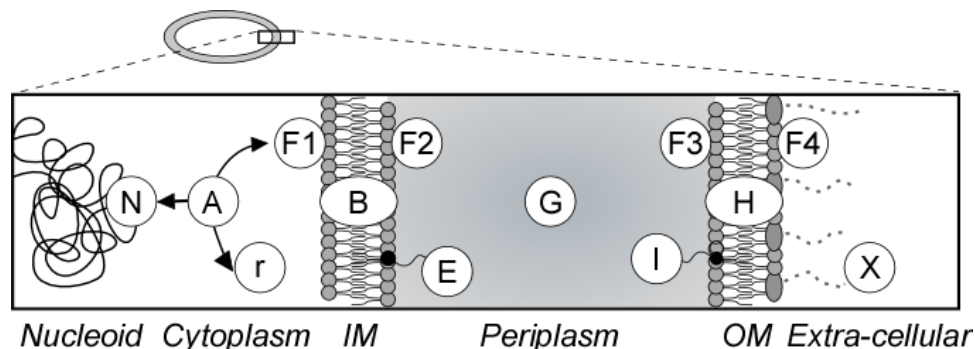
**Πίνακας 2.2 - Ανάλυση των ψευδογονιδίων στο στέλεχος *E.coli* K-12**

<sup>ο</sup> Γονίδια απαραίτητα για την αερόβια ανάπτυξη του βακτηρίου σε πλούσιο θρεπτικό μέσο; <sup>1</sup> Γονίδια για τα οποία έχει επιβεβαιωθεί πειραματικά η κυτταρική τους λειτουργία; <sup>2</sup> Κοινές πρωτεΐνες ανάμεσα σε 43 *E.coli* στελέχη. (Πίνακας 6.2); <sup>3</sup> Πρωτεΐνες που έχουν ανιχνευθεί πειραματικά από πρωτεομική ανάλυση (Iwasaki et al, 2010), \* Ταξινομημένες προερχόμενες από την βάση δεδομένων Uniprot (Αύγουστος 2013)

	Uniprot	EcoGene	Κοινός μεταξύ Uniprot/EcoGene	τόπος	Ψευδογονίδια (πρόβλεψη) (Ochman et al, 2006)	Συνολικός αριθμός πιθανών ψευδογονιδίων
<b>Σύνολο</b>	<b>126</b>	<b>20</b>		<b>11</b>	<b>126</b>	<b>207</b>
Έχει βρεθεί να εκφράζεται σε επίπεδο πρωτεΐνης (Uniprot) *	-	4		-	25	25
Έχει βρεθεί να εκφράζεται σε επίπεδο mRNA (Uniprot) *	1	1		1	-	1
Προέκυψε από ομολογία με άλλα στελέχη (Uniprot) *	1	1		-	23	24
Έχει προβλεφθεί (Uniprot) *	-	3		-	20	20
<b>Αβέβαιο (Uniprot) *</b>	<b>124</b>	<b>11</b>		<b>10</b>	<b>58</b>	<b>137</b>
Απαραίτητο για την βιωσιμότητα του κυττάρου (Ochman, 2006) <sup>ο</sup>	-	-		-	12	12
Η λειτουργία του είναι επιβεβαιωμένη πειραματικά (Ochman, 2006) <sup>1</sup>	-	2		-	21	21
<b>Κοινό πρωτεΐνωμα <sup>2</sup></b>	<b>8</b>	<b>-</b>		<b>-</b>	<b>31</b>	<b>36</b>
Ανιχνεύθηκαν σε πρωτεομική ανάλυση (Iwasaki, 2010) <sup>3</sup>	3	2		-	37	37

### 2.3 Πλήρης ταξινόμηση των πρωτεϊνών του βακτηρίου *E.coli* σε υποκυτταρικά διαμερίσματα

Τα υποκυτταρικά διαμερίσματα του *E.coli* κωδικοποιήθηκαν σε 13 κατηγορίες (10 που αφορούν τον ΚΦ και 3 το κυτταρόπλασμα (Εικόνα 2.1), ξεκινώντας και επεκτείνοντας τις κατηγορίες του EchoLOCATION (Horler et al, 2009). Κάθε κατηγορία σχετίζεται και με ένα όρο οντολογίας (Πίνακας 2.3; GO term; (GOConsortium, 2012)). Πιο αναλυτικά οι κατηγορίες είναι: ριβοσωμικές πρωτεΐνες (r) οι οποίες μαζί με μόρια rRNA συγκροτούν τις υπομονάδες του ριβοσώματος; πρωτεΐνες του πυρηνοειδούς (N), σε αυτές ανήκουν πρωτεΐνες που προσδένονται πάνω στο DNA/RNA όπως DNA ελικάσες, πολυμεράσες, παράγοντες σίγμα (sigma factors), ένζυμα επιδιόρθωσης του DNA (repair enzymes) και μεταγραφικοί παράγοντες (transcription factors). Πρωτεΐνες για τις οποίες υπάρχει πειραματική ένδειξη ότι βρίσκονται στο κυτταρόπλασμα ή για τις οποίες δεν βρήκαμε κάποια πληροφορία για το που βρίσκονται ταξινομήθηκαν ως κυτταροπλασματικές (A). Σε γενικές γραμμές όλες οι πρωτεΐνες του πυρηνοειδούς και οι περιφερικές πρωτεΐνες θεωρούμε ότι μπορούν να βρεθούν και σε κυτταροπλασματική κατάσταση.



Εικόνα 2.1 – Κατηγορίες υποκυτταρικού εντοπισμού.

Το κύτταρο του βακτηρίου *E.coli* αποτελείται από το κυτταρόπλασμα (cytoplasm) το οποίο περικλείεται από την ΠΜ (IM) και την ΕΜ (OM). Οι πρωτεΐνες μπορούν να ταξινομηθούν σε 13 κατηγορίες N: πρωτεΐνες του πυρηνοειδούς, r: ριβοσωμικές πρωτεΐνες, A: κυτταροπλασματικές, F1: περιφερικές πρωτεΐνες της ΠΜ από την πλευρά του κυτταροπλάσματος, B: διαμεμβρανικές πρωτεΐνες της ΠΜ, F2: περιφερικές πρωτεΐνες της ΠΜ από την πλευρά του περιπλάσματος, E: λιποπρωτεΐνες της ΠΜ, G: περιπλασματικές πρωτεΐνες, I: λιποπρωτεΐνες της ΕΜ, F3: περιφερικές πρωτεΐνες της ΕΜ από την πλευρά του περιπλάσματος, H: διαμεμβρανικές πρωτεΐνες της ΕΜ, F4: περιφερικές πρωτεΐνες της ΕΜ από την πλευρά του εξωκυττάρου χώρου, X: εξωκυττάρια πρωτεΐνες.

**Πίνακας 2.3 - Αντιστοιχία ανάμεσα στην ονοματολογία του STEPdb και των όρων οντολογίας (gene ontology)**

STEPdb (κωδικός ενός γράμματος)	STEPdb (Πλήρες όνομα)	Κωδικός όρου οντολογίας (Gene Ontology ID)	Περιγραφή όρου οντολογίας
N	Nucleoid	<a href="#">9295</a>	Nucleoid
r	Ribosome	<a href="#">5840</a>	Ribosome
A	Cytoplasmic	<a href="#">44444</a>	Cytoplasmic part
F1	Peripheral inner membrane protein facing the cytoplasm	<a href="#">31234</a>	Extrinsic to internal side of plasma membrane
B	Integral Inner Membrane	<a href="#">5887</a>	Integral to plasma membrane
F2	Peripheral inner membrane protein facing the periplasm	<a href="#">31232</a>	Extrinsic to external side of plasma membrane
E	Inner Membrane Lipoprotein	<a href="#">31233</a>	Intrinsic to external side of plasma membrane
G	Periplasmic	<a href="#">42597</a>	Periplasmic space
I	Outer Membrane Lipoprotein	<a href="#">31246</a>	Intrinsic to internal side of cell outer membrane
F3	Peripheral outer membrane protein facing the periplasm	<a href="#">31245</a>	Extrinsic to internal side of cell outer membrane
H	Outer Membrane b-barrel protein	<a href="#">45203</a>	Integral to cell outer membrane
F4	Peripheral outer membrane protein facing the extracellular space	<a href="#">31242</a>	Extrinsic to external side of cell outer membrane
X	Extracellular	<a href="#">44420</a>	Extracellular matrix part
		<a href="#">19861</a>	Flagellum
		<a href="#">9289</a>	Pilus

## 2.4 Ασυνέπεια μεταξύ παλαιότερων ταξινομήσεων

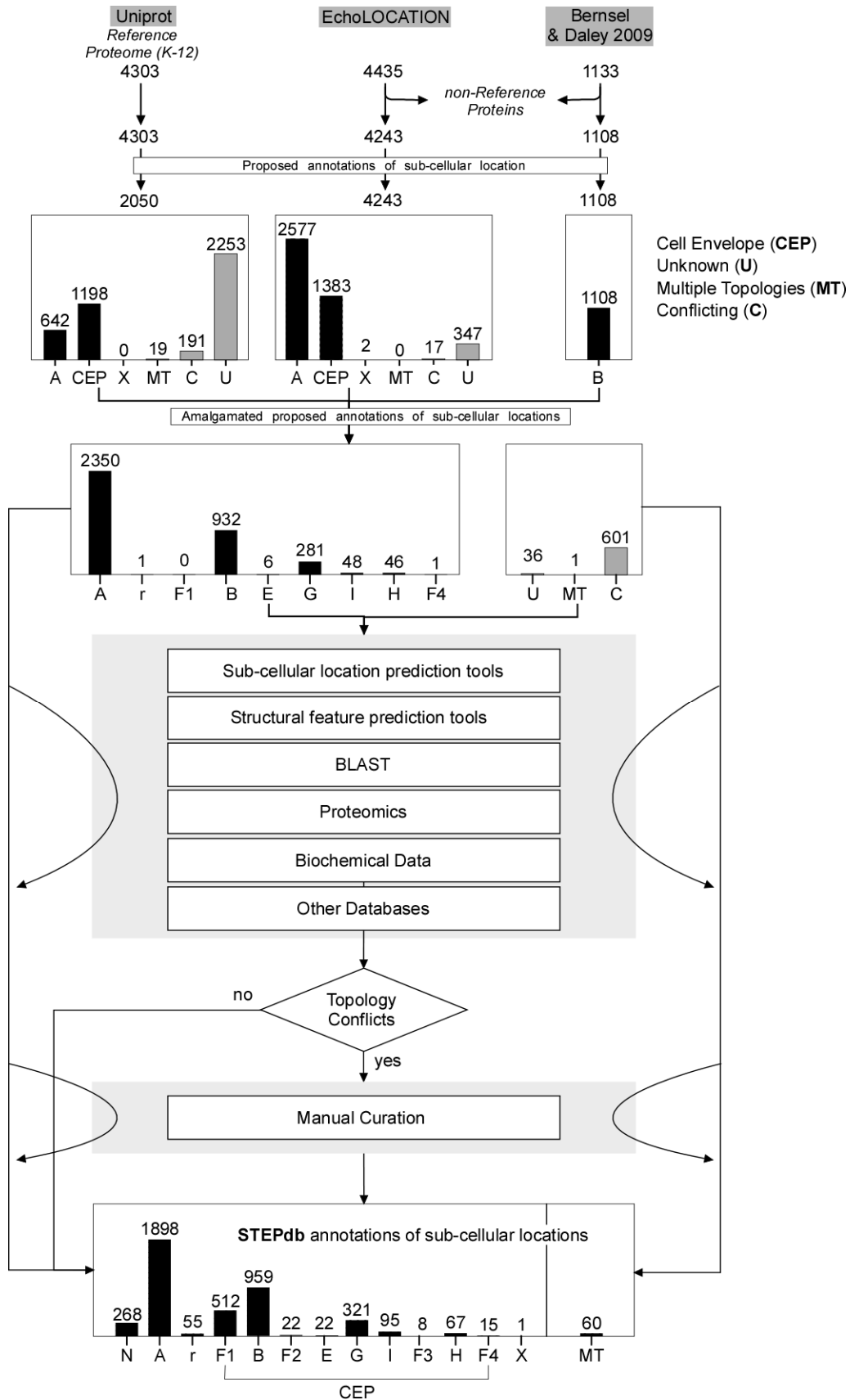
Αρχικά συγκρίναμε τις υποκυτταρικές ταξινομήσεις τριών πηγών. Η βάση δεδομένων UniProt ταξινομεί 48% του πρωτεϊνώματος αναφοράς *E.coli* K-12 (Πίνακας 2.6). Η βάση δεδομένων EchoLOCATION οργανώνει 4345 πρωτεΐνες σε 11 υποκυτταρικές κατηγορίες. Η θεωρητική ανάλυση των Bernsel και Daley (2009), χαρακτηρίζει 1133 μεμβρανικές πρωτεΐνες. Αντιστοιχίσαμε 4243 από τις 4345 πιθανά κωδικοποιούσες αλληλουχίες στο EchoLOCATION ενώ 1108 από 1133 προτεινόμενες μεμβρανικές πρωτεΐνες (Bernsel et al, 2009) στο πρωτεϊνώμα

αναφοράς του *E.coli* K-12 (Πίνακας 2.6; Εικόνα 2.2). Οι πρωτεΐνες που δεν αντιστοιχήθηκαν στο πρωτεϊνώμα αναφοράς του *E.coli* K-12 δεν αναλύθηκαν περαιτέρω. Πιο αναλυτικά σε αυτές συμπεριλαμβάνονται άγνωστες κωδικοποιούσες περιοχές πιθανά ψευδογονίδια, διπλές εγγραφές στο EchoLOCATION, πρωτεΐνες που έχουν διαγραφεί από το πρωτεϊνώμα αναφοράς ή ανήκουν σε άλλα στελέχη *E.coli* σύμφωνα με το Uniprot.

Η αντιπαράθεση των υποκυτταρικών ταξινομήσεων των τριών πηγών (που αντιστοιχήθηκαν στο πρωτεϊνώμα αναφοράς του *E.coli* K-12) αποκάλυψε ότι υπήρχαν πρωτεΐνες με: α) ταύτιση προτεινόμενων ταξινομήσεων (“Matching”) β) διαφορές στις προτεινόμενες υποκυτταρικές ταξινομήσεις (“Conflicting”) και γ) απουσία υποκυτταρικής ταξινόμησης (“Unknown”) από οποιαδήποτε πηγή (Εικόνα 2.2 και Εικόνα 2.3). Για περίπου το ~14% του πρωτεϊνώματος αναφοράς (~15% του βασικού πρωτεϊνώματος) οι υπάρχουσες ταξινομήσεις ήταν αντικρουόμενες (Πίνακας 2.5).

Για μέρος των πρωτεϊνών η ταξινόμηση των πηγών βασίζεται σε πειραματικά δεδομένα (398 στο Uniprot; 506 στο EchoLOCATION; μόνο 105 πρωτεΐνες κοινές; Εικόνα 2.4A). Για τις υπόλοιπες πρωτεΐνες η ταξινόμηση είναι θεωρητική είτε προέρχεται από πρόβλεψη βιοπληροφορικών εργαλείων. Το Uniprot ορίζει τρία επίπεδα θεωρητικής ταξινόμησης για τις πρωτεΐνες που έχουν που δεν επαληθευτεί πειραματικά: «θεωρητική» (potential) (η ταξινόμηση είναι προϊόν αλγόριθμου πρόβλεψης), «πιθανή» (probable) (υπάρχει έστω κάποια πειραματική ένδειξη), «Βάση ομοιότητας» (by similarity) (ενδείξεις υπάρχουν για ομόλογες πρωτεΐνες σε άλλα βακτηριακά στελέχη). Το EchoLOCATION υιοθετεί δύο επίπεδα απόδειξης της ταξινόμησης το «θεωρητικό» για όσες έχουν προβλεφθεί από βιοπληροφορικά μοντέλα και το «πειραματικό».

Για την εμπειριστατωμένη ταξινόμηση των άγνωστων πρωτεϊνών αλλά και για τη διασάφηση των αντιφατικών ταξινομήσεων αποφασίσαμε να συγκεντρώσουμε πειραματικά δεδομένα και να επανεξετάσουμε *ab initio* τις υποκυτταρικές θέσεις των πρωτεϊνών.



---

**Εικόνα 2.2 – Εμπειριστικώς ταξινομημένη υποκυτταρική ταξινόμηση του συνολικού πρωτεϊνώματος του *E.coli*.**

Για την ανάλυση χρησιμοποιήθηκε το πρωτεϊνωμα αναφοράς (reference proteome) του *E.coli* K-12 διαθέσιμο από το Uniprot (Αύγουστος 2013;(Dimmer et al, 2012)). Υπάρχουσα ταξινόμηση για 2050 πρωτεΐνες συλλέχθηκε από το Uniprot (48% του πρωτεϊνώματος; **Πίνακας 2.6**). Αντίστοιχη ταξινόμηση διαθέσιμη για 3957 (από 4345) στο EchoLOCATION (Horler et al, 2009) χρησιμοποιήθηκε. Ο προσδιορισμός των διαμεμβρανικών πρωτεϊνών της ΠΜ βασίστηκε σε μια πρωτεομική ανάλυση που αναφέρει 1133 διαμεμβρανικές πρωτεΐνες της ΠΜ (Bernsel et al, 2009).

Οι ορολογίες των δύο βάσεων δεδομένων αντιστοιχίστηκαν στην κατηγοριοποίηση των 13 υποκυτταρικών θέσεων (**Εικόνα 2.1, Εικόνα 2.2**). Ο συνδυασμός των τριών αυτών πηγών συνεισφέρει πιθανή ταξινόμηση για 4267 πρωτεΐνες αφήνοντας 36 πρωτεΐνες με άγνωστο υποκυτταρικό εντοπισμό και 601 πρωτεΐνες με αμφιλεγόμενες ταξινομήσεις λόγω ασυμφωνίας μεταξύ των πηγών (Table I). Για την ταξινόμηση των άγνωστων πρωτεϊνών και την διασάφηση των αντιθέσεων στις υπάρχουσες ταξινομήσεις χρησιμοποιήσαμε βιοπληροφορικά εργαλεία τα οποία προβλέπουν υποκυτταρικό εντοπισμό ή άλλα δομικά χαρακτηριστικά καθώς και στοίχιση αλληλουχιών. (**Εικόνα 2.2**). Τα βασικά βιοπληροφορικά εργαλεία που χρησιμοποιήσαμε ήταν: SignalP, TatP, LipoP, Phobius (Bendtsen et al, 2005b; Juncker et al, 2003; Kall et al, 2007; Petersen et al, 2011) τα οποία αναγνωρίζουν σηματοδοτικές αλληλουχίες, το PSORT-B (Gardy et al, 2003) για την πρόβλεψη υποκυτταρικού εντοπισμού και τα TMHMM και Phobius (Kall et al, 2004; Kall et al, 2007) για την πρόβλεψη διαμεμβρανικών περιοχών. Ένα σύνολο από συμπληρωματικά βιοπληροφορικά εργαλεία (Prediction Tools 2) αποτέλεσαν τα: ProtScale για υπολογισμό υδροφοβικότητας των πρωτεϊνών (Wilkins et al, 1999), SOSUI, ClubSub και LocTree3 για πρόβλεψη υποκυτταρικού εντοπισμού (Goldberg et al, 2014; Imai et al, 2008; Paramasivam et al, 2011), AmphipaSeek (Sapay et al, 2006) για την πρόβλεψη αμφίφιλων α-ελικών ενώ χρησιμοποιήσαμε BLAST (Camacho et al, 2009) για τον προσδιορισμό των πιθανών autotransporters. Η αντιπαράθεση της ταξινόμησης των βάσεων δεδομένων και των προβλέψεων οδήγησε σε συγκρουόμενες πιθανές ταξινομήσεις (**Πίνακας 2.1**). Για να διασαφηνιστούν αυτές οι διαφορές προχωρήσαμε σε εκτεταμένη βιβλιογραφική αναζήτηση αλλά και στην συλλογή πειραματικών δεδομένων από πρωτεομικές, γονιδιωματικές και βιοχημικές αναλύσεις (**Πίνακας 6.1**).

---

## 2.5 Συνδυαστική ανάλυση για την εμπειριστατωμένη *de novo* ταξινόμηση του πρωτεϊνώματος του *E.coli* K-12.

Προς την πλήρη επανεξέταση και ταξινόμηση του πρωτεϊνώματος *E.coli* ακολουθήσαμε μια συνδυαστική ανάλυση στην οποία συμπεριλάβαμε βιοπληροφορικά εργαλεία (BE) για την πρόβλεψη υποκυτταρικού εντοπισμού ή δομικών στοιχείων, αναζητήσεις ομοιότητας και πολλαπλής στοίχισης ακολουθιών (sequence similarity, BLAST), πρωτεϊνωματικά και γονιδιωματικά δεδομένα, βάσεις δεδομένων και ως βασικό μέσο την εκτεταμένη βιβλιογραφική έρευνα (Εικόνα 2.2). Με ένα δέντρο απόφασης συνοψίζουμε την διαδικασία εκτίμησης και ταξινόμησης της πληροφορίας από τις διάφορες πηγές βασιζόμενοι στο βαθμό αξιοπιστίας που συνοδεύει αυτά τα δεδομένα (Εικόνα 2.3).

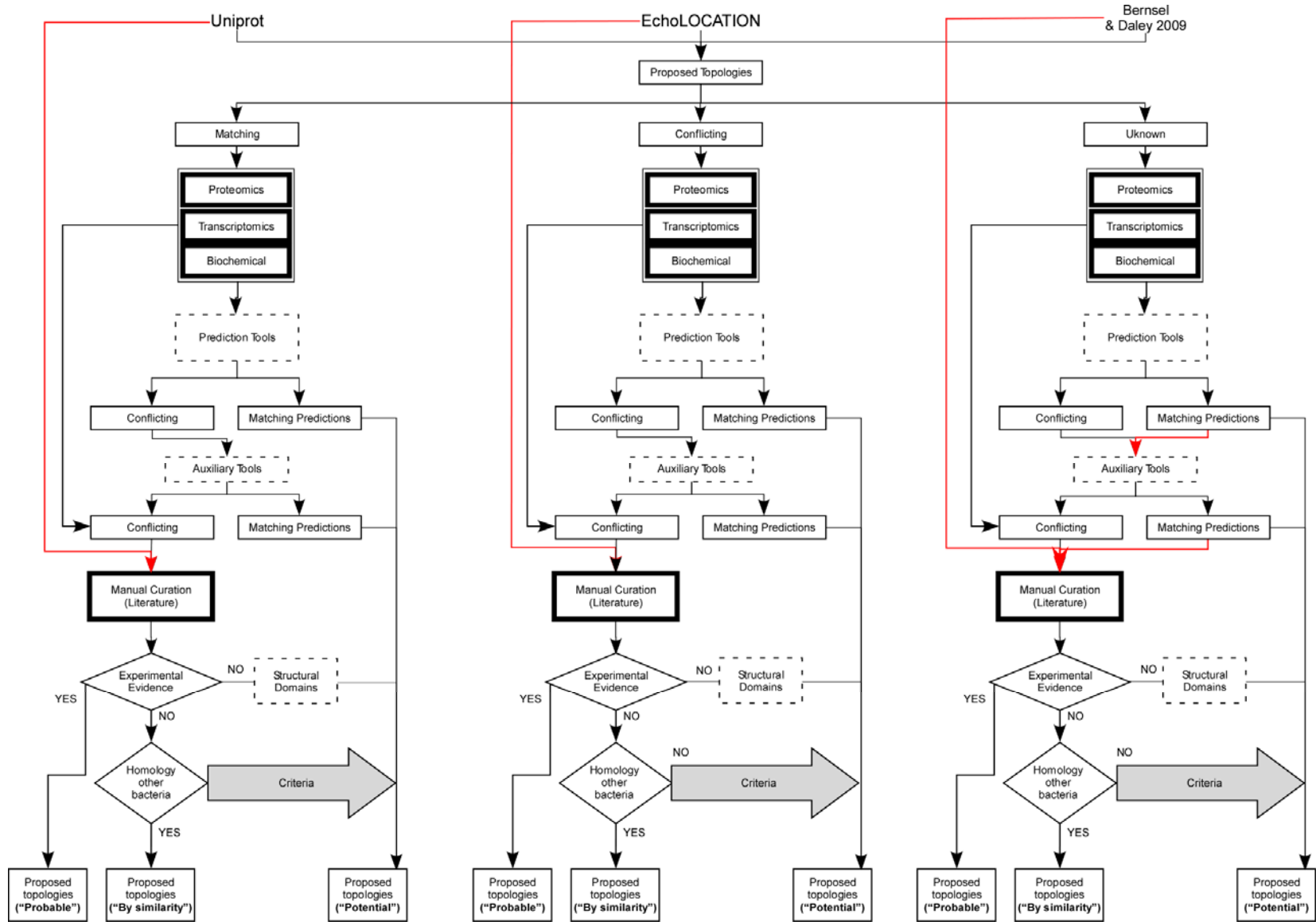
Σε ορισμένες ήδη υπάρχουσες ταξινομήσεις οι πρωτεΐνες ήταν καταχωρημένες ως πειραματικά επαληθευμένες. Στην πλειοψηφία αυτών των περιπτώσεων υιοθετήσαμε την ήδη υπάρχουσα ταξινόμηση, παρόλο που σε ελάχιστες περιπτώσεις η προτεινόμενη τοποθεσία αναθεωρήθηκε λόγω περισσότερων πειραματικών τεκμηρίων. Στις ενότητες που ακολουθούν περιγράφεται η διαδικασία ανάλυσης που ακολουθήθηκε καθώς επίσης παραθέτονται αντίστοιχα παραδείγματα. Καταχωρήσαμε την πληροφορία που συλλέξαμε (δημοσιεύσεις, πειραματικές μετρήσεις, προβλέψεις) αλλά και την τελική υποκυτταρική ταξινόμηση του πρωτεϊνώματος του *E.coli* στην οποία καταλήξαμε, σε μια βάση δεδομένων την οποία ονομάσαμε STEPdb (δες ενότητα 2.13).

## 2.6 Ταξινόμηση «άγνωστων» πρωτεϊνών

Για την ταξινόμηση των 36 πρωτεϊνών με άγνωστο υποκυτταρικό εντοπισμό (Πίνακας 2.6), σε πρώτο στάδιο χρησιμοποιήσαμε μοντέλα πρόβλεψης ενώ σαν δεύτερο εργαλείο την βιβλιογραφική αναζήτηση. Η αναζήτηση μας οδήγησε στο να ταξινομήσουμε τις: MntS (κυταροπλασματική), OmpP (πρωτεΐνη της εξωτερικής μεμβράνης) and SgrT (περιφερική πρωτεΐνη; F1) ως πειραματικά αποδεδειγμένες. Για την πρωτεΐνη YshB υπάρχει πρόβλεψη για μία πιθανή αμφίφιλη έλικα (θέσεις 1-10)(Sapay et al, 2006) και ως εκ τούτου την ταξινομήσαμε ως περιφερική πρωτεΐνη (F1). Οι υπόλοιπες πρωτεΐνες ταξινομήθηκαν ως «θεωρητικές» (potential) με βάση τις προβλέψεις των BE (οι περισσότερες από αυτές ως κυταροπλασματικές).







---

**Εικόνα 2.3 – Διάγραμμα ροής της διαδικασίας υποκυτταρικής ταξινόμησης του πρωτεϊνώματος *E.coli*.**

Σε αυτό το δέντρο απόφασης συνοψίζουμε τα βήματα που ακολουθήσαμε προς τον πλήρη χαρακτηρισμό του βακτηρίου *E.coli* ως προς την υποκυτταρική τοποθεσία. Οι βασικές πηγές που χρησιμοποιήσαμε (UniProt, EchoLOCATION, θεωρητικό ΜΠ (3)) συνδυάστηκαν και οι διαθέσιμες υποκυτταρικές ταξινομήσεις σε αυτά, συγκρίθηκαν. Οι πρωτεΐνες κατηγοριοποιήθηκαν σε τρεις υποπεριπτώσεις α) αυτές με ταυτόσημες ταξινομήσεις ανάμεσα στις τρεις πηγές ("Matching"), β) αυτές με αντικρουόμενες προτεινόμενες ταξινομήσεις ("Conflicting"), και γ) στις άγνωστες ("Unknown"). Για όσες πρωτεΐνες συνοδεύονταν με πειραματικές ενδείξεις από τις βάσεις δεδομένων UniProt και EchoLOCATION υιοθετήσαμε τις προτεινόμενες υποκυτταρικές θέσεις. Για τις περιπτώσεις β) και γ) συνδυάσαμε βιοπληροφορικά μοντέλα πρόβλεψης υποκυτταρικού εντοπισμού, πειραματικά δεδομένα (βιοχημικά, πρωτεϊνωματικά και γονιδιωματικά) αλλά και εκτεταμένη βιβλιογραφική αναζήτηση.

---

## 2.7 Διαλεύκανση αντιφάσεων στις υπάρχουσες ταξινομήσεις

Για 601 πρωτεΐνες υπήρξαν αντικρουόμενες ταξινομήσεις («Conflicting») (Εικόνα 2.3, Πίνακας 2.6). Για να επιλύσουμε τις διαφορές ακολουθήσαμε μια σειρά από βήματα:

**1. Πειράματα μεγάλης κλίμακας:** Αρχικά αναλύσαμε δεδομένα από πειράματα ευρείας κλίμακας: πρωτεϊνωματικά, βιοχημικά (Daley et al, 2005; Gonnet et al, 2004), γονιδιωματικά (Ishihama, 2012) και πειράματα μικροσκοπίας (Πίνακας 6.1).

Πρωτεΐνες για τις οποίες βρέθηκαν αποδεικτικά στοιχεία κατηγοριοποιήθηκαν ως πειραματικές (“experimental”) στο αντίστοιχο υποκυτταρικό διαμέρισμα, εκτός εάν πιο εξειδικευμένες και στοχευόμενες βιοχημικές αναλύσεις τις έχουν ανιχνεύσει σε άλλη τοποθεσία (δες βήμα 4). Υπήρξαν περιπτώσεις όπου μια πρωτεΐνη βρέθηκε σε δύο διαφορετικές θέσεις από δύο διαφορετικές πρωτεϊνωματικές αναλύσεις. Τέτοια παραδείγματα υπήρξαν οι: BtuB η οποία ανιχνεύθηκε στην EM και στο κυτταρόπλασμα, η Fis που απομονώθηκε σε μεμβρανικά κλάσματα αλλά και σε νουκλεοειδή (Πίνακας 6.1). Απουσία επιπλέον τεκμηρίων οι πρωτεΐνες αυτές σημειώθηκαν ως πιθανές να ανήκουν σε ένα από αυτά τα διαμερίσματα.

**2. Βιοπληροφορικά εργαλεία:** Όσες πρωτεΐνες παρέμειναν με αντικρουόμενες ταξινομήσεις εξετάστηκαν με την βοήθεια μοντέρνων BE. Στην περίπτωση συμφωνίας των προβλέψεων (π.χ. SignalP, LipoP και Phobius προβλέπουν την ύπαρξη ΣΠ; TMHMM και Phobius προβλέπουν ΔΠ) οι πρωτεΐνες κατηγοριοποιήθηκαν ως «θεωρητικές» (“potential”). Στην αντίθετη περίπτωση όπου ακόμα και τα βασικά εργαλεία απέτυχαν να συμφωνήσουν στις προβλεπόμενες ταξινομήσεις τότε καταφύγαμε στην χρήση *βοηθητικών βιοπληροφορικών εργαλείων* (δες 6.1.2 και Εικόνα 2.3).

**3. Δομικά στοιχεία τα οποία σχετίζονται με υποκυτταρικό εντοπισμό:** Συγκεκριμένα δομικά χαρακτηριστικά των πρωτεϊνών είναι ενδεικτικά της λειτουργίας τους και κατά συνέπεια της θέσης τους μέσα στο κύτταρο. Αυτά περιλαμβάνουν διαμεμβρανικές έλικες, βήτα βαρέλια αμφίφιλες α-έλικες και τριτοταγή δομικά χαρακτηριστικά όπως η δομική περιοχή «autotransporter».

Μερικές περιφερικές πρωτεΐνες της ΠΜ αλληλεπιδρούν με αυτή μέσω αμφίφιλων ελίκων (amphipathic helices) (King et al, 1999; Parlitz et al, 2007; Phoenix et al, 1990; Shiomi et al, 2008a; Sung et al, 2009; Walz et al, 2002). Η παρούσα ανάλυση συγκεντρώνει εννέα από αυτές: Dhna

(Villegas et al, 2011), PbpB (Sung et al, 2009), FtsA (Shiomi et al, 2008a) , MinD (King et al, 1999), GlpD (Walz et al, 2002), FtsY (Parlitz et al, 2007), Rne (Murashko et al, 2012), Rnb (Lu et al, 2013) και MinE (Shih et al, 2011). Πρωτεΐνες με αντίστοιχα χαρακτηριστικά βρίσκονται αγκυροβολημένες και στην εξωτερική στοιβάδα της ΠΜ, στο περίπλασμα (π.χ. DacA (Phoenix et al, 1990)).

Το δομικό στοιχείο «autotransporter», βρίσκεται στο καρβοξυτελικό άκρο (C-terminus) των autotransporters (AT) (Henderson et al, 2000). Στην βάση δεδομένων STEPdb συμπεριλαμβάνονται 10 AT πρωτεΐνες οι οποίες είναι ομόλογες με την καλά χαρακτηρισμένη πρωτεΐνη Ag43 και περιέχουν επίσης το δομικό στοιχείο «at-1» (InterPro family) (Hunter et al, 2009) ενώ τρεις από αυτές (YcgI, YcgV and YdeU) δεν έχουν πρόβλεψη για ΣΠ.

Δομικά στοιχεία πρόσδεσης σε μόρια πεπτιδογλυκάνης (peptidoglycan(PG)-binding domains) μπορούν επίσης να αποτελέσουν δείκτες για πρωτεΐνες που είναι περιφερικά αγκυροβολημένες στην EM (F3;Εικόνα 2.1). Στην παρούσα ανάλυση αναφέρουμε πέντε τέτοιες πρωτεΐνες. Τρεις από αυτές (YbiS, ErfK and YcfS) προσδένονται χημικά πάνω στην λιποπρωτεΐνη του Braun (Braun's lipoprotein ,Lpp) πάνω στο πλέγμα της πεπτιδογλυκάνης (ΠΠ)(Sanders et al, 2013). Η MotB είναι ταυτόχρονα δεμένη στο ΠΠ μέσω της καρβοξυτελική της ουράς (C-tail) και στην ΠΜ μέσω ενός αμινοτελικής ΜΠ (O'Neill et al, 2011). Η λιποπρωτεΐνη Pal είναι αγκυροβολημένη στη EM και αλληλεπιδρά με το ΠΠ το καρβοξυτελικό της άκρο (Godlewska et al, 2009).

Το μοτίβο H-T-H (helix-turn-helix) είναι ένα από τα γνωστά δομικά στοιχεία αλληλεπίδρασης με το DNA. Κάποιοι παράγοντες σίγμα (RpoS, RpoE, RpoH) περιέχουν το H-T-H μοτίβο το οποίο μεσολαβεί για την αλληλεπίδρασης με την πρωτεΐνη RpoA (RNA polymerase subunit) αλλά και με το στοιχείο -35 του υποκινητή του DNA. Άλλες πρωτεΐνες που είναι γνωστές για την αλληλεπίδραση τους με DNA ανήκουν στην οικογένεια των ιστόνων (histone-like proteins) (Pettijohn, 1988), η οποία χαρακτηρίζεται από μία συντηρημένη αλληλουχία είκοσι αμινοξέων. Οι πρωτεΐνες DbhA, DbhB, IhfA και IhfB στο *E.coli* ανήκουν στην συγκεκριμένη οικογένεια.

Τέλος το δομικό στοιχείο BON (Bacterial OsmY and Nodulation) σχετίζεται με την αλληλεπίδραση με τα φωσφολιπίδια των μεμβρανών (Yeats et al, 2003). Στο *E.coli* εκτός από την OsmY η οποία περιέχει δύο BON δομικά στοιχεία οι YraP και YgaU περιέχει επίσης από δύο και ένα αντίστοιχα BON στοιχεία.

## 2.8 Βιβλιογραφική αναζήτηση πρωτεϊνών

Για τις υπόλοιπες πρωτεΐνες με αδιευκρίνιστη υποκυτταρική ταξινόμηση πραγματοποιήσαμε εξαντλητική βιβλιογραφική αναζήτηση. Η αναζήτηση οδήγησε στην εύρεση νέων πειραματικών τεκμηρίων για 1205 πρωτεΐνες (Εικόνα 2.4A), η οποία βασίστηκε σε 118 δημοσιευμένες μελέτες. Για 152 πρωτεΐνες που παρέμειναν αταξινομήτες εφαρμόσαμε μια σειρά από κριτήρια (Πίνακας 2.4) και τις κατηγοριοποιήσαμε ως «θεωρητικές» (potential). Η αναζήτηση προσέθεσε πειραματική επιβεβαίωση ακόμα και για πρωτεΐνες για τις οποίες οι τρεις αρχικές πηγές ταυτίζονταν («Ταυτόσημες προτεινόμενες ταξινομήσεις»; Πίνακας 2.6).

Συνοψίζοντας τα αποτελέσματα της διαδικασίας που περιγράψαμε για την επίλυση των διαφορών ανάμεσα στις τρεις πηγές από τις οποίες ξεκινήσαμε την ανάλυση μας: για 200 πρωτεΐνες βρέθηκαν ικανοποιητικά τεκμήρια για την υποκυτταρική τους ταξινόμηση ενώ για 227 η ταξινόμηση βασίστηκε σε επιπλέον κριτήρια (Πίνακας 2.4). Παραδείγματα πρωτεϊνών για τις οποίες διαλευκάναμε τις αντικρουόμενες ταξινομήσεις ή προβλέψεις των BE βρίσκονται στον Πίνακα 2.5.

**Πίνακας 2.4 – Κριτήρια απόφασης για την ταξινόμηση των πρωτεϊνών με βάση τις προβλέψεις των βιοπληροφορικών εργαλείων (BE).**

Βιοπληροφορικά Εργαλεία - Κριτήρια	Ταξινόμηση
<ol style="list-style-type: none"> <li>1. TMHMM δεν προβλέπει ΔΠ</li> <li>2. Phobius προβλέπει τουλάχιστον μία ΔΠ</li> <li>3. <b>και αντίστροφα</b></li> </ol>	Κυτταροπλασματική («θεωρητική»)
<ol style="list-style-type: none"> <li>1. Βασικά BE: κυτταρόπλασμα</li> <li>2. Βοηθητικά BE : «Cell Membrane» ή πρωτεΐνη έχει ανιχνευθεί σε κλάσματα της ΠΜ</li> </ol>	Περιφερική («θεωρητική»)
<ol style="list-style-type: none"> <li>1. SignalP, LipoP, Phobius προβλέπουν ΣΠ</li> <li>2. TMHMM προβλέπει μία ΔΠ</li> <li>3. PSORTB: περίπλασμα</li> </ol>	Περιπλασμική («θεωρητική»)
<ol style="list-style-type: none"> <li>1. Βασικά BE : κυτταρόπλασμα &amp;</li> <li>2. Βοηθητικά BE : «Cell Membrane»</li> </ol>	Κυτταροπλασματική («θεωρητική»)
<ol style="list-style-type: none"> <li>1. Δύο από τα τρία εργαλεία: TMHMM, Phobius, LipoP προβλέπουν ΔΠ</li> </ol>	Λιποπρωτεΐνη της ΠΜ («θεωρητική»)
<ol style="list-style-type: none"> <li>1. LipoP προβλέπει σηματοδοτικό πεπτιδίο τύπου II με Ασπαραγίνη ή Γλουταμίνη στη θέση +2 του ώριμου τμήματος</li> </ol>	Λιποπρωτεΐνη της ΠΜ («θεωρητική»)
<ol style="list-style-type: none"> <li>1. LipoP προβλέπει ΣΠ τύπου II με οποιοδήποτε άλλο κατάλοιπο στην θέση +2 του ώριμου τμήματος</li> </ol>	Λιποπρωτεΐνη της EM («θεωρητική»)

**Πίνακας 2.5 – Παραδείγματα πρωτεϊνών με αντικρουόμενες προτεινόμενες υποκυτταρικές ταξινομήσεις οι οποίες επανεξετάστηκαν και επιλύθηκαν.**

Η σύγκριση των τριών αρχικών πηγών (Uniprot, EchoLOCATION, θεωρητικό μεμβρανικό πρωτεΐνωμα (Bernsel et al, 2009) ανέδειξε αντιφάσεις μεταξύ των προτεινόμενων υποκυτταρικών ταξινομήσεων (conflicts). Ο πίνακας παρουσιάζει 21 παραδείγματα πρωτεϊνών με αντιφατικές ταξινομήσεις ανάμεσα στις τρεις πηγές οι οποίες στην συνέχεια επιλύθηκαν και υποκυτταρικές θέσεις των πρωτεϊνών αναθεωρήθηκαν. Για να διευκολύνουμε την σύγκριση ανάμεσα στις τρεις πηγές αντιστοιχίσαμε τις δικές τους ορολογίες σε αυτή που υιοθετήσαμε στην παρούσα ανάλυση («ορολογία STEPdb»).

Αναγνωριστικό πρωτεΐνης (Uniprot )	Υποκυτταρική Ταξινόμηση (STEPdb)	Υποκυτταρική Ταξινόμηση (Uniprot) [ορολογία STEPdb]	Υποκυτταρική Ταξινόμηση (EchoLOCATION) [ορολογία STEPdb]	Υποκυτταρική Ταξινόμηση (Bernsel et al) [ορολογία STEPdb]	Υποκυτταρική Ταξινόμηση (Επίπεδο αξιοπιστίας)	Κωδικός Δημοσίευσης (PMID)
AAER_ECOLI	N,B	U	A	B	Πυρηνοειδές («πειραματικό»)	[23138451]
ACRR_ECOLI	N	U	A	B	Πυρηνοειδές («πειραματικό»)	[23138451]
ALKB_ECOLI	N	U	A	B	Πυρηνοειδές («πιθανό»)	[20084272]
BIRA_ECOLI	N,B	U	A	B	Πυρηνοειδές («πειραματικό»)	[23138451]
CHEA_ECOLI	F1	A	A	B	Περιφερική πρωτεΐνη της ΠΜ από τη μεριά του κυτταροπλάσματος («πιθανό»)	[19766000]
CPXP_ECOLI	F2	G	F1/F2	U	Περιπλασματική («πιθανό»)	[17904518]
CSGD_ECOLI	N,F1	F1/F2	A	U	Πυρηνοειδές («πειραματικό»), Περιφερική πρωτεΐνη της ΠΜ από τη μεριά του κυτταροπλάσματος («πειραματικό»)	[20874755], [16513732], [23138451]
CYSB_ECOLI	N,F1	A	A	B	Πυρηνοειδές («πειραματικό»), Περιφερική πρωτεΐνη της ΠΜ από τη μεριά του κυτταροπλάσματος («πειραματικό»)	[23138451], [23230279]
DACA_ECOLI	F2	F1	B	U	Περιφερική πρωτεΐνη της ΠΜ από τη μεριά του περιπλάσματος («πειραματικό»)	[2194801], [20192190], [20545860]
DMLR_ECOLI	N	U	A	B	Πυρηνοειδές («πειραματικό»)	[23138451]
DMSA_ECOLI	F2	F1	G	U	Περιφερική πρωτεΐνη της ΠΜ από τη μεριά του κυτταροπλάσματος ή περιπλάσματος. («πιθανό»)	[11389150]
ENO_ECOLI	A,F1,F4	F4	A	U	Περιφερική πρωτεΐνη της ΠΜ από τη	[23230279]

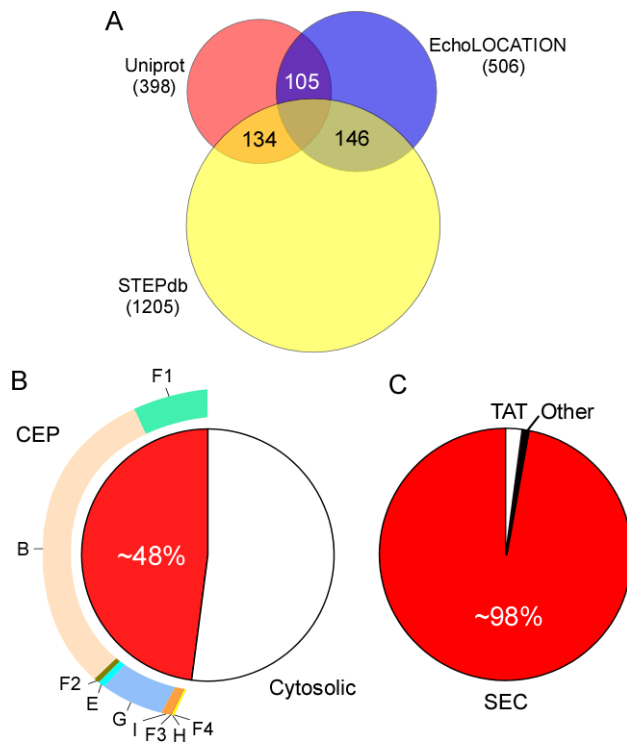
GSPD_ECOLI	H	H	G	U	μεριά του κυτταροπλάσματος («πειραματικό»)	[21931548]
HLYE_ECOLI	G,X	B	X	U	Εξωτερική μεμβράνη («πιθανό»)	[10027972]
HYBA_ECOLI	F2	G	G	B	Περίπλασμα, Εξωκυττάριος χώρος («πιθανό»)	[15911532]
MLRA_ECOLI	N,F1	U	A	B	Εξωτερική μεμβράνη («πειραματικό»)	[20874755]
NFNB_ECOLI	F1	U	A	B	Περιφερική πρωτεΐνη της ΠΜ από τη μεριά του κυτταροπλάσματος («πιθανό»)	[23230279]
PBP2_ECOLI	B	F1/F2	B	B	Περιφερική πρωτεΐνη της ΠΜ από τη («πειραματικό»)	[15919657]
WZA_ECOLI	I, F4	B	I	U	Πλασματική μεμβράνη («πειραματικό»)	[PMC3315050], [10619844]
YCBB_ECOLI	F3	B	G	U	Εξωτερική μεμβράνη, λιποπρωτεΐνη («πιθανό»),Επιφάνεια του κυττάρου («πιθανό»)	[23832002]
					Περιφερική πρωτεΐνη της ΠΜ από τη μεριά του περιπλάσματος («πειραματικό»)	

**Πίνακας 2.6 - Σύνοψη της ήδη υπάρχουσας ταξινόμησης και σύγκριση με την ταξινόμηση στη βάση δεδομένων STEPdb.**

Οι βάσεις δεδομένων Uniprot και EchoLOCATION συνεισέφεραν 2050 και 4243 υποκυτταρικές ταξινομήσεις πρωτεϊνών ενώ το θεωρητικό ΜΠ (Bernsel et al, 2009) 1108 μεμβρανικές πρωτεΐνες. Συνδυαστικά από τις τρεις πηγές συλλέξαμε πληροφορία για 4267 υποκυτταρικές ταξινομήσεις πρωτεϊνών. Η σύγκριση μεταξύ των τριών πηγών αποκάλυψε: ένα υποσύνολο πρωτεϊνών για τις οποίες υπήρχε συμφωνία ανάμεσα στις προτεινόμενες υποκυτταρικές ταξινομήσεις («Ταυτόσημες προτεινόμενες ταξινομήσεις»), ένα άλλο υποσύνολο πρωτεϊνών με διαφορετικές προτεινόμενες ταξινομήσεις μεταξύ των πηγών («Αντικρουόμενες προτεινόμενες ταξινομήσεις»). Πρωτεΐνες με διαθέσιμη ταξινόμηση μόνο από μια πηγή αναφέρονται ως «Μοναδικές πρωτεΐνες (συνολικά)». Πρωτεΐνες που αναφέρονται ως μεμβρανικές μόνο από μια από τις τρεις πηγές ενώ στις υπόλοιπες αποτελούν άγνωστες «Μοναδικές πρωτεΐνες (μεμβρανικές)». Η συνδυαστική ανάλυση που ακολουθήσαμε συνεισέφερε 36 πρωτεΐνες με άγνωστο υποκυτταρικό εντοπισμό τις οποίες ταξινομήσαμε *de novo* («*de novo* ταξινομήσεις»), για 674 πρωτεΐνες η προτεινόμενη ταξινόμηση από τις τρεις πηγές, αναθεωρήθηκε («Αναθεωρημένες ταξινομήσεις») και για 601 πρωτεΐνες έγινε διασάφηση των συγκρουόμενων ταξινομήσεων («Αντιφάσεις που επιλύθηκαν»).

	Uniprot	EchoLOCATION	Bernsel & Daley (2009)	Σύνολο
Πρωτεΐνωμα Αναφοράς ( <i>E.coli</i> K-12)	4303	4345*	1133	4303
Ταυτόσημες προτεινόμενες ταξινομήσεις	1613	1646	850	1652
Μοναδικές πρωτεΐνες (μεμβρανικές)	11	29	4	44
Μοναδικές πρωτεΐνες (συνολικά)	12	1998	4	2014
Αντικρουόμενες προτεινόμενες ταξινομήσεις	425	599	254	601
Σύνολο προτεινόμενων ταξινομήσεων	2050	4243	1108	4267
<b>% του πρωτεϊνώματος αναφοράς</b>	<b>48%</b>	<b>98%</b>	<b>26%</b>	<b>99%</b>
Άγνωστες	2253	60	-	36
Άγνωστες και αντικρουόμενες προτεινόμενες ταξινομήσεις	2678	659	254	637
<b>% του πρωτεϊνώματος αναφοράς</b>	<b>62%</b>	<b>15%</b>	<b>6%</b>	<b>18%</b>
Συνολική συνεισφορά στις είδη υπάρχουσες ταξινομήσεις (STEPdb)	3352	1333	560	1311
<i>De novo</i> ταξινομήσεις (STEPdb)	2253	60	84	36
Αναθεωρημένες ταξινομήσεις (STEPdb)	674	674	222	674
Αντιφάσεις που επιλύθηκαν (STEPdb)	425	599	254	601
Πειραματικά επαληθευμένες πρωτεΐνες				1205
Προστιθέμενες βιβλιογραφικές αναφορές				118
<b>% του πρωτεϊνώματος αναφοράς</b>	<b>76.89%</b>	<b>35.37%</b>	<b>7.87%</b>	<b>32.81%</b>





**Εικόνα 2.4 – Σύνοψη της υποκυτταρικής ταξινόμησης των πρωτεϊνών και των πειραματικά επαληθευμένων πρωτεϊνών.**

A) Πρωτεΐνες με πειραματικά επαληθευμένο υποκυτταρικό εντοπισμό στις δύο βάσεις δεδομένων. B) Πρωτεΐνωμα του ΚΦ και η κατανομή των πρωτεϊνών στα διάφορα υποκυτταρικά διαμερίσματα. C) Ποσοστό πρωτεϊνών που χρησιμοποιούν τα βασικά εκκριτικά συστήματα στο *E.coli*.

## 2.9 Πρωτεΐνες με πολλαπλές υποκυτταρικές τοποθεσίες.

Η ανάλυση μας έδειξε ότι 60 πρωτεΐνες του K-12 μπορούν να βρίσκονται σε πολλαπλές υποκυτταρικές τοποθεσίες (Εικόνα 2.2, "MT: multiple topologies") καθιστώντας το πρωτεΐνωμα πολύ πιο δυναμικό από τις μέχρι τώρα εκτιμήσεις. Πρωτεΐνες με παραπάνω από μία θέσεις μέσα στο κύτταρο δεν υπάρχουν σαν ξεχωριστή κατηγορία στο Uniprot και EchoLOCATION.

Πιο αναλυτικά 37 πρωτεΐνες αποτελούν περιφερικές πρωτεΐνες της ΠΜ αλλά είναι επίσης πρωτεΐνες του πυρηνοειδούς (N,F1). Πέντε πρωτεΐνες (ArcD, YgjI, ClocB, RodZ and CadC) έχουν επιβεβαιωθεί πειραματικά ως μεμβρανικές (Bendezu et al, 2009b; Daley et al, 2005; Rauschmeier et al, 2014). Για την πρωτεΐνη SecM και έχει επιβεβαιωθεί πειραματικά ότι μπορεί να βρίσκεται στο κυτταρόπλασμα και στο περίπλασμα (Link et al, 1997; Rajarandi et al, 1991) (A,G). Οι πρωτεΐνες OsmY, ChiA και HlyE έχουν ταυτοποιηθεί ως περιπλασματικές και εξωκυττάρειες πρωτεΐνες (G,X). Συγκεκριμένα η ChiA εκκρίνεται όταν ενεργοποιείται το εκκριτικό σύστημα T2S (Francetic et al, 2000). Η κυτταροπλασματική πρωτεΐνη ArfB (peptidyl-tRNA hydrolase) προσδένεται επίσης και στα ριβοσώματα μέσω του καρβοξυτελικού της άκρου (C-tail) (Handa et al, 2011). Τρεις λιποπρωτεΐνες

της EM (Wza, CsgG and Lpp), οι οποίες υπό φυσιολογικές συνθήκες βρίσκονται αγκυροβολημένες στην εσωτερική στοιβάδα της EM, έχει αποδειχτεί ότι μπορούν και εκκρίνονται στην επιφάνεια του κυττάρου (Cowles et al, 2011; Drummelsmith et al, 2000; Robinson et al, 2006). Δύο πρωτεΐνες του μαστιγίου (flagellum) FliK και ο παράγοντας αντί-σίγμα FlgM, βρίσκονται στο κυτταρόπλασμα και εκκρίνονται σε κατάλληλη χρονική στιγμή από το μαστίγιο. Δύο μεμβρανικές πρωτεΐνες, CyoA και YiaD, έχουν δειχθεί πειραματικά ότι είναι υποστρώματα της πεπτιδάσης τύπου II, καθιστώντας τις επίσης λιποπρωτεΐνες. Τέλος η πρωτεΐνη Eno (enolase) παρατηρείται σε τρεις καταστάσεις: κυτταροπλασματική, περιφερική στην ΠΜ αλλά και προσδεμένη στην επιφάνεια του κυττάρου (Boel et al, 2004).

## 2.10 Περιφερικές πρωτεΐνες

Μια ιδιαίτερη κατηγορία πρωτεϊνών με διπλή ιδιότητα είναι οι περιφερικές πρωτεΐνες (PIM proteins), καθώς αλληλεπιδρούν με την μεμβράνη αλλά ταυτόχρονα διατηρούν την ικανότητα τους να είναι διαλυτές στο κυτταρόπλασμα (Papanastasiou et al, 2013). Οι πρωτεΐνες αυτές αποτελούν το δίαυλο επικοινωνίας μεταξύ του κυτταροπλάσματος και του ΚΦ για πολλές κυτταρικές λειτουργίες (Papanastasiou et al, 2013). Οι περιφερικές πρωτεΐνες είναι ελλιπώς ταξινομημένες στις σύγχρονες βάσεις δεδομένων πιθανόν λόγω απουσίας κατάλληλων BE για την πρόβλεψη τους. Το EchoLOCATION αναφέρει ότι υπάρχουν 10 πρωτεΐνες που σχετίζονται με την μεμβράνη (“membrane-associated”) ενώ το Uniprot καταγράφει συνολικά 139 τέτοιες πρωτεΐνες (“membrane-related”) εκ των οποίων 127 αναφέρει ότι αλληλεπιδρούν με την εξωτερική στοιβάδα της ΠΜ (“associate with the IM from the cytoplasmic side”).

Πειραματικά οι περιφερικές πρωτεΐνες μπορούν να προσδιοριστούν μέσω της αλληλεπίδρασης τους με την μεμβράνη (Papanastasiou et al, 2013). Η ανάλυση μας ανέδειξε συνολικά 550 περιφερικές πρωτεΐνες εκ των οποίων 37 αποτελούν πρωτεΐνες με πολλαπλές τοποθεσίες, 392 είναι πειραματικά επαληθευμένες, 76 κατηγοριοποιήθηκαν ως «πιθανές» (“probable”), 20 ως «θεωρητικές» (“potential”) και 26 «βάση ομοιότητας» (“by similarity”).

## 2.11 Σύγκριση και αντιστοίχιση δύο πρωτεϊνωμάτων *E.coli*

Το πρωτεϊνωμα του στελέχους *E.coli* K-12 (MG1655) αντιπαρατέθηκε με το στέλεχος BL21-DE3. Το εργαστηριακό στέλεχος BL21-DE3 αποτελεί τον πιο συχνά χρησιμοποιούμενο ξενιστή για την παραγωγή γονιδιακά ανασυνδυασμένων πρωτεϊνών και θεωρείται γενικά πιο σταθερό σε σχέση με το

στέλεχος K-12 λόγω της έλλειψης κάποιων πρωτεολυτικών ενζύμων αλλά και λόγω μειωμένης παραγωγής οξικού άλατος (Veit et al, 2007; Yoon et al, 2012). Συγκρίνοντας τα δύο στελέχη προέκυψε ότι το πάνω από το 90% των πρωτεϊνών είναι κοινές (common proteome) ενώ απομένουν 266 και 359 μοναδικές πρωτεΐνες στο K-12 και BL21-DE3 αντίστοιχα (Πίνακας 2.1). Για τις κοινές πρωτεΐνες έγινε αναγωγή της υποκυτταρικής ταξινόμησης από το στέλεχος K-12 στο BL21-DE3. Για τις μοναδικές πρωτεΐνες του BL21-DE3 ακολουθήθηκε η ίδια αντίστοιχη διαδικασία ταξινόμησης με αυτή των άγνωστων πρωτεϊνών στο K-12 (δες ενότητα 2.6).

## **2.12 Οργάνωση των εκκριτικών μηχανισμών στο *E.coli* και προσδιορισμός των πρωτεϊνών που συμμετέχουν σε αυτά.**

Η ταξινόμηση στην οποία καταλήξαμε για το πρωτεϊνωμα του K-12 αποκαλύπτει ότι ένα αξιοσημείωτο ποσοστό της τάξης του 48% συντίθεται στο κυτταρόπλασμα αλλά μεταφέρεται στα διάφορα υποκυτταρικά διαμερίσματα του ΚΦ (Εικόνα 2.4B). Πριν μεταφερθούν στις τελικές τους θέσεις στον ΚΦ οι πρωτεΐνες πρέπει να ξεπεράσουν τις βιολογικές μεμβράνες οι οποίες αποτελούν σημαντικά ενεργειακά εμπόδια. Το *E.coli* έχει εξελίξει διάφορους εκκριτικούς μηχανισμούς οι οποίοι, ως επί το πλείστον, δαπανώντας ενέργεια είτε μετατοπίζουν πρωτεΐνες-υποστρώματα διαμέσου των μεμβρανών είτε τις ενσωματώνουν σε αυτές. Η βάση δεδομένων STEPdb που υλοποιήσαμε συγκεντρώνει 12 γνωστά εκκριτικά μονοπάτια στο K-12 (Εικόνα 2.5) (Barnhart et al, 2006; Chatzi et al, 2013; Douzi et al, 2012; Henderson et al, 2000; Lee et al, 2006; Miyamoto et al, 2007; Prehna et al, 2012; Selkrig et al, 2012; Solov'eva et al, 2012; Van Gerven et al, 2011). Επιχειρήσαμε επίσης να προσδιορίσουμε, όπου είναι δυνατόν, τις αντίστοιχες πρωτεΐνες-υποστρώματα που χρησιμοποιούν τα συγκεκριμένα μονοπάτια.

### **2.12.1. Εκκριτικά Μονοπάτια στο *E.coli* K-12**

Τα εκκριτικά μονοπάτια στο K-12 μπορούν να διαχωριστούν σε δύο βασικές κατηγορίες τα Sec-εξαρτώμενα (LOL, BAM, TAM, CU, T5SS, Curli, OMV, και T2SS) και τα μη εξαρτώμενα από το Sec (Flagellum και TAT) (Εικόνα 2.5). Στην πρώτη περίπτωση το σύστημα Sec αναλαμβάνει να μεταφέρει τις πρωτεΐνες διαμέσου της ΠΜ, κατόπιν δευτερεύοντα συστήματα αναλαμβάνουν τη διαλογή πρωτεϊνών στο περίπλασμα, την ΕΜ αλλά και εκτός κυτάρου. Σε γενικές γραμμές η συγκεκριμένη κατηγοριοποίηση ισχύει και στα υπόλοιπα βακτήρια μολονότι είναι γνωστό ότι σε κάποια στελέχη υπάρχουν επιπλέον μονοπάτια τα οποία απουσιάζουν στο K-12. Συνοψίζοντας, η

---

ανάλυση μας δείχνει ότι το Sec σύστημα εξυπηρετεί το μεγαλύτερο μέρος των πρωτεϊνών (>98%; Εικόνα 2.4C; Εικόνα 2.5) .

### 2.12.2. Εκκριτικές πρωτεΐνες

Τα εκκριτικά μονοπάτια οργανώθηκαν σε δύο επίπεδα τα οποία λειτουργούν σε συνεργασία με δύο αντίστοιχα επίπεδα στόχευσης (chaperone/targeting). Τα δύο επίπεδα έκκρισης αναφέρονται στο πρώτο βήμα της έκκρισης από την ΠΜ (Sec, TAT ή YidC) και το δεύτερο βήμα της έκκρισης από την ΕΜ (π.χ. BAM, TAM, CU). Αντίστοιχα τα δύο επίπεδα στόχευσης αναφέρονται σε διαλυτές πρωτεΐνες που λειτουργούν σαν μοριακοί οδηγοί που καθοδηγούν τις πρωτεΐνες στα αντίστοιχα συστήματα πάνω στην μεμβράνη (π.χ SecB για το SecA/SecYEG σύστημα, SRP για το SecYEG σύστημα, Skp για το BAM κλπ).

Η προσδιορισμός των πρωτεϊνών που χρησιμοποιεί κάθε εκκριτικό σύστημα βασίστηκε κυρίως στα BE που προβλέπουν καλά χαρακτηρισμένα μοτίβα ή δευτεροταγείς δομές, όπως τα ΣΠ τύπου Sec/Tat και οι ΔΠ (Bendtsen et al, 2005b; Juncker et al, 2003; Petersen et al, 2011), σε *in vivo* (Hiniker et al, 2004) και *in vitro* (Martinez-Hackert et al, 2009) πρωτεϊνωματικά/βιοχημικά δεδομένα και τέλος σε πρωτεομικές αναλύσεις (Baars et al, 2006; Ishihama et al, 2008; Iwasaki et al, 2010; Masuda et al, 2009). Να σημειωθεί ότι υπάρχουν πρωτεΐνες που εκκρίνονται μέσω του Sec συστήματος χωρίς να έχουν προφανή σηματοδοτικές αλληλουχίες όπως η SodA στο *Rhizobium* (ομόλογες πρωτεΐνες στο K-12 είναι οι SodM και SodF (Krehenbrink et al, 2011)).

Δύο υποκατηγορίες πρωτεϊνών χρησιμοποιούν το Sec σύστημα (Dalbey et al, 2012). Στις τύπου I (π.χ. περιπλασμικές, πρωτεΐνες της ΕΜ και εξωκυτάρειες πρωτεΐνες) τα ΣΠ αφαιρούνται χημικά με την βοήθεια του ενζύμου της πεπτιδάσης τύπου I (SPaseI). Η αντίδραση αυτή λαμβάνει χώρα στην επιφάνεια της ΠΜ. Οι τύπου II πρωτεΐνες, εξελικτικά έχουν διαφορετικά ΣΠ τα οποία κόβονται από το ένζυμο της πεπτιδάσης τύπου II (SPaseII). Σε αυτήν την κατηγορία ανήκουν οι λιποπρωτεΐνες της ΠΜ και ΕΜ (Miyamoto et al, 2007). Ορισμένες λιποπρωτεΐνες εκκρίνονται σε αναδιπλωμένη μορφή με την συνεργασία του συμπλόκου Sec και της πρωτεΐνης YidC (Froderberg et al, 2004).

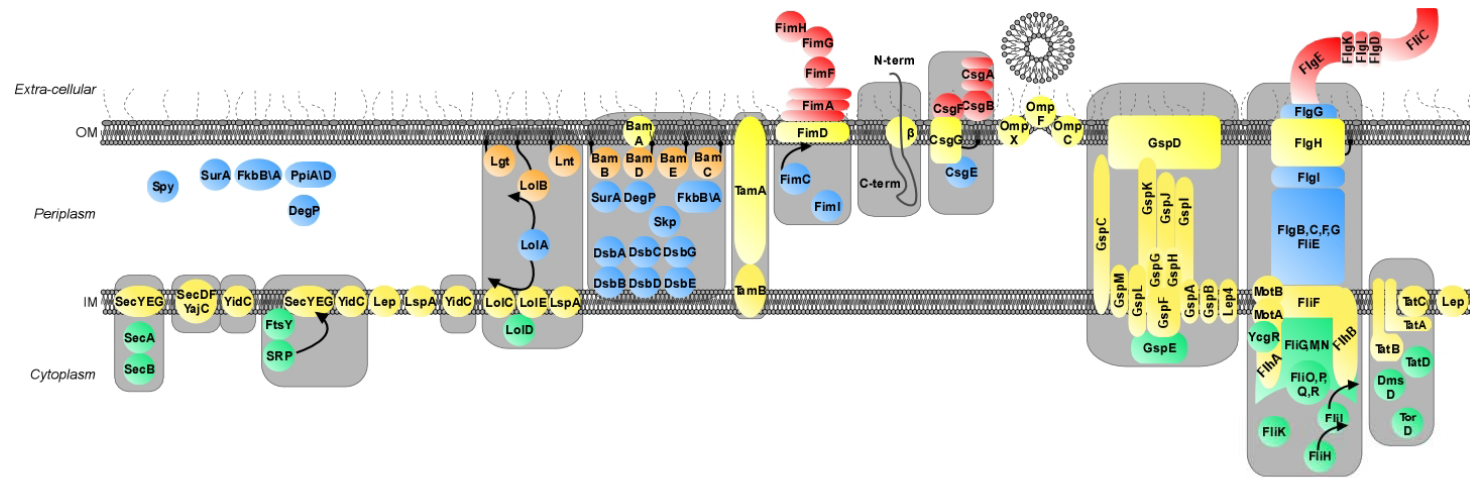
Οι μεμβρανικές πρωτεΐνες πιστεύεται ότι ακολουθούν το μονοπάτι Srp και ότι εισέρχονται στην ΜΠ με συν-μεταφραστικό τρόπο μετατόπισης από το Sec σύστημα (Tian et al, 2009). Απουσία πειραματικών δεδομένων υιοθετήσαμε αυτήν την θεωρία και κατηγοριοποιήσαμε όλες τις μεμβρανικές

---

πρωτεΐνες ως Srp/Sec εκκρινόμενες. Εξαίρεση αποτελεί η μεμβρανική πρωτεΐνη HyaA, με μια ΔΠ, η οποία φαίνεται οδηγείται στην ΠΜ ακολουθώντας το μονοπάτι Srp ενώ εκκρίνεται από το σύστημα TAT (Tullman-Ercek et al, 2007). Μέσω φθορίζουσας επισήμανσης προσδιορίστηκαν μεμβρανικές πρωτεΐνες οι οποίες είναι YidC εξαρτώμενες (YidC-dependent) (Gray et al, 2011). Επίσης για ~400 μεμβρανικές πρωτεΐνες μελετήθηκε η ικανότητα ενσωμάτωσης τους στην ΠΜ σε συνθήκες απουσίας της YidC (Gray et al, 2011). Τα αποτελέσματα της μελέτης έδειξαν ότι 77 πρωτεΐνες εξαρτώνται από την YidC όπως επίσης και μερικές μικρές πρωτεΐνες όπως η M13 και Pf3 (Fontaine et al, 2011; Gray et al, 2011). Αξίζει να αναφέρουμε ότι αρνητικά φορτία στα περιπλασματικά τμήματα ή στις ΔΠ φαίνεται ότι σχετίζονται με την επιλεκτικότητα που παρουσιάζουν τα υποστρώματα στην πρωτεΐνη YidC (Dalbey et al, 2014).

Συνολικά, 33 πρωτεΐνες ακολουθούν το εκκριτικό μονοπάτι Tat οι οποίες στην πλειοψηφία τους έχουν χαρακτηριστικά ΣΠ τύπου Tat (Tullman-Ercek et al, 2007). Ιδιαίτερο ενδιαφέρον παρουσιάζουν κάποιες πρωτεΐνες που χρησιμοποιούν το συγκεκριμένο εκκριτικό σύστημα παράλο που στερούνται ΣΠ (π.χ. DmsB). Για αυτές έχει διαπιστωθεί ότι εκκρίνονται ως σύμπλοκα με άλλες πρωτεΐνες του εκκριτικού μονοπατιού Tat ("piggy-backing") (π.χ. DmsA) (Bilous et al, 1988; Neumann et al, 2009).

Τέλος βρήκαμε τουλάχιστον δύο πρωτεΐνες (HlyE και YebF) οι οποίες εκκρίνονται στην επιφάνεια του κυττάρου μέσω κυστιδίων της EM (outer membrane vesicles) (Lee et al, 2007a; Lee et al, 2007b; Prehna et al, 2012).



**Export Systems**



**Protein Topologies**

<ul style="list-style-type: none"> <li><span style="color: yellow;">●</span> outer membrane β-barrel protein (H)</li> <li><span style="color: orange;">●</span> outer membrane lipoprotein (I)</li> <li><span style="color: red;">●</span> peripherally associated with the outer membrane facing the extra-cellular space (F4)</li> </ul>	<ul style="list-style-type: none"> <li><span style="color: blue;">●</span> periplasmic protein (G)</li> <li><span style="color: yellow;">●</span> integral Inner membrane protein (B)</li> <li><span style="color: green;">●</span> peripherally associated with the plasma membrane facing the cytoplasm (F1)</li> </ul>
--	---

---

**Εικόνα 2.5 - Τα βασικά εκκριτικά μονοπάτια στο βακτήριο *E.coli***

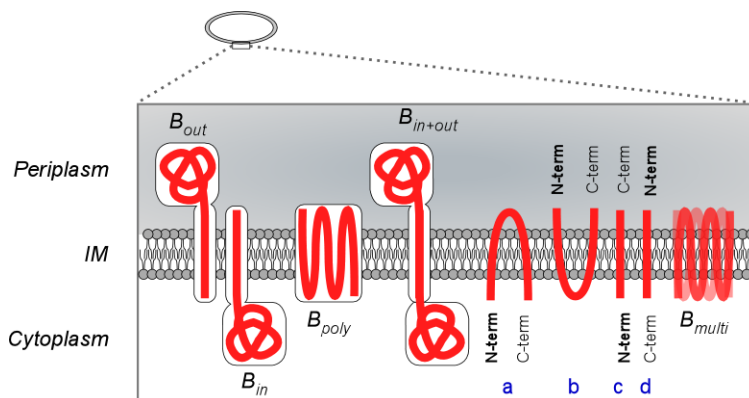
Σχηματική απεικόνιση των μέχρι σήμερα γνωστών εκκριτικών μηχανισμών στο *E.coli* με τις αντίστοιχες δομικές υπομονάδες αλλά και τα στοιχεία στόχευσης (targeting) και μοριακής καθοδήγησης (chaperoning). Τα συστήματα έκκρισης μπορούν να διαχωριστούν σε Sec εξαρτώμενα (Sec depended) και μη εξαρτώμενα από το Sec (non-Sec depended). Τα Sec εξαρτώμενα συστήματα είναι: SEC, το βασικό και απαραίτητο εκκριτικό μονοπάτι μέσω του οποίου το μεγαλύτερο μέρος των πρωτεϊνών εκκρίνεται ή εισέρχεται μέσα στην ΠΜ; SRP-SEC, το συ-μεταφραστικό εκκριτικό μονοπάτι, γνωστό για την στόχευση και ενσωμάτωση των μεμβρανικών πρωτεϊνών στην ΠΜ; LOL, το σύστημα διαλογής των λιποπρωτεϊνών; BAM, το σύμπλοκο για την συναρμολόγηση των βήτα βαρελοειδών πρωτεϊνών της EM (outer membrane  $\beta$ -barrels); TAM, υπομονάδα για την μετατόπιση και συναρμολόγηση των αυτομεταφορέων (autotransporters) στην εξωτερική μεμβράνη; CU, εκκριτικά μονοπάτια τύπου chaperone usher; T5S, εκκριτικό σύστημα τύπου πέντε γνωστό επίσης ως το εκκριτικό μονοπάτι των αυτομεταφορέων (autotransporters) (αποτελείται από τρία δομικά στοιχεία: την αλληλουχία οδηγό, το τμήμα «επιβάτη» (passenger domain) και το τμήμα βήτα ); Curli, εξωκυττάρειες αμυλοειδείς ίνες (extracellular amyloid fibers); OMV, κυστίδια της εξωτερικής μεμβράνης (outer membrane vesicles); T2S, εκκριτικό σύστημα τύπου δύο το οποίο μεσολαβεί για την έκκριση αναδιπλωμένων πρωτεϊνών στο περίπλασμα. Τα μη εξαρτώμενα από το Sec εκκριτικά συστήματα (non-Sec depended) είναι: Flagellum, οργανίδιο το οποίο λειτουργεί ως γεννήτρια κίνησης ενώ είναι ικανό να εκκρίνει μερικές από τις υπομονάδες από τις οποίες αποτελείται; TAT, ελάσσων σύστημα έκκρισης Tat (twin arginine translocation system). Η μεμβρανική πρωτεΐνη YidC εμπλέκεται στην εισαγωγή μερικών περιπτώσεων πρωτεϊνών σε συνεργασία με το μεμβρανικό σύμπλοκο SecYEG. Είναι γνωστό ότι μπορεί να λειτουργήσει και ανεξάρτητα και συνεπώς ορίζει από μόνη της ένα ξεχωριστό εκκριτικό μονοπάτι. Επιπλέον στοιχεία δρουν βοηθητικά στα αντίστοιχα εκκριτικά συστήματα λειτουργώντας ως μοριακοί οδηγοί (chaperones) και ως στοιχεία στόχευσης. Στην παρούσα απεικόνιση παρουσιάζονται: η πρωτεΐνη SecB μοριακός οδηγός του κυτταροπλάσματος, η περιφερική πρωτεΐνη Srp (signal recognition ribonucleoprotein particle), οι πρωτεΐνες μοριακοί οδηγοί του περιπλάσματος (Skr και DegP), οι ισομεράσες πεπτιδικών δεσμών (SurA, FkbAB, και PpiAD) και οι περιπλασμικές οξειδοαναγωγάσες δισουλφιδίων (disulfide oxidoreductases) DsbABCD. Ο πλήρης κατάλογος των πρωτεϊνών είναι καταχωρημένος και διαθέσιμος στη βάση δεδομένων STEPdb (<http://stepdb.eu>).

---

### 2.13 Η βάση δεδομένων STEPdb (Sub-cellular Topologies of E.coli Polypeptides database)

Η τελική υποκυτταρική ταξινόμηση των πρωτεϊνών του *E.coli* και η πληροφορία που συλλέχθηκε στην πορεία βιβλιογραφικής αναζήτησης καταχωρήθηκε και οργανώθηκε σε μια βάση δεδομένων (STEPdb). Βρίσκεται διαθέσιμη μέσω μιας ιστοσελίδας που κατασκευάσαμε (στην ηλεκτρονική διεύθυνση <http://stepdb.eu>) η οποία ενσωματώνει επίσης βοηθητικά βιοπληροφορικά εργαλεία όπως το “BLAST2STEP” και το “Predict Topology”.

Στο STEPdb η πληροφορία είναι οργανωμένη σε πίνακες που είναι προσπελάσιμοι μέσω ενός καταλόγου συνδέσμων στα αριστερά του παράθυρου. Οι υπηρεσίες της ιστοσελίδας είναι χωρισμένες σε δύο ομάδες Α) τα στελέχη *E.coli* (“Strains”) η οποία συγκεντρώνει πίνακες που οργανώνουν τις περιφερικές πρωτεΐνες, τις Sec/TAT εκκρινόμενες πρωτεΐνες, τις μεμβρανικές πρωτεΐνες και μια αντιστοίχιση ανάμεσα στα δύο στελέχη *E.coli* και Β) τα βοηθητικά εργαλεία και τους διαθέσιμους πίνακες προς λήψη (“Downloads and Tools”) ανάμεσα στους οποίους περιλαμβάνονται λίστες πειραματικών δεδομένων και δημοσιεύσεις.



**Εικόνα 2.6 – Προσανατολισμός μεμβρανικών πρωτεϊνών**

Οι μεμβρανικές πρωτεΐνες κατηγοριοποιήθηκαν με βάση τον προσανατολισμό τους στην ΠΜ. Bin, Bout μεμβρανικές πρωτεΐνες μονής ΔΜ (μονοτοπικές) με το διαλυτό τους κομμάτι κυρίως στο κυτταρόπλασμα ή στο περίπλασμα αντίστοιχα; Bpoly μεμβρανικές πρωτεΐνες με πολλαπλές ΔΠ (πολυτοπικές); Bin+out μονοτοπικές μεμβρανικές πρωτεΐνες με διαλυτές περιοχές στο περίπλασμα αλλά και στο κυτταρόπλασμα; Ταξινόμηση μεμβρανικών πρωτεϊνών με βάση τη τοποθεσία των άκρων τους: a,b και τα δύο άκρα στο κυτταρόπλασμα και περίπλασμα αντίστοιχα; c,d άκρα σε αντίθετες θέσεις σε σχέση με το επίπεδο της ΠΜ, Bmulti διπλός προσανατολισμός.



Η κατηγορία «K-12» περιλαμβάνει τις λίστες πρωτεϊνών για υποκατηγορίες όπως οι περιφερικές, οι μεμβρανικές και οι Sec/Tat εκκρινόμενες καθώς επίσης και μια σύνοψη των χαρακτηριστικών τους. Κάτω από την κατηγορία των μεμβρανικών πρωτεϊνών βρίσκεται μια βοηθητική εφαρμογή η οποία σχεδιάζει την τοπολογία ~700 μεμβρανικών (“Topology”) για τις οποίες ο προσανατολισμός τους πάνω στην ΠΜ έχει επιβεβαιωθεί πειραματικά. Στη συγκεκριμένη απεικόνιση ο αριθμός των ΔΜ κάθε πρωτεΐνης βασίστηκε στην πρόβλεψη του βιοπληροφορικού εργαλείου Phobius ενώ ο αντίστοιχος προσανατολισμός πάνω στην ΠΜ βασίστηκε σε πειραματικά δεδομένα ανίχνευσης του καρβοξυτελικού άκρου (C-termini) (Daley et al, 2005). Συγκρίναμε τα πειραματικά δεδομένα με τις προβλέψεις δύο BE και φαίνεται ότι το TMHMM προβλέπει σωστό προσανατολισμό για το 78% των περιπτώσεων(3) ενώ το Phobius για το ~81%. Επιπλέον το STEPdb οργανώνει τις μεμβρανικές πρωτεΐνες με κριτήριο Α) τη θέση των δύο άκρων τους (κυτταρόπλασμα/περίπλασμα) και Β) του αριθμού των ΔΠ τους (μονοτοπικές με μια ΔΠ, και πολυτοπικές με περισσότερες από μία ΔΠ) (Εικόνα 2.6).

### 2.13.1. Βιοπληροφορικά εργαλεία που συμπεριλαμβάνονται στο STEPdb

Το STEPdb ενσωματώνει τέσσερα BE: TMHMM, Phobius και SignalP, τα οποία προβλέπουν ΣΠ και ΔΠ και το IUPred το οποίο προβλέπει περιοχές με αταξία (disordered regions) (Dosztanyi et al, 2005a). Ο χρήστης έχει την δυνατότητα να εκτελέσει ταυτόχρονα τα παραπάνω εργαλεία και να συγκρίνει τα αποτελέσματα τους μέσω του συνδέσμου «Predict Topology».

Αναζητήσεις ομοιότητας αλληλουχιών μέσα στο STEPdb μπορούν να εκτελεστούν μέσω του βοηθητικού εργαλείου BLAST2STEP. Το εργαλείο αυτό μπορεί να προβλέψει το υποκυτταρικό εντοπισμό μιας άγνωστης πρωτεΐνης με βάση την εμπειριστατωμένη ταξινόμηση των πρωτεϊνών του μοντέλου οργανισμού *E.coli*.

### 2.13.2. Δομικές και φυσικοχημικές ιδιότητες των πρωτεϊνών

Πρωτεΐνες που εμπεριέχουν περιοχές με δομική αταξία ανήκουν στην κατηγορία των εγγενώς δομικά διαταραγμένων πρωτεϊνών (intrinsically disordered – IDPs; (Tompa, 2002)). Ορισμένες από αυτές τις πρωτεΐνες έχουν σημαντικές λειτουργίες μέσα στο κύτταρο (Tompa, 2002). Οι εγγενώς άτακτες πρωτεΐνες συμμετέχουν σε σηματοδοτικά (Kishii et al, 2007) και ρυθμιστικά μονοπάτια (Dunker et al, 2008) και συχνά προσδένονται στο DNA (Chang et al, 2010),

---


στα ριβοσώματα (Handa et al, 2011), σε μικρά μόρια και πρωτεΐνες (Gajiwala et al, 2000). Η βάση δεδομένων STEPdb ενσωματώνει το εργαλείο IUPred (Dosztanyi et al, 2005a) το οποίο προβλέπει περιοχές με μεγάλη πιθανότητα αταξίας.

Φυσικοχημικές ιδιότητες όπως υδροφοβικότητα και διαλυτότητα, αντανακλώνται από αλλά και καθορίζουν τα δομικά χαρακτηριστικά πρωτεϊνών, όπως η ανάγκη για μοριακό οδηγό κατά την έκκριση Niwa et al (Niwa et al, 2009). Η βάση δεδομένων STEPdb συνοψίζει τη διαλυτότητα των πρωτεϊνών όπως αυτές μετρήθηκαν σε ένα *in vitro* σύστημα όπου απουσιάζουν οι μοριακοί οδηγοί. Ενδιαφέρονται συμπεράσματα προέκυψαν από την ανάλυση της διαλυτότητας των πρωτεϊνών όπως: το μεμβρανικό πρωτεΐνωμα αποτελείται κυρίως από πρωτεΐνες με χαμηλή διαλυτότητα (solubility <30%), ενώ 65% των ριβοσωμικών πρωτεϊνών είναι σε μεγάλο βαθμό διαλυτές (solubility >70%).

### 2.13.3. Απεικόνιση πρωτεϊνικών συμπλόκων

Στο STEPdb γίνεται μια πρώτη προσπάθεια να οργανωθούν τα πρωτεϊνικά σύμπλοκα (Paranastasiou et al, 2013) στα υποκυτταρικά διαμερίσματα που λειτουργούν. Το STEPdb συγκεντρώνει και απεικονίζει τα πρωτεϊνικά σύμπλοκα του *E.coli* τα οποία προήλθαν κυρίως από την βάση δεδομένων EcoCyc (957 σύμπλοκα). Σε αυτά προσθέσαμε 61 καινούργια σύμπλοκα τα οποία προέκυψαν στην πορεία της βιβλιογραφικής αναζήτησης και δεν υπήρχαν καταχωρημένα στη βάση δεδομένων EcoCyc. Ένα ενδιαφέρον παράδειγμα είναι η Psd η οποία πρωτεολύεται και σχηματίζει σύμπλοκο με τα δύο μέρη του ευατού της (Tyhach et al, 1979).

Με βάση την τοποθεσία των υπομονάδων κάθε συμπλόκου ορίσαμε την υποκυτταρική τοποθεσία του συμπλόκου. Ενδιαφέρον παρουσιάζουν τα σύμπλοκα που διασχίζουν και τις δύο κυτταρικές μεμβράνες όπως το μαστίγιο (flagellum). Στη βάση δεδομένων STEPdb είναι επίσης διαθέσιμη μια συνολική απεικόνιση των συμπλόκων του *E.coli* ("Cell Atlas").



Search:  in K-12

[Basic](#) | [Advanced](#)

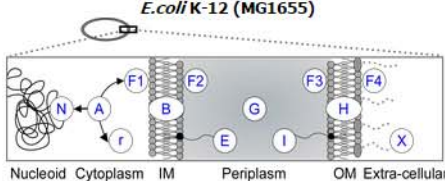
**Strains**

**K-12**

- Cell Atlas
- Peripherome
- Complexome
- Membranome (IM)
  - Sequences
  - Topology
- Secretomes
  - SEC
  - Features
  - Sequences
- TAT
  - Features
  - Sequences
- Multiple Locations
- Export Systems
- Solubility

BL21

K-12 vs BL21



**E.coli K-12 (MG1655)**

Nucleoid Cytoplasm IM Periplasm OM Extra-cellular

[Cell Atlas](#)

[show distribution](#)

Subcellular Localization

STEPdb		Subcellular Localization															
Protein ID (Uniprot)	Accession (Uniprot)	Gene Name (Uniprot)	Order Locus Name	Protein Name (Uniprot)	Full Name	Symbol	Uniprot	EchoLOCATION	Bernsel et al	Extracellular Appendage	First Level Targeting Chaperoning	First Level Secretion System	Second Level Targeting Chaperoning	Second Level Secretion System	K-12 Core Proteome (yes/no)	K-12 Basic Proteome (yes/no)	
<a href="#">more info</a>	<a href="#">SECE_ECOLI</a>	<a href="#">P0AG96</a>	<i>prfG</i>	b3981 JW3944	Preprotein translocase subunit SecE	Integral Inner Membrane	<b>B</b>	Cell inner membrane   Multi-pass membrane protein	Integral Membrane Protein	IM (predicted), Detected in IM Fraction		SRP (exp)	SEC			yes	yes
<a href="#">more info</a>	<a href="#">SECY_ECOLI</a>	<a href="#">P0AGA2</a>	<i>prfA</i>	b3300 JW3262	Preprotein translocase subunit SecY	Integral Inner Membrane	<b>B</b>	Cell inner membrane   Multi-pass membrane protein	Integral Membrane Protein (Experimental)	IM (predicted), Detected in IM Fraction		SRP	SEC			yes	yes
<a href="#">more info</a>	<a href="#">SECG_ECOLI</a>	<a href="#">P0AG99</a>	<i>secG</i>	b3175 JW3142	Protein-export membrane protein SecG (P12) (Preprotein translocase band 1 subunit)	Integral Inner Membrane	<b>B</b>	Cell inner membrane   Multi-pass membrane protein	Integral Membrane Protein	IM (predicted), Detected in IM Fraction		SRP	SEC			yes	yes

Records 1 to 3 of 3 Page Size

Εικόνα 2.7 – Η ιστοσελίδα της βάσης δεδομένων STEPdb

Στιγμιότυπο από την ιστοσελίδα του STEPdb.

## 2.14 Συζήτηση

Η απαρίθμηση των πρωτεϊνών που ανήκουν σε κάθε υποκυτταρικό διαμέρισμα είναι το πρώτο βήμα για την κατανόηση των φυσικοχημικών, λειτουργικών και δομικών τους χαρακτηριστικών. Επίσης, αποτελεί απαραίτητη προεργασία για οποιαδήποτε πειραματική μελέτη ευρεία κλίμακας (π.χ. πρωτεομικές αναλύσεις) με μελλοντικό στόχο την *in silico* μοντελοποίηση και κατανόηση των κυττάρων.

Επιπλέον, γίνεται ολοένα και πιο έκδηλο ότι αρκετές πρωτεΐνες υποβάλλονται σε δυναμικές αλλαγές τόσο στην θέση όσο και στη λειτουργία τους. Η δυναμική φύση αυτών των πρωτεϊνών ρυθμίζεται από διαφορετικά ερεθίσματα και είναι θεμελιώδης για τις κυτταρικές λειτουργίες. Τέτοιο παράδειγμα αποτελούν κάποιοι μεταγραφικοί παράγοντες που βρίσκονται δεσμευμένοι στην ΠΜ μεμβράνη και κατόπιν ερεθίσματος ελευθερώνονται στο κυτταρόπλασμα για να προσδεθούν στα αντίστοιχα σημεία του DNA/RNA (Ostrovsky de Spicer et al, 1993; Raffaelli et al, 1999).

Η υποκυτταρική τοποθέτηση των πρωτεϊνικών αλληλεπιδράσεων και συμπλόκων, θέτει τα θεμέλια για την απεικόνιση των κυτταρικών λειτουργιών και των μεταβολικών μονοπατιών του κυττάρου. Η επισταμένη ταξινόμηση των πρωτεϊνών στα διάφορα υποκυτταρικά διαμερίσματα αποτελεί για πρώτη φορά τη βάση για μια προκαταρκτική τοποθέτηση των πρωτεϊνικών συμπλόκων με βάση την ταξινόμηση των αντίστοιχων υπομονάδων. Αναφέρουμε ως παράδειγμα μια εφαρμογή της μεθόδου για την πρόσφατη χαρτογράφηση των περιφερικών πρωτεϊνών (ΚΕΦΑΛΑΙΟ 4; (Papanastasiou et al, 2013)). Μέσω συστηματική αναζήτησης των πρωτεϊνικών αλληλεπιδράσεων (Kerrien et al, 2012) και συμπλόκων (Keseler et al, 2013) των περιφερικών πρωτεϊνών οδήγησε στην οργάνωση τους σε εννέα κατηγορίες ανάλογα με την κυτταρική λειτουργία στην οποία συμμετέχουν.

Η ολοκληρωμένη ταξινόμηση σε υποκυτταρικά διαμερίσματα αποτελεί ακρογωνιαίο λίθο για την μετέπειτα ανάλυση οποιουδήποτε κυττάρου. Οι όλο και αυξανόμενες εφαρμογές της πρωτεομικής απαιτεί καλά χαρακτηρισμένα πρωτεϊνώματα. Στην παρούσα μελέτη παρουσιάσαμε μια ολοκληρωμένη συλλογή δεδομένων (STEPdb), η οποία ταξινομεί το σύνολο των πρωτεϊνών του *E.coli* K-12 σε υποκυτταρικές τοποθεσίες.

Προς την κατεύθυνση αυτή συνδυάστηκαν βιοπληροφορικά εργαλεία πρόβλεψης με υπάρχουσες ταξινομήσεις από βάσεις δεδομένων (Dimmer et al, 2012; Horler et al, 2009) οι οποίες επιβεβαιώθηκαν με βιοχημικά, πρωτεϊνωματικά δεδομένα αλλά και εκτεταμένη βιβλιογραφική αναζήτηση. Συνολικά η αναζήτηση αυτή συνεισφέρει 1547 πρωτεΐνες με πειραματικά επιβεβαιωμένες τοποθεσίες στο κύτταρο οι οποίες βασίστηκαν σε 397 δημοσιευμένες μελέτες.

Η συνδυαστική ανάλυση που ακολουθήσαμε συνολικά: αναθεωρεί τις υπάρχουσες ταξινομήσεις για το ~15% των πρωτεϊνών του *E.coli* K-12 (Πίνακας 2.6; 674 of 4303 proteins), ταξινομεί 36 άγνωστες πρωτεΐνες (1%) (Πίνακας 2.6) και επιλύει αντιφάσεις (ανάμεσα σε BE και υπάρχουσες ταξινομήσεις) για 601 από τις 4303 πρωτεΐνες.

Η μεθοδολογία που ακολουθήσαμε απέδειξε ότι τα βιοπληροφορικά εργαλεία δεν είναι επαρκή από μόνα τους να προβλέψουν τον υποκυτταρικό εντοπισμό των πρωτεϊνών του *E.coli* K-12. Αυτό αντανακλάται από τις αντιφάσεις ανάμεσα στις προβλέψεις των μέχρι σήμερα διαθέσιμων βιοπληροφορικών εργαλείων αλλά και από την έλλειψη εξειδικευμένων BE (π.χ. για περιφερικές πρωτεΐνες). Αντίστοιχο παράδειγμα είναι η ασυμφωνία ανάμεσα στα BE Phobius και TMHMM σε σχέση με τον αριθμό των ΔΠ που προβλέπουν. Μάλιστα σε παλαιότερες εκδόσεις των συγκεκριμένων εργαλείων συχνά το ΣΠ προβλεπόταν σαν ΔΠ στο αμινοτελικό άκρο.

Προς την πλήρη υποκυτταρική ταξινόμηση των πρωτεϊνών του *E.coli* η βιβλιογραφική αναζήτηση αλλά και η εκτίμηση πειραματικών δεδομένων αποδείχτηκε αναγκαία. Αυτό ισχύει ιδιαίτερα για την κατηγορία των περιφερικών πρωτεϊνών η οποία δεν έχει μέχρι σήμερα κάποιο ανιχνεύσιμο κοινό χαρακτηριστικό που να περιγράφεται από αντίστοιχα βιοπληροφορικά εργαλεία. Σημαντική για την ταξινόμηση των πρωτεϊνών είναι επίσης και η γνώση του εκκρηκτικού μονοπατιού που ακολουθούν.

Τέλος η βάση δεδομένων STEPdb που δημιουργήσαμε, αναμένεται να αποτελέσει εργαλείο για μια πρώτη αναζήτηση άγνωστων πρωτεϊνών. Η επισταμένη υποκυτταρική ταξινόμηση του *E.coli* K-12 αναμένεται αν αποτελέσει την βάση για εμπεριστατωμένα παραδείγματα πρωτεϊνών που ενδέχεται να οδηγήσουν στην ανάπτυξη νέων βιοπληροφορικών εργαλείων αλλά και στην βελτίωση των ήδη υπαρχόντων. Ενδιαφέρουσα εφαρμογή θα μπορούσε να είναι η πρόβλεψη περιφερικών πρωτεϊνών.



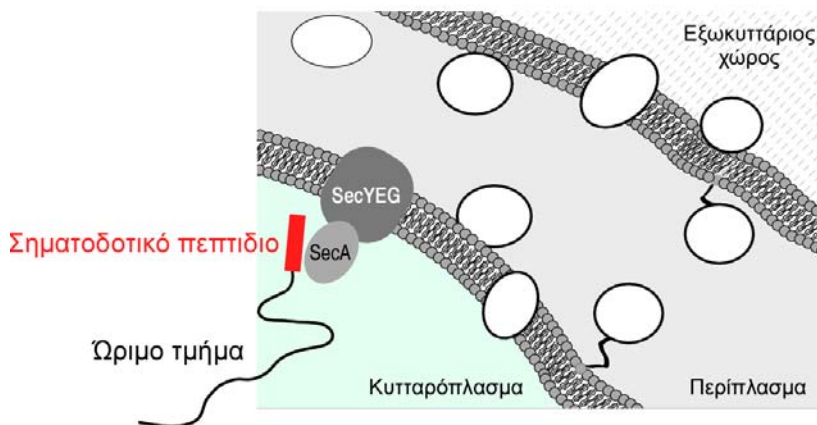
---

### **ΚΕΦΑΛΑΙΟ 3 Αποτελέσματα – Μελέτη των υποστρωμάτων του συστήματος έκκρισης Sec με στόχο τον χαρακτηρισμό νέων σημάτων έκκρισης στην περιοχή του ώριμου τμήματος.**

Πάνω από το 40% των πρωτεϊνών που παράγονται σε ένα οργανισμό εκκρίνονται από το κυτταρόπλασμα και οδηγούνται στον ΚΦ, στην επιφάνεια του κυτάρου αλλά και έξω από αυτό. Στα βακτήρια έχουν αναγνωριστεί περίπου 16 εκκριτικοί μηχανισμοί (Paranikou et al, 2007) με βασικότερο το σύστημα Sec μέσω του οποίου το μεγαλύτερο μέρος των εκκρινόμενων πρωτεϊνών διασχίζουν την πλασματική μεμβράνη (ΠΜ) ή ενσωματώνονται σε αυτήν (Εικόνα 3.1). Οι πρωτεΐνες που ακολουθούν το μονοπάτι Sec συνθέτονται σε μια προ-μορφή στην οποία φέρουν μια αμινοτελική προέκταση γνωστή ως ΣΠ (Blobel et al, 1975; Dalbey et al, 2012; Hegde et al, 2006).

Τα ΣΠ επιτρέπουν την αναγνώριση αλλά και την στόχευση των εκκρινόμενων πρωτεϊνών στο διαμεμβρανικό κανάλι το οποίο διεξάγει την διαδικασία της έκκρισης (translocase). Ο καθολικός παράγοντας SRP (Signal Recognition Particle) αναγνωρίζει ορισμένα ΣΠ καθώς προκύπτουν από τα ριβοσώματα κλητεύοντας την εισαγωγή των προ-πρωτεϊνών στην μεμβράνη (Dalbey et al, 2012). Μετά την έκκριση τους τα ΣΠ αποκόπτονται από εξειδικευμένες πεπτιδάσες στην επιφάνεια της ΠΜ (Economidou, 1999). Το εναπομείναν κομμάτι της πρωτεΐνης, το οποίο ονομάζεται ΩΤ, απελευθερώνεται από την ΠΜ ώστε να αναδιπλωθεί στην τελική του μορφή.

Τα ΣΠ δεν αποτελούν συντηρημένα μοτίβα παρουσιάζουν όμως κοινές φυσικοχημικές ιδιότητες (Paetzel et al, 2002). Αυτές συνοψίζονται σε τρεις περιοχές: 1) το θετικά φορτισμένο αμινοτελικό άκρο (n-domain), 2) ένα κεντρικό υδρόφοβο τμήμα (h-domain) και γ) ένα υδρόφιλο καρβοξυτελικό άκρο (c-domain) το οποίο περιέχει συντηρημένα μοτίβα που αναγνωρίζουν οι πεπτιδάσες (SPases) (Paetzel et al, 2002) (Ενότητα 1.1.1, Εικόνα 1.5).



**Εικόνα 3.1 – Σχηματική απεικόνιση της τομής ενός Gram<sup>-</sup> κυτάρου – Sec σύστημα έκκρισης.**

Τομή ενός Gram<sup>-</sup> βακτηρίου το οποίο αποτελείται από το κυτταρόπλασμα το οποίο περικλείεται από δύο βιολογικές μεμβράνες την ΠΜ και την ΕΜ. Στη συγκεκριμένη απεικόνιση διακρίνονται τα διάφορα υποκυτταρικά διαμερίσματα στα οποία μπορεί να εκκριθεί μια πρωτεΐνη. Το μεγαλύτερο μέρος των πρωτεϊνών χρησιμοποιούν το βασικό σύστημα έκκρισης Sec το οποίο βρίσκεται στην ΠΜ και αποτελείται από το μεμβρανικό σύμπλοκο SecYEG και τον κινητήρα SecY. Οι πρωτεΐνες που αναγνωρίζονται από το σύστημα Sec αποτελούνται από μια σηματοδοτικό πεπτιδίο (ΣΠ) και από το ώριμο τμήμα (ΩΤ).

Τα συντηρημένα χαρακτηριστικά του ΣΠ έχουν επιτρέψει την ανάπτυξη αξιόπιστων βιοπληροφορικών εργαλείων τα οποία μπορούν και προβλέπουν τις θέσεις αποκοπής (cleavage sites) (Juncker et al, 2003; Kall et al, 2007; Petersen et al, 2011) ενώ άλλα προβλέπουν την τελική θέση μέσα στο κύτταρο (Gardy et al, 2003; Imai et al, 2008). Μερικά από αυτά τα βιοπληροφορικά εργαλεία αξιολογούν χαρακτηριστικά όπως η δευτεροταγής δομή, το συνολικό φορτίο, αριθμό διαδοχικών φορτισμένων κατάλοιπα και υδροφοβικότητα.

Πρόσφατα στοιχειοθετήθηκε ο ρόλος των ΩΤ στην στόχευση των πρωτεϊνών ακόμα και σε συνθήκες απουσίας του ΣΠ (Gouridis et al, 2009). Η παρατήρηση αυτή οδήγησε στο συμπέρασμα ότι ίσως υπάρχουν άγνωστα χαρακτηριστικά στα ώριμα τμήματα των προ-πρωτεϊνών τα οποία μεσολαβούν στόχευση τους στην μεμβράνη (Gouridis et al, 2009; Kajava et al, 2000; Kim et al, 2000; Li et al, 1988; Safdar et al, 2010; Summers et al, 1989).

Το πρώιμο ΩΤ έχει βρεθεί ότι είναι εξαιρετικά σημαντικό καθώς μεταλλάξεις στην συγκεκριμένη περιοχή, όπως η εισαγωγή θετικών φορτίων, μπορεί να παρεμποδίζει την έκκριση (Kajava et al, 2001; Kato et al, 1992; Kim et al, 2000; MacIntyre et al, 1990; Summers et al, 1989). Μοριακοί οδηγοί όπως η SecB (Bassilana et al, 1992; Fekkes et al, 1999; Khokhlova et al, 2003;



Kim et al, 2000) και η SecA (Baud et al, 2002; Lill et al, 1990) εμπλέκονται στην αναγνώριση των ΩΤ όμως η μοριακή βάση της αλληλεπίδρασης παραμένει εντελώς άγνωστη.

Στατιστική ανάλυση που έγινε σχετικά με την αμινοξική σύσταση του πρώιμου ΩΤ προτείνει ότι τα Gram- βακτήρια τείνουν να διατηρούν αρνητικό καθαρό φορτίο (Choo et al, 2008a; Kajava et al, 2001). Μεταλλάξεις όπου βασικά κατάλοιπα (Αργινίνη και Λυσίνη) προστέθηκαν σε διάφορα σημεία στις πρώτες 30 θέσεις του ΩΤ της αλκαλικής φωσφατάσης μείωσαν την απόδοση της έκκρισης (Kajava et al, 2001). Γονιδιακά ανασυνδυασμένες κυτταροπλασματικές πρωτεΐνες που δεν μπορούσαν να εκκριθούν, παρά την παρουσία ΣΠ, οδηγήθηκαν σε έκκριση έπειτα από αντικατάσταση των βασικών κατάλοιπων που βρίσκοντας στις δύο πρώτες θέσεις του ΩΤ με όξινα (Ασπαρτικό και Γλουταμινικό οξύ) (Tian et al, 2009)

Σε αντίθεση με τα ΣΠ τα οποία μπορούν να προβλεφθούν με μεγάλη ακρίβεια, τα ΩΤ των πρόδρομων πρωτεϊνών στερούνται κάποιας προφανής χαρακτηριστικής αλληλουχίας και κατά συνέπεια δεν έχει αναπτυχθεί μέχρι σήμερα κάποιο βιοπληροφορικό εργαλείο για την πρόβλεψη τους. Στις ενότητες που ακολουθούν θα παρουσιάσουμε τεκμήρια που αντικρούουν την συγκεκριμένη θεώρηση. Η ανάλυση μας οδηγείται στο συμπέρασμα ότι οι ώριμες μορφές των εκκρινόμενων πρωτεϊνών έχουν χαρακτηριστικά τα οποία τις καθιστούν διαχωρίσιμες από τις κυτταροπλασματικές. Με την χρήση μεθόδων μηχανικής μάθησης αναπτύξαμε νέα μοντέλα τα οποία προβλέπουν με υψηλό ποσοστό επιτυχίας τις εκκρινόμενες από το Sec σύστημα πρωτεΐνες. Τα μοντέλα εκπαιδεύτηκαν απουσία του ΣΠ και βασίστηκαν αποκλειστικά στην αλληλουχία των ΩΤ. Με αυτόν τον τρόπο καταφέραμε να απομονώσουμε την πληροφορία που εμπεριέχεται στα ΩΤ και να αναδείξουμε νέα χαρακτηριστικά.

Η προσέγγιση μας βασίστηκε στην ανάπτυξη μοντέλων διαχωρισμού τα οποία εκπαιδεύτηκαν με το βιοπληροφορικό πακέτο GEMS (Statnikov et al, 2005), ένα εργαλείο ταξινόμησης που βασίζεται σε μηχανές διανυσμάτων υποστήριξης (SVMs: Support Vector Machines) και ενσωματώνει τεχνικές ενδοπιστοποίησης (cross-validation) και επιλογής χαρακτηριστικών.

Για την πρόβλεψη των ΩΤ στην διαδικασία της εκπαίδευσης λάβαμε υπόψη πληροφορία χαμηλού επιπέδου όπως η πρωτοταγής αμινοξική αλληλουχία, σύσταση σε διπεπτίδια και τριπεπτίδια και φυσικοχημικές ιδιότητες των πλευρικών αλυσίδων των αμινοξέων, αντιθέτως δεν συμπεριλάβαμε καμία πληροφορία που μαρτυρά την ύπαρξη ΣΠ.

Τα χαρακτηριστικά που επιλέχθηκαν κατά την διαδικασία της εκπαίδευσης υποδηλώνουν ότι στις αλληλουχίες υπήρξε εξελικτική πίεση η οποία οδήγησε στην διαφοροποίηση τους από αυτές των κυτταροπλασματικών. Η διαφορετική αμινοξική σύσταση των ώριμων εκκρινόμενων πρωτεϊνών ενδεχομένως αντανακλά το αποτέλεσμα της εξισορρόπησης δύο ταυτόχρονων εξελικτικών διαδικασιών. Πρώτον της βελτιστοποίησης της έκκρισης με επιλογή αμινοξέων που διατηρούν την αλληλουχία σε μια σχετικά «ξεδίπλωτη» κατάσταση προκειμένου να αναγνωρισθεί από την μεταθετάση Sec. Δεύτερον την διατήρηση της ικανότητας τους να αναδιπλωθούν στην τελική τους λειτουργική δομή. Αυτό θα εξηγούσε γιατί τα χαρακτηριστικά που παρατηρούμε είναι διάσπαρτα και δεν είναι ευδιάκριτα και μεγάλωφωνα. Οι αλλαγές αυτές αντανακλώνται μέσα από την διαφορετική αμινοξική σύσταση των πρωτεϊνών την οποία στη συνέχεια θα προσπαθήσουμε συσχετίσουμε με την πιθανότητα αταξίας των αλληλουχιών.

### 3.1 Επιλογή δεδομένων

Η ανάλυση μας εστιάζεται στο πρωτεϊνώμα του Gram- βακτηρίου *Escherichia coli*, έναν από τους βασικούς οργανισμούς που έχει χρησιμοποιηθεί σαν μοντέλο για την μελέτη των μηχανισμών έκκρισης. Εκκρινόμενες πρωτεΐνες του συστήματος Sec, από τρία διαφορετικά υποκυτταρικά διαμερίσματα συμπεριλήφθηκαν στην ανάλυση: 1) περιπλασματικές (G), 2) πρωτεΐνες της εξωτερικής μεμβράνης (H) και 3) λιποπρωτεΐνες της εσωτερικής/εξωτερικής μεμβράνης (E,I) (Εικόνα 2.1 και Πίνακας 6.7). Τα σύνολα αυτά προέκυψαν από την υποκυτταρική ταξινόμηση του συνολικού πρωτεϊνώματος του βακτηρίου (βάση δεδομένων STEPdb) όπου και επιλέξαμε μόνο όσες πρωτεΐνες εκκρίνονται από το σύστημα Sec και έχουν ΣΠ. Οι πρωτεΐνες που εκκρίνονται από το σύστημα Tat δεν συμπεριλήφθηκαν στην ανάλυση παρόλο που κάποιες από αυτές φαίνεται να μπορούν εναλλακτικά να εκκριθούν από το Sec σύστημα (Tullman-Ercek et al, 2007). Πρωτεΐνες που ακολουθούν το μονοπάτι Ssr αποκλείστηκαν και αυτές από την ανάλυση καθώς φέρουν ΣΠ τα οποία έχουν έκτεταμένα υδρόφοβα τμήματα (h-domain).

### 3.2 Συντηρημένα αμινοξέα στις αλληλουχίες των εκκρινόμενων πρωτεϊνών, σύγκριση με κυτταροπλασματικές.

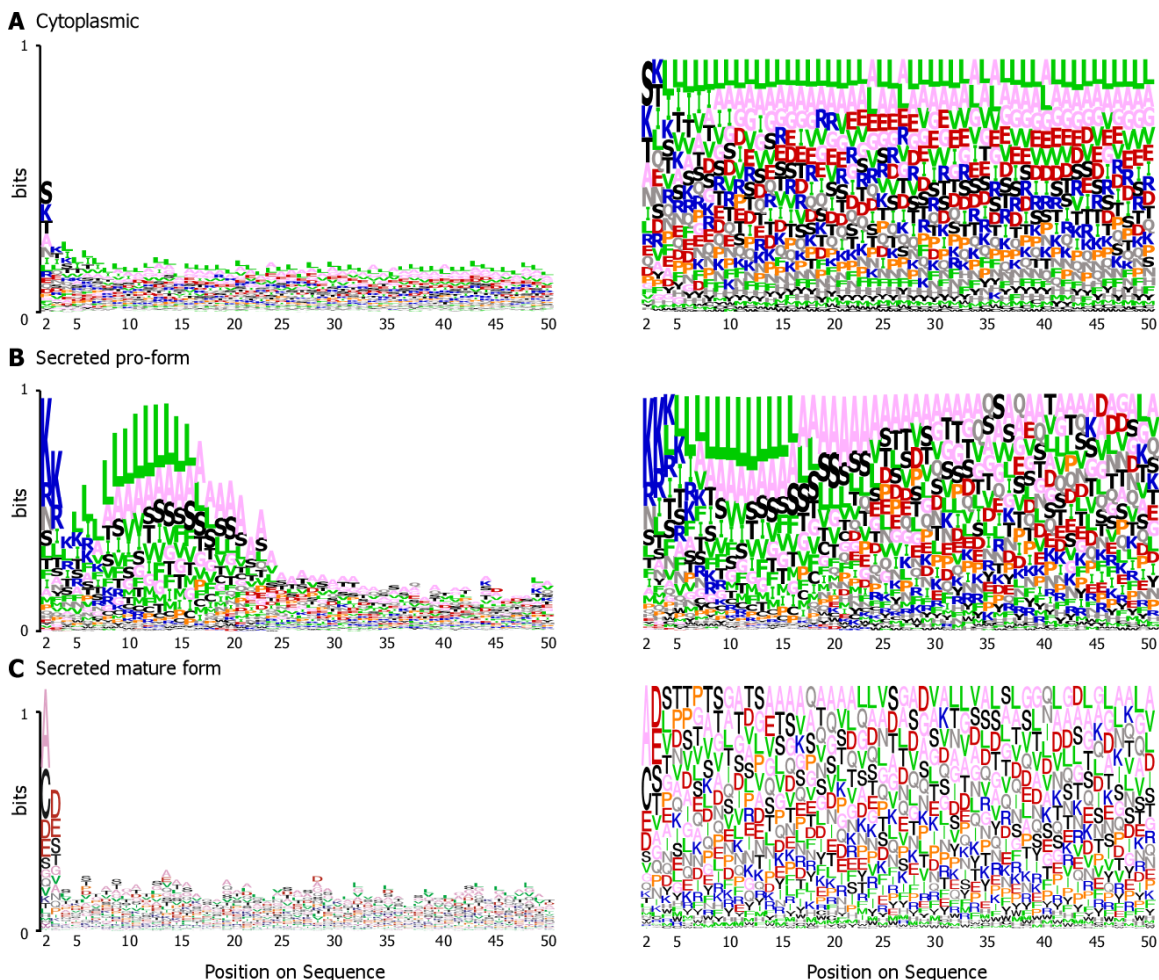
Αναλύσαμε της αλληλουχίες των εκκρινόμενων πρωτεϊνών ως προς την αμινοξική τους σύσταση ανά θέση χρησιμοποιώντας το WebLogo (Crooks et al, 2004). Το εργαλείο αυτό

σχεδιάζει γραφήματα που αναδεικνύουν συντηρημένα μοτίβα σε αλληλουχίες. Οι εκκρινόμενες πρωτεΐνες στοιχήθηκαν στην αρχή του ΣΠ αλλά και στην αρχή του ΩΤ και συγκριθήκαν με τις κυτταροπλασματικές (Εικόνα 3.2).

Η ανάλυση ανέδειξε τις διαφορές στη σύσταση των αμινοξέων κάθε κατηγορίας πρωτεϊνών. Οι κυτταροπλασματικές πρωτεΐνες δεν παρουσιάζουν συντηρημένα κατάλοιπα εκτός από Σερίνες (S) και Λυσίνες (K) στην δεύτερη θέση (μετά την εναρκτήρια Μεθιονίνη η οποία δεν απεικονίζεται), ενώ κατά μήκος της υπόλοιπης αλληλουχίας φαίνεται να παρουσιάζουν μεγαλύτερη συχνότητα μικρά κατάλοιπα όπως η Αλανίνη και την Γλυκίνη (A,G) και το υδρόφοβο κατάλοιπο της Λυσίνης (L).

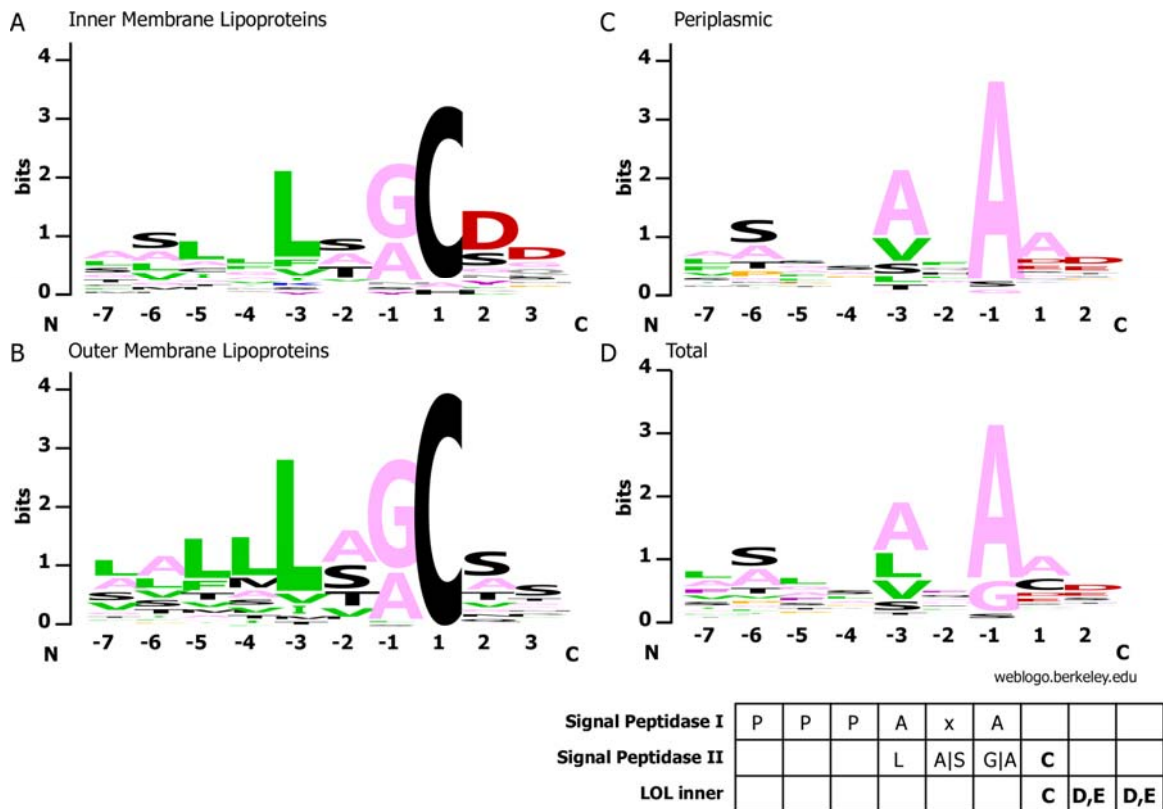
Η στοιχίση των πρόδρομων μορφών (το ΣΠ συμπεριλαμβάνεται) αποδεικνύει ότι τα ΣΠ παρουσιάζουν συντηρημένες φυσικοχημικές ιδιότητες παρόλο που δεν είναι συγκεκριμένα αμινοξέα συντηρημένα σε συγκεκριμένες θέσεις (Nesmeyanova et al, 1997). Οι περιοχές του πεπτιδίου σήματος (n,h,c-domain) είναι ευδιάκριτες παρόλο που είναι προφανές ότι δεν είναι απόλυτα στοιχισμένες πιθανώς λόγω μεγάλης ποικιλίας στο μήκος των ΣΠ και κατ' επέκταση στο μήκος των αντίστοιχων περιοχών (1.5 - Χαρακτηριστικά εκκρινόμενων πρωτεϊνών).

Διαφορετική εικόνα παρουσιάζουν τα ΩΤ των εκκρινόμενων πρωτεϊνών με αρνητικά φορτισμένα αμινοξέα στις δύο πρώτες θέσεις (Ασπαρτικό και Γλουταμινικό οξύ) ενώ η ακόλουθη θέση φαίνεται να αποτελούν τυχαία επιλογή αμινοξέων. Παρατηρήσαμε ότι στη θέση +1 επιλέγονται με μεγαλύτερη συχνότητα τα αμινοξέα της Αλανίνης (A) και της Κυστεΐνης (C). Τα κατάλοιπα αυτά είναι γνωστό ότι αποτελούν μέρος καλά χαρακτηρισμένων μοτίβων που αναγνωρίζονται από τις εξειδικευμένες πεπτιδάσες τύπου I και II (μοτίβο -3 -1 και lipobox αντίστοιχα). Στη συνέχεια διερευνήσαμε την πιθανή προέλευση των αρνητικά φορτισμένων αμινοξέων στις δύο πρώτες θέσεις των ώριμων τμημάτων (Εικόνα 3.3). Αποδώσαμε το αρνητικό φορτίο του πρώιμου ώριμου τμήματος στο υποσύνολο των λιποπρωτεϊνών. Υπάρχουν δύο τύποι λιποπρωτεϊνών αυτές που αφού εκκριθούν παραμένουν αγκυροβολημένες στην ΠΜ ενώ άλλες ακολουθούν το μονοπάτι LolA για να αγκυροβοληθούν στην ΕΜ. Είναι γνωστό ότι στην πρώτη περίπτωση η πρωτεΐνη φέρουν αρνητικά φορτισμένα κατάλοιπα στις θέσεις +2 και +3 τα οποία σημαίνουν την διαφυγή από το μονοπάτι LolA.



**Εικόνα 3.2 - Συντηρημένα αμινοξικά χαρακτηριστικά των εκκρινόμενων και κυτταροπλασματικών πρωτεϊνών**

Γραφήματα (logo graphs) της εντροπίας της πληροφορία (αριστερά) και της συχνότητας των αμινοξέων (δεξιά) για τις αλληλουχίες των: A) κυτταροπλασματικών, B) προδρομων μορφών και C) ώριμων μορφών. Τα γραφήματα των προδρομων αναδεικνύουν τις συντηρημένες φυσικοχημικές ιδιότητες των ΣΠ παρόλο που τα αντίστοιχα αμινοξέα δεν είναι πλήρως συντηρημένα ανά θέση. Οι δύο πρώτες θέσεις των ώριμων τμημάτων παρουσιάζουν προτίμηση σε αρνητικά φορτισμένα αμινοξέα (Ασπαρτικό και Γλουταμινικό οξύ). Η δεύτερη θέση των κυτταροπλασματικών πρωτεϊνών εμφανίζει την μεγαλύτερη συντήρηση σε Σερίνες και Λυσίνες ενώ οι υπόλοιπες θέσεις δεν είναι συντηρημένες αλλά παρουσιάζουν μια μεγαλύτερη συχνότητα σε υδρόφοβα/μικρά αμινοξέα (Λευκίνη, Αλανίνη, Γλουταμίνη).



Εικόνα 3.3 – Συντηρημένα μοτίβα που αναγνωρίζονται από τις πεπτιδάσες τύπου I και II.

Στοιχισμός των αμινοξικών αλληλουχιών στην περιοχή του σημείου αποκοπής του ΣΠ από τις πεπτιδάσες τύπου I (SPaseI) και II (SPaseII) για τις: A) λιποπρωτεΐνες της πλασματική μεμβράνης, B) λιποπρωτεΐνες της εξωτερικής μεμβράνης C) Περιπλασμικές D) εκκρινόμενες πρωτεΐνες. Ο πίνακας συνοψίζει τα μοτίβα που αναγνωρίζονται από τις αντίστοιχες πεπτιδάσες. Οι λιποπρωτεΐνες αποκόπτονται από την πεπτιδάση τύπου II η οποία αναγνωρίζει ένα κατάλοιπο κυστεΐνης στην θέση +1 του ΩΤ ενώ οι περιπλασμικές πρωτεΐνες αναγνωρίζονται αντίστοιχα από την πεπτιδάση τύπου I η οποία είναι γνωστό ότι αναγνωρίζει το μοτίβο -1 -3 (Αλανίνες στις θέσεις -1,-3 ενώ είναι αδιάφορο το κατάλοιπο στην θέση -2). Οι λιποπρωτεΐνες της ΠΜ αφού εκκριθούν παραμένουν αγκυροβολημένες στην ΠΜ σε αντίθεση με τις λιποπρωτεΐνες της ΕΜ οι οποίες ακολουθούν το μονοπάτι LolA και αγκυροβολούνται στην ΕΜ. Είναι γνωστό ότι στην πρώτη περίπτωση η πρωτεΐνες φέρουν αρνητικά φορτισμένα κατάλοιπα στις θέσεις +2 και +3 τα οποία σημαίνουν την διαφυγή από το μονοπάτι LolA (Paetzel et al, 2002).

### 3.3 Επιλογή της ελάχιστης αλληλουχίας προς ανάλυση

Αρχικά προσδιορίσαμε το ελάχιστο μήκος του ΩΤ που περιέχει όλη την πληροφορία σχετιζόμενη με έκκριση. Για το λόγο αυτό αναλύσαμε βιοχημικά δεδομένα και τα συνδυάσαμε με παρατήρηση της εξελικτικής διαδικασίας. Πειράματα περικοπής του ΩΤ υποδεικνύουν ότι ώριμα

τμήματα με μήκος τουλάχιστον ~50 αμινοξέων διατηρούν την εκκριτική τους ικανότητα (Bassilana et al, 1992; Hemm et al, 2008b; Kato et al, 1992) κατά συνέπεια και την ελάχιστη απαραίτητη πληροφορία για την διαδικασία της έκκρισης.

Η θεωρητική ανάλυση του μήκους των ΩΤ των εκκριτικών πρωτεϊνών του *E.coli* έδειξε ότι σε 23 πρωτεΐνες τα ΩΤ είναι μήκους μικρότερου των 70 αμινοξέων. (Πίνακας 6.3). Οι λιποπρωτεΐνες EcpA και EcpB έχουν τα πιο μικρά ΩΤ μήκους 23 και 27 αμινοξέα αντίστοιχα. Ενόψει αυτών των παρατηρήσεων επιλέξαμε να αναλύσουμε τα πρώτα 100 αμινοξέα των πρόδρομων μορφών, δηλαδή ΩΤ μήκους από 68 έως 80 αμινοξέα.

### **3.4 Εκπαίδευση μοντέλων διαχωρισμού χρησιμοποιώντας το GEMS σαν εργαλείο μηχανικής μάθησης**

Προς την ανάδειξη νέων σημάτων έκκρισης στα ώριμα τμήματα των εκκρινόμενων πρωτεϊνών χρησιμοποιήσαμε μεθόδους μηχανικής μάθησης για να διαχωρίσουμε τις δύο κατηγορίες πρωτεϊνών, εκκρινόμενες και κυτταροπλασματικές.

Η εκπαίδευση των μοντέλων διαχωρισμού έγιναν με το GEMS (Gene Expression Model Selector) σαν βασικό εργαλείο μηχανικής μάθησης. Το εργαλείο αυτό αναπτύχθηκε αρχικά για την ταξινόμηση δεδομένων από πειράματα μικρο-συστοιχιών (Statnikov et al, 2005). Αποτελεί ένα ολοκληρωμένο βιοπληροφορικό πακέτο το οποίο περιλαμβάνει αλγόριθμους για την επιλογή χαρακτηριστικών, ενδοπιστοποίηση (cross-validation) και υπολογισμού της απόδοσης (για τεχνικές λεπτομέρειες δες ενότητα 6.2).

Έχοντας ως στόχο να αναδείξουμε νέα σήματα στις αλληλουχίες των ΩΤ αποφασίσαμε να συμπεριλάβουμε στην εκπαίδευση τα αμινοξέα από την θέση +3 έως την +102, καθότι οι δύο πρώτες θέσεις αφορούν μοτίβα που δεν σχετίζονται με την διαδικασία της έκκρισης (δες ενότητα 3.2). Επιπλέον η εναρκτήρια Μεθιονίνη παραλείπεται από την ανάλυση (πρόδρομες και κυτταροπλασματικές πρωτεΐνες).

### **3.5 Μεταβλητές Εκπαίδευσης**

Εκπαιδεύσαμε μοντέλα χρησιμοποιώντας διαφορετικές μεταβλητές στην εκπαίδευση με στόχο να αναδείξουμε να πιθανά χαρακτηριστικά των εκκρινόμενων πρωτεϊνών. Συγκεκριμένα χρησιμοποιήσαμε την α) αμινοξική αλληλουχία των πρωτεϊνών (δες 6.2.1) β) σύσταση σε

---

διπεπτίδια γ) σύσταση σε τριπεπτίδια γ) ψευδο-αμινοξική σύσταση (δες 6.2.8 Ψεύδο-αμινοξική σύσταση (Pseudo amino-acid Composition)).

Για την εκπαίδευση μοντέλων διαχωρισμού που βασίζονται στις πρωτεϊνικές αλληλουχίες εφαρμόσαμε μια αριθμητική αναπαράσταση των αμινοξέων. Συνοπτικά κάθε αμινοξύ κωδικοποιήθηκε ως ένας μοναδικός δυαδικός αριθμός (Πίνακας 6.4 – Απλή αναπαράσταση των αμινοξέων) ενώ κάθε αμινοξική αλληλουχία ως μια αλληλουχία δυαδικών αριθμών.

Στη συνέχεια αποφασίσαμε να διερευνήσουμε την σημασία των φυσικοχημικών ιδιοτήτων των αμινοξέων. Ομαδοποιήσαμε τα αμινοξέα με βάση τις ιδιότητες αυτές όπως για παράδειγμα το Ασπαρτικό και Γλουταμινικό οξύ ως αρνητικά φορτισμένα κατάλοιπα. Οι φυσικοχημικές ιδιότητες των αμινοξέων είναι αλληλεπικαλύπτουσες, για παράδειγμα το κατάλοιπο της Φαινυλαλανίνη μπορεί να θεωρηθεί αρωματικό (περιέχει αρωματικό δακτύλιο) αλλά και υδρόφοβο. Για αναπαραστήσουμε μερικές από τις πολλαπλές ιδιότητες των αμινοξέων δοκιμάσαμε δύο εναλλακτικές ομαδοποιήσεις σε 9 (συμπαγής αναπαράσταση; Πίνακας 6.5) και σε 11 ιδιότητες (χαλαρή αναπαράσταση; Πίνακας 6.6). Μερικές διαφορές στις δύο ομαδοποιήσεις είναι για παράδειγμα η ομαδοποίηση των θετικά (Λυσίνη, Αργινίνη) ή των αρνητικά φορτισμένων αμινοξέων (Ασπαρτικό και Γλουταμινικό οξύ) στην πρώτη περίπτωση έναντι της δεύτερης. Μια άλλη διαφοροποίηση είναι η ομαδοποίηση των αρωματικών αμινοξέων (Φαινυλαλανίνη, Τυροσίνη και Τρυπτοφάνη) στην «συμπαγή» αναπαράσταση ενώ στη «χαλαρή» αναπαράσταση η Φαινυλαλανίνη λαμβάνεται υπόψη ως υδρόφοβο κατάλοιπο.

### 3.5.1. Μοντέλα πρόδρομης και ώριμης μορφής

Χρησιμοποιήσαμε την κωδικοποίηση των αμινοξικών αλληλουχιών (μεταβλητές εκπαίδευσης) για να εκπαιδεύσουμε μοντέλα που διαχωρίζουν: α) τις πρόδρομες μορφές β) τα ώριμα τμήματα των εκκρινόμενων πρωτεϊνών από τις κυτταροπλασματικές.

Δοκιμάσαμε να διαχωρίσουμε τις πρόδρομες μορφές από τις κυτταροπλασματικές, χρησιμοποιώντας την απλή αλλά και την ομαδοποιημένη αναπαράσταση των αμινοξέων (χαλαρή και συμπαγής; Πίνακας 6.5 και 6.6) και η απόδοση των μοντέλων στην δεύτερη περίπτωση ήταν μεγαλύτερη (~97% έναντι ~99%) ενώ ο αριθμός των επιλεγμένων χαρακτηριστικών μικρότερος (134 έναντι 131; δεδομένα δεν παρουσιάζονται).

Συλλογικά τα δεδομένα αυτά αποδεικνύουν ότι η ελάχιστη απαραίτητη πληροφορία για τον διαχωρισμό των δύο κατηγοριών εμπεριέχεται στην πρωτοταγή αλληλουχία των πρωτεϊνών ενώ είναι ενδεικτικό ότι μπορεί να συνοψιστεί στις φυσικοχημικές ιδιότητες των αμινοξέων που τις απαρτίζουν (Εικόνα 3.4).

Σχεδόν σε όλες τις περιπτώσεις προέκυψαν γραμμικά μοντέλα δηλαδή η συνάρτηση διαχωρισμού είναι ένα πολυώνυμο πρώτου βαθμού το οποίο γεωμετρικά αντιστοιχεί σε ένα υπερεπίπεδο στο N-διάστατο χώρο. Η παρατήρηση αυτή υποδηλώνει γραμμική ανεξαρτησία των χαρακτηριστικών δηλαδή η επιλογή συγκεκριμένου αμινοξέος σε μια θέση δεν εξαρτάται από κάποια άλλη.

Η σύγκριση δύο αντίστοιχων μοντέλων διαχωρισμού τις πρόδρομης και ώριμης μορφής των εκκρινόμενων πρωτεϊνών προβάλλει κάποια πιθανά χαρακτηριστικά των ώριμων τμημάτων (Εικόνα 3.4). Στο μοντέλο πρόδρομης τα χαρακτηριστικά με την μεγαλύτερη βαρύτητα επιλέγονται στην περιοχή του ΣΠ. Τα χαρακτηριστικά αυτά βρίσκονται σε συμφωνία με τις ιδιότητες των τριών περιοχών από τα οποία αποτελείται ένα ΣΠ (Paetzel et al, 2002) και μπορούν να συνοψιστούν σε: θετικά φορτισμένα αμινοξέα στις θέσεις 3-5 (n-domain), υδρόφοβα στις θέσεις 4-15 (h-domain) και Αλανίνη, Κυστεΐνη, Σερίνη στις θέσεις 15-23 (c-domain).

Λιγότερο «ηχηρά» χαρακτηριστικά επιλέγονται στην περιοχή του ΩΤ τα οποία συνοψίζονται πολικά κατάλοιπα όπως Σερίνη (S), Γλουταμίνη (Q) και Θρεονίνη (T). Επιλέγονται επίσης εξίσου σημαντικά χαρακτηριστικά με αρνητική βαρύτητα όπως υδρόφοβα αμινοξέα και Αργινίνη στο πρώιμο ΩΤ (θέσεις 1-33 κατά προσέγγιση). Θετικά φορτισμένα αμινοξέα στην περιοχή αυτή έχουν παρατηρηθεί και πειραματικά ότι παρεμποδίζουν την έκκριση όταν εισαχθούν σε σειρά (Kajava et al, 2000; Li et al, 1988; MacIntyre et al, 1990) ενώ αντίστροφα η αντικατάστασή τους από όξινα κατάλοιπα φαίνεται να αποκαθιστά στην έκκριση (Summers et al, 1989). Τα χαρακτηριστικά αυτά αναδεικνύονται περισσότερο όταν το ΣΠ είναι απών (μοντέλο ώριμης μορφής; Εικόνα 3.4B).

Το πρώιμο ΩΤ των Gram έχει προκύψει αρνητικά φορτισμένο από μια προγενέστερη στατιστική ανάλυση (Kajava et al, 2000) ενώ έχει σχετιστεί και πειραματικά με την διαδικασία της έκκρισης. Με βάση την δική μας ανάλυση και βιβλιογραφική αναζήτηση, το αρνητικό φορτίο στο αμινοτελικό άκρο των ΩΤ οφείλεται σε σήματα στόχευσης που φέρουν οι λιποπρωτεΐνες (αρνητικά φορτισμένα αμινοξέα στις θέσεις +1 και +2; δες Ενότητα 3.2). Έχοντας αφαιρέσει τις θέσεις αυτές



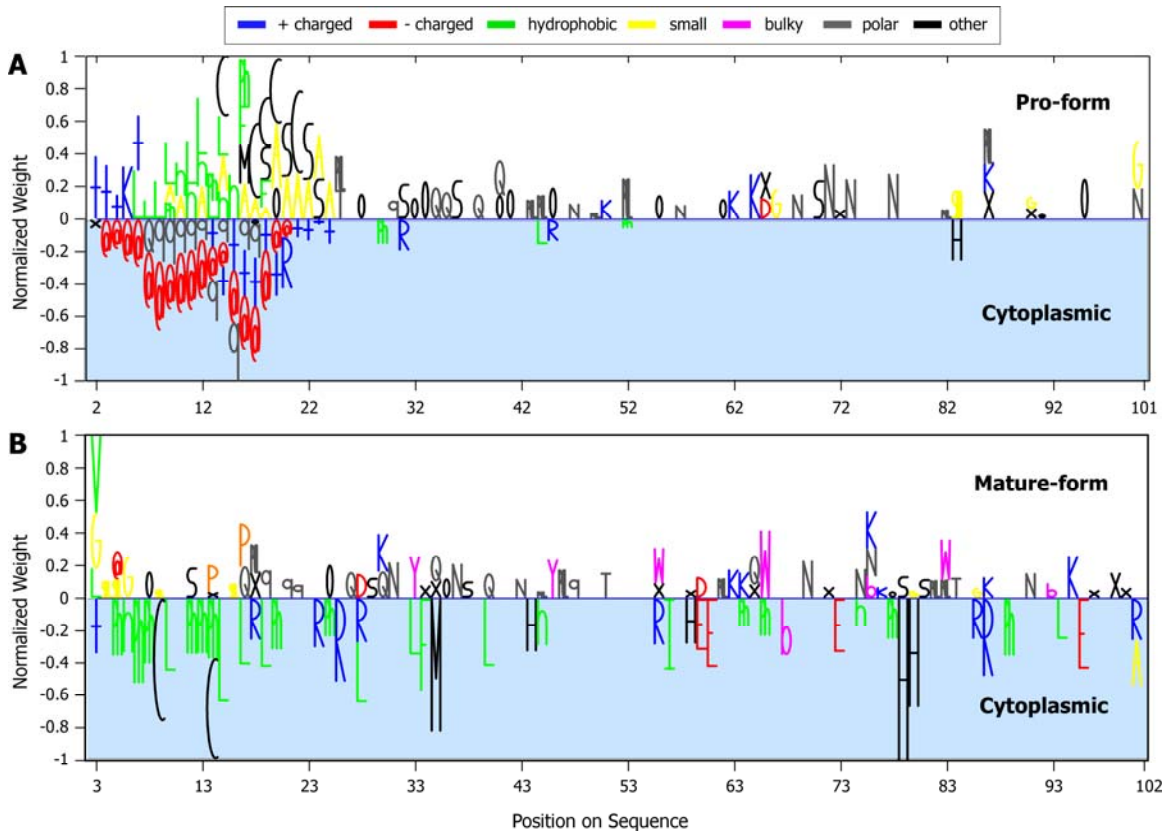
από την διαδικασία της εκπαίδευσης καταφέραμε να αναδείξουμε λιγότερο «ηχηρά» χαρακτηριστικά όπως η προτίμηση σε συγκεκριμένα θετικά και αρνητικά φορτισμένα αμινοξέα σε κάθε κλάση (Αργινίνη και Γλουταμινικό οξύ [RE] στις κυτταροπλασματικές, Λυσίνη και Ασπαρτικό οξύ [KD] στα ώριμα τμήματα).

Προς ενίσχυσης της παρατήρησης, μοντέλα που εκπαιδεύτηκαν με την συμπαγή αναπαράσταση (9 χαρακτηριστικά; Πίνακας 6.5) όπου τα δύο αυτά ζεύγη αμινοξέων ήταν ομαδοποιημένα είχε απόδοση 10% χαμηλότερη σε σχέση με την χαλαρή αναπαράσταση (11 χαρακτηριστικά; Πίνακας 6.6) όπου τα αμινοξέα αυτά αναπαρίστανται ξεχωριστά. Τέλος ειδικά επιλεγμένα χαρακτηριστικά μοιάζουν να είναι Προλίνες (P) στις 14-18 και τα αρωματικά αμινοξέα Τρυπτοφάνη (W) και Τυροσίνη (Y).

**Πίνακας 3.1 – Σύγκριση συμπαγούς και χαλαρής ομαδοποίησης αμινοξέων σε μοντέλα διαχωρισμού ώριμων τμημάτων από κυτταροπλασματικές πρωτεΐνες**

A/A	Ομαδοποίηση Αμινοξέων	Σύνολο Εκπαίδευσης	Σύνολο Αξιολόγησης	Πειραματικά Δεδομένα	Τάξη συνάρτησης πυρήνα
1	<i>συμπαγής (Πίνακας 7.4)</i>	84.93	66.63	52.05	1
2	<i>χαλαρή (Πίνακας 7.5)</i>	99.79	82.03	90.29	2

Συνοψίζοντας τα χαρακτηριστικά που επιλέγονται στην περιοχή του ΩΤ δεν φαίνεται να ομαδοποιούνται σε περιοχές με καλά συντηρημένα αμινοξέα ή ιδιότητες αμινοξέων ανά θέση όπως στην περίπτωση του πεπτιδίου σήματος (Εικόνα 3.4A) αλλά είναι διάσπαρτα. Το μοναδικό τμήμα με ιδιαίτερα ξεχωριστές ιδιότητες αποτελεί το πρώιμο ΩΤ (1-33) το οποίο φαίνεται να είναι απαγορευτικό σε κατάλοιπα Αργινίνης και σε υδρόφοβα αμινοξέα (L,I,F). Παρά το γεγονός ότι τα επιλεγμένα χαρακτηριστικά δεν συνοψίζονται σε αναγνωρίσιμα μοτίβα, η προτίμηση σε συγκεκριμένα αμινοξέα, διαφορετικά ανά κατηγορία, αντανακλά την διαφορετική σύσταση των πρωτεϊνών.



---

### Εικόνα 3.4- Απεικόνιση μοντέλων πρόδρομης και ώριμης μορφής

Απεικονίσαμε τα πιο σημαντικά χαρακτηριστικά, που επιλέχθηκαν, για τον διαχωρισμό των: Α) πρόδρομων και Β) ώριμων μορφών των εκκρινόμενων πρωτεϊνών από τις κυτταροπλασματικές πρωτεΐνες.

Τα δύο μοντέλα εκπαιδεύτηκαν χρησιμοποιώντας μόνο την πληροφορία της αμινοξικής αλληλουχίας των πρωτεϊνών με απλή (Πίνακας 6.4) αλλά ως ομαδοποιημένη αναπαράσταση (Πίνακας 6.5 και 6.6). Και τα δύο μοντέλα που προέκυψαν ήταν γραμμικά ενώ χρησιμοποιήθηκε ο αλγόριθμος HITON-PC για την επιλογή των χαρακτηριστικών.

Στη συγκεκριμένη αναπαράσταση κάθε επιλεγμένο χαρακτηριστικό (αμινοξύ) απεικονίζεται με ένα έως δύο γράμματα/σύμβολα στην αντίστοιχη θέση στην αλληλουχία. Στις θέσεις που επιλέχθηκαν παραπάνω από ένα χαρακτηριστικά τότε αυτά απεικονίζονται ως μια στοίβα συμβόλων. Το συνολικό ύψος κάθε στοίβας είναι ίσο με το κανονικοποιημένο άθροισμα όλων των βαρών στη συγκεκριμένη θέση (δες ενότητα 2.6.7) ενώ το σχετικό ύψος κάθε γράμματος είναι ανάλογο του συντελεστή βάρους του συγκεκριμένου

Τα επιλεγμένα χαρακτηριστικά ακολουθούν ένα χρωματικό κώδικα που σχετίζεται με τις φυσικοχημικές ιδιότητες των αμινοξέων που κωδικοποιούν (π.χ. πράσινο για υδρόφοβα). Όσα χαρακτηριστικά έχουν θετικούς συντελεστές βάρους απεικονίζονται πάνω από το άξονα  $x=0$  και ευνοούνται στις εκκρινόμενες πρωτεΐνες ενώ το αντίστροφο ισχύει για τα χαρακτηριστικά με αρνητικά βάρη. Η απεικόνιση αυτή μας οδηγεί σε κάποια συμπεράσματα για τα χαρακτηριστικά των ώριμων μορφών των εκκρινόμενων πρωτεϊνών όπως ότι στο πρώιμο ΩΤ (θέσεις 1-33) αποφεύγονται τα υδρόφοβα αμινοξέα και το θετικά φορτισμένο κατάλοιπο της Αργινίνη (R).

**Σύμβολα: @: (D,E); +: (K,R); smI: (V,G,A,P); sm: (A,G); h: (I,L,V,M); ph: (L,I,F); b: (Y,W,F); o: (T,S); x:(Y,T,S); pol: (N,Q,C); q: (N,Q,H)**

---

### 3.5.2. Μοντέλα με άλλες μεταβλητές

Το μοντέλα ώριμης και πρόδρομης μορφής εκπαιδεύτηκαν με την πληροφορία των 100 πρώτων αμινοξέων των αλληλουχιών. Το μοντέλο ώριμης μορφής, παρόλο που αγνοεί το βασικό σήμα στόχευσης το ΣΠ, μπορεί και προβλέπει τις αλληλουχίες των ΩΤ με AUC (area under the curve; δες ενότητα 6.2.4) ίσο με 90% (~74% σε άγνωστα δεδομένα; Πίνακας 3.4 και 3.3). Αποτυγχάνει όμως να αναδείξει συγκεκριμένα αμινοξικά μοτίβα. Προς την καλύτερη κατανόηση των χαρακτηριστικών που διέπουν τις εκκρινόμενες πρωτεΐνες δοκιμάσαμε εναλλακτικές μεταβλητές στην διαδικασία της εκπαίδευσης: α) σύσταση σε διπεπτίδια β) σύσταση σε τριπεπτίδια γ) ψευδο-αμινοξική σύσταση (δες 6.2.8). Πραγματοποιήσαμε διαφορετικούς συνδυασμούς των παραπάνω χαρακτηριστικών με ή χωρίς την πληροφορία της αμινοξικής αλληλουχίας και συνοψίζουμε τα μοντέλα που προέκυψαν στον Πίνακα 3.3.

Συγκρίνοντας τα αποτελέσματα, ενδιαφέρον παρουσιάζει το μοντέλο το οποίο εκπαιδεύτηκε αποκλειστικά με την πληροφορία ψευδο-αμινοξικής σύστασης. Έχει αντίστοιχα υψηλή απόδοση με το μοντέλο ώριμης μορφής (~91% και 86% σε άγνωστα δεδομένα) παρόλο που χρησιμοποιεί μόνο 17 επιλεγμένα χαρακτηριστικά. Τα 15 από τα 17 χαρακτηριστικά αντιστοιχούν στις κανονικοποιημένες τιμές αμινοξικής σύστασης (δες 6.2.8). Με μεγαλύτερο συντελεστή βάρους επιλέχτηκαν η Λευκίνη (L), προλίνη (P) και το Γλουταμινικό οξύ (E) (τα δεδομένα δεν παρουσιάζονται). Με βάση τα δεδομένα αυτά μπορούμε να συμπεράνουμε ότι η αμινοξική σύσταση των εκκρινόμενων πρωτεϊνών διαφέρει από αυτή των κυτταροπλασματικών και η πληροφορία αυτή αρκεί για να διαχωρίσει τις δύο κατηγορίες. Αντίστοιχα υψηλή απόδοση επιτυγχάνεται με τον συνδυασμό της αμινοξικής σύστασης και της σύστασης σε διπεπτίδια και τριπεπτίδια (~90% και ~85% σε άγνωστα δεδομένα; Πίνακας 3.3.). Στο μοντέλο αυτό επιλέγονται 40 χαρακτηριστικά εκ των οποίων: τα τριπεπτίδια «ΥΜΥ» και «QKK» επιλέγονται με θετικούς συντελεστές βάρυτητας ενώ το διπεπτίδιο «FP» με αρνητικό βάρος. Το «FP» ίσως εμπεριέχεται σε υδρόφοβες έλικες με την Προλίνη να λειτουργεί ως στοιχείο καμψής της έλικας.

### 3.6 Μελέτη της ικανότητας αναδίπλωσης των εκκρινόμενων πρωτεϊνών

Οι περισσότερες εκκριτικές πρωτεΐνες που χρησιμοποιούν το σύστημα Sec, ακολουθούν το μετα-μεταφραστικό μονοπάτι έκκρισης, δηλαδή πρώτα συνθέτονται στα ριβοσώματα και έπειτα οδηγούνται στο μεμβρανικό σύμπλοκο SecYEG. Μέχρι πρόσφατα επικρατούσε η αντίληψη ότι οι

εκκριτικές πρωτεΐνες αναδιπλώνονται γρήγορα και βασίζονται σε μοριακούς οδηγούς όπως η πρωτεΐνη SecB (Randall et al, 2002) και ο Trigger factor (Gouridis et al, 2009) για να παραμείνουν «ξεδίπλωτες» και να μπορέσουν να εκκριθούν.

Πρόσφατα, σε ένα *in vitro* πείραμα μεγάλης κλίμακας μετρήθηκε η διαλυτότητα των πρωτεϊνών του *E.coli* απουσία μοριακών οδηγών. Με βάση την υποκυτταρική ταξινόμηση της βάσης δεδομένων STEPdb (Orfanoudaki et al, 2014) το 45% των εκκρινόμενων πρωτεϊνών φαίνεται να είναι διαλυτές πρωτεΐνες χωρίς τη βοήθεια μοριακών οδηγών (Niwa et al, 2009).

Η ανάλυση των εκκρινόμενων πρωτεϊνών χρησιμοποιώντας μεθόδους μηχανικής μάθησης που περιγράψαμε παραπάνω, κατέληξε στο συμπέρασμα ότι η αμινοξική σύσταση των ΩΤ διαφέρει από αυτή των κυτταροπλασματικών πρωτεϊνών (Εικόνα 3.2 και Εικόνα 3.3). Λαμβάνοντας υπ όψιν τις παρατηρήσεις αυτές αποφασίσαμε να διερευνήσουμε την αμινοξική σύσταση των εκκρινόμενων πρωτεϊνών και πως αυτή σχετίζεται με ικανότητα τους για αναδίπλωση.

Η πιθανότητες αναδίπλωσης κωδικοποιούνται στην αλληλουχία μιας πρωτεΐνης (Dosztanyi et al, 2005b; Tompa, 2002). Για να σταθεροποιηθεί δομικά μια πολυπεπτιδική αλυσίδα απαιτείται ο σχηματισμός μεγάλου αριθμού αμινοξικών αλληλεπιδράσεων. Κάθε αλληλεπίδραση δύο αμινοξέων συνεισφέρει στη μείωση της συνολικής ενέργειας της πρωτεΐνης. Η τελική διαμόρφωση μιας πρωτεΐνης μπορεί να εξακριβωθεί με τεχνικές όπως η κρυσταλλογραφία ακτίνων-Χ (X-ray crystallography), η φασματοσκοπία πυρηνικού μαγνητικού συντονισμού (NMR spectroscopy) και η ηλεκτρονική μικροσκοπία (electron microscopy).

Η συνολική ενέργεια αλληλεπίδρασης των αμινοξέων μιας αναδιπλωμένης πρωτεΐνης εξαρτάται από την σύσταση αλλά και την τελική της διαμόρφωση. Μπορεί να υπολογιστεί κατά προσέγγιση από τα πεδία δυνάμεων των λυμένων δομών ενάν υπολογιστεί ο αριθμός των αμινοξικών αλληλεπιδράσεων. Όμως η εκτίμηση αυτή καθίσταται δυνατή μόνο στην περίπτωση των πρωτεϊνών για τις οποίες έχει προσδιοριστεί πειραματικά η δομή τους. Επίσης για μια μεγάλη κατηγορία πρωτεϊνών είναι αδύνατη η παρατήρηση της δομής τους καθώς είναι ενεργειακά ασταθείς (IUPs: intrinsically disordered proteins).

Πρόσφατα προτάθηκε μια μέθοδος εκτίμησης της συνολικής ενέργειας αλληλεπιδράσεων χρησιμοποιώντας την αμινοξική σύσταση μιας πρωτεΐνης (Dosztanyi et al, 2005b). Πιο αναλυτικά

η συνολική ενέργεια αλληλεπίδρασης των αμινοξέων μιας αλληλουχίας μπορεί να εκφραστεί συναρτήσει της συχνότητας των αμινοξέων που την απαρτίζουν:

$$\sum_{ij} n_i P_{ij} n_j, \text{ όπου } n_i \text{ είναι η συχνότητα του αμινοξέος } i \text{ και } P \text{ ο πίνακας πιθανότητας}$$

ενέργειας.

Κάθε στοιχείο  $P_{ij}$  του πίνακα  $P$  εκφράζει την εξάρτηση της ενέργειας του αμινοξέος τύπου  $i$  από την αμινοξική σύσταση της αλληλουχίας σε αμινοξέα τύπου  $j$ . Ο πίνακας  $P$  (Εικόνα 3.5) εκτιμήθηκε από την επίλυση ενός συστήματος τετραγωνικών εξισώσεων:

$$\frac{E_{estimated}}{L} = \sum_{ij} n_i P_{ij} n_j,$$

όπου  $E_{estimated}$  είναι η συνολική ενέργεια αλληλεπιδράσεων όπως έχει υπολογιστεί από τα πεδία δυνάμεων σε λυμένες δομές πρωτεϊνών (δες Παράρτημα C).

Η παραπάνω μέθοδος, υπολογίζει την συνολική ενέργεια μιας αλληλουχίας από την αντίστοιχη αμινοξική σύσταση τεκμηριώνοντας έτσι της έννοια της εσωτερικής αταξίας. Η ιδιότητα της εσωτερικής αταξίας εμπεριέχεται στην αλληλουχία καθώς η αντίστοιχη αμινοξική σύσταση δεν επιτρέπει σχηματισμό ικανού αριθμού αλληλεπιδράσεων.

Table 2. P energy predictor matrix

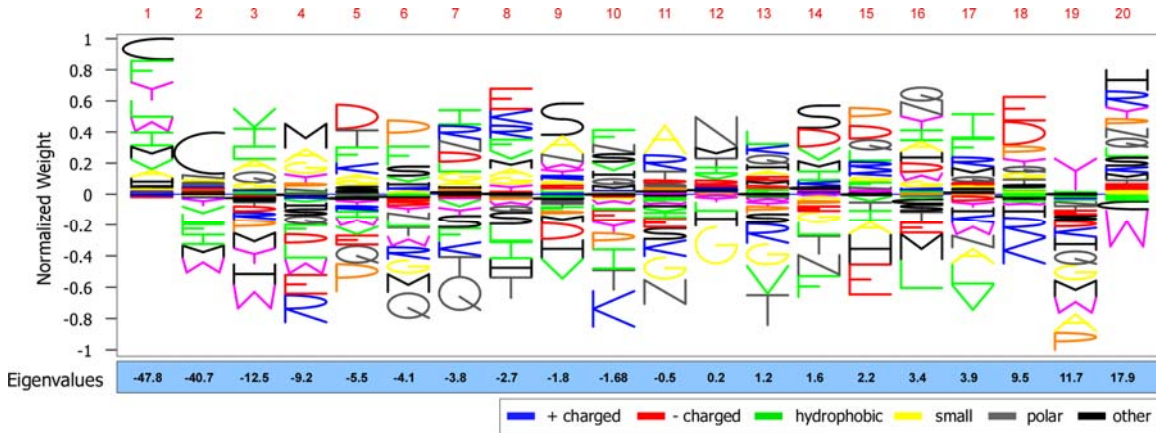
	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	-1.65	-2.83	1.16	1.80	-3.73	-0.41	1.90	-3.69	0.49	-3.01	-2.08	0.66	1.54	1.20	0.98	-0.08	0.46	-2.31	0.32	-4.62
C	-2.83	-39.58	-0.82	-0.53	-3.07	-2.96	-4.98	0.34	-1.38	-2.15	1.43	-4.18	-2.13	-2.91	-0.41	-2.33	-1.84	-0.16	4.26	-4.46
D	1.16	-0.82	0.84	1.97	-0.92	0.88	-1.07	0.68	-1.93	0.23	0.61	0.32	3.31	2.67	-2.02	0.91	-0.65	0.94	-0.71	0.90
E	1.80	-0.53	1.97	1.45	0.94	1.31	0.61	1.30	-2.51	1.14	2.53	0.20	1.44	0.10	-3.13	0.81	1.54	0.12	-1.07	1.29
F	-3.73	-3.07	-0.92	0.94	-11.25	0.35	-3.57	-5.88	-0.82	-8.59	-5.34	0.73	0.32	0.77	-0.40	-2.22	0.11	-7.05	-7.09	-8.80
G	-0.41	-2.96	0.88	1.31	0.35	-0.20	1.09	-0.65	-0.16	-0.55	-0.52	-0.32	2.25	1.11	0.84	0.71	0.59	-0.38	1.69	-1.90
H	1.90	-4.98	-1.07	0.61	-3.57	1.09	1.97	-0.71	2.89	-0.86	-0.75	1.84	0.35	2.64	2.05	0.82	-0.01	0.27	-7.58	-3.20
I	-3.69	0.34	0.68	1.30	-5.88	-0.65	-0.71	-6.74	-0.01	-9.01	-3.62	-0.07	0.12	-0.18	0.19	-0.15	0.63	-6.54	-3.78	-5.26
K	0.49	-1.38	-1.93	-2.51	-0.82	-0.16	2.89	-0.01	1.24	0.49	1.61	1.12	0.51	0.43	2.34	0.19	-1.11	0.19	0.02	-1.19
L	-3.01	-2.15	0.23	1.14	-8.59	-0.55	-0.86	-9.01	0.49	-6.37	-2.88	0.97	1.81	-0.58	-0.60	-0.41	0.72	-5.43	-8.31	-4.90
M	-2.08	1.43	0.61	2.53	-5.34	-0.52	-0.75	-3.62	1.61	-2.88	-6.49	0.21	0.75	1.90	2.09	1.39	0.63	-2.59	-6.88	-9.73
N	0.66	-4.18	0.32	0.20	0.73	-0.32	1.84	-0.07	1.12	0.97	0.21	0.61	1.15	1.28	1.08	0.29	0.46	0.93	-0.74	0.93
P	1.54	-2.13	3.31	1.44	0.32	2.25	0.35	0.12	0.51	1.81	0.75	1.15	-0.42	2.97	1.06	1.12	1.65	0.38	-2.06	-2.09
Q	1.20	-2.91	2.67	0.10	0.77	1.11	2.64	-0.18	0.43	-0.58	1.90	1.28	2.97	-1.54	0.91	0.85	-0.07	-1.91	-0.76	0.01
R	0.98	-0.41	-2.02	-3.13	-0.40	0.84	2.05	0.19	2.34	-0.60	2.09	1.08	1.06	0.91	0.21	0.95	0.98	0.08	-5.89	0.36
S	-0.08	-2.33	0.91	0.81	-2.22	0.71	0.82	-0.15	0.19	-0.41	1.39	0.29	1.12	0.85	0.95	-0.48	-0.06	0.13	-3.03	-0.82
T	0.46	-1.84	-0.65	1.54	0.11	0.59	-0.01	0.63	-1.11	0.72	0.63	0.46	1.65	-0.07	0.98	-0.06	-0.96	1.14	-0.65	-0.37
V	-2.31	-0.16	0.94	0.12	-7.05	-0.38	0.27	-6.54	0.19	-5.43	-2.59	0.93	0.38	-1.91	0.08	0.13	1.14	-4.82	-2.13	-3.59
W	0.32	4.26	-0.71	-1.07	-7.09	1.69	-7.58	-3.78	0.02	-8.31	-6.88	-0.74	-2.06	-0.76	-5.89	-3.03	-0.65	-2.13	-1.73	-12.39
Y	-4.62	-4.46	0.90	1.29	-8.80	-1.90	-3.20	-5.26	-1.19	-4.90	-9.73	0.93	-2.09	0.01	0.36	-0.82	-0.37	-3.59	-12.39	-2.68

The pairwise energy per amino acid is estimated as a quadratic form in the amino acid composition vector using the elements of this matrix.

### Εικόνα 3.5 – Πίνακας πιθανότητας ενέργειας αλληλεπίδρασης

Πιθανότητα ενέργειας αλληλεπίδρασης ανά ζεύγους αμινοξέων όπως έχει υπολογιστεί από την επίλυση ενός συστήματος εξισώσεων που προέκυψαν από τον υπολογισμό ενέργειας αλληλεπίδρασης των αμινοξέων από δομές πρωτεϊνών (Dosztanyi et al, 2005b). Ο πίνακας είναι συμμετρικός και πολλαπλασιαζόμενος με το διάνυσμα που αναπαριστά την αμινοξική σύσταση (ποσοστό κάθε αμινοξέος επί του συνόλου των αμινοξέων) υπολογίζει την συνολική πιθανή ενέργεια αναδίπλωσης μιας πρωτεΐνης.

Υπολογίζοντας το σύστημα συντεταγμένων που ορίζεται από τα ιδιοδιανύσματα του πίνακα  $P$  (δες παράρτημα Β) μπορούμε να αναλύσουμε την έννοια της αναδίπλωσης μιας πρωτεΐνης στους βασικούς τις άξονες. Οι βασικοί άξονες αναδίπλωσης με βάση τον πίνακα  $P$  και οι αντίστοιχες ιδιοτιμές απεικονίζονται στην Εικόνα 3.6. Στην συγκεκριμένη αναπαράσταση κάθε ιδιοδιάνυσμα του  $P$  απεικονίζεται με μια στοίβα γραμμάτων (αμινοξέων) ενώ στο κάτω μέρος σημειώνονται οι αντίστοιχες ιδιοτιμές.



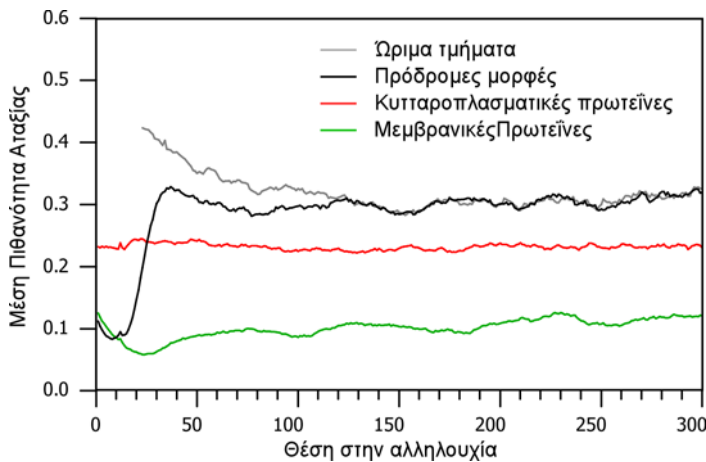
**Εικόνα 3.6 – Βασικοί άξονες ενέργειας αναδίπλωσης πρωτεϊνών.**

Ο πίνακας πιθανότητας ενέργειας αλληλεπίδρασης ανά ζεύγος αμινοξέων  $P$  (Εικόνα 3.5) μπορεί να αναλυθεί σε 20 βασικούς άξονες (principal components) οι οποίοι αντιστοιχούν στα αντίστοιχα ιδιοδιανύσματα (eigenvectors) του πίνακα (δες Παράρτημα Β). Στη συγκεκριμένη αναπαράσταση τα ιδιοδιανύσματα έχουν κανονικοποιηθεί από το -1 στο 1. Στον πρώτο άξονα βλέπουμε ότι ο συνδυασμός υδρόφοβων αμινοξέων (LVIF) με υδρόφοβα αρωματικά κατάλοιπα (WY) και Κυστεΐνη μπορούν να οδηγήσουν μια πρωτεΐνη σε πιο σταθερή αναδίπλωση (αρνητικές ενέργειες αλληλεπίδρασης αμινοξέων). Συγκεκριμένα τα κατάλοιπα Κυστεΐνης συνεισφέρουν σχηματίζοντας δισουλφιδικούς δεσμούς μεταξύ τους ενώ τα υδρόφοβα κατάλοιπα με υδρόφοβες αλληλεπιδράσεις μεταξύ τους.

Ερμηνεύοντας τους άξονες αναδίπλωσης θα χρειαστεί να πούμε ότι οι αρνητικές τιμές αντιστοιχούν και σε αρνητικές ενέργειες αλληλεπίδρασης κατά συνέπεια ενέργειες που ευνοούν την αναδίπλωση μιας πρωτεΐνης. Αντίθετα αμινοξέα με θετικές ενέργειες αλληλεπίδρασης συνεισφέρουν θετικά στην συνολική ενέργεια μιας πρωτεΐνης και στην αποσταθεροποίησή της. Έτσι για παράδειγμα πρωτεΐνες πλούσιες σε υδρόφοβα αμινοξέα (F,L,I,V) και κατάλοιπα Κυστεΐνης (πρώτος άξονας; Εικόνα 3.6) έχουν περισσότερες προοπτικές για ευνοϊκές αλληλεπιδράσεις (αρνητικές ενέργειες) που συνεπάγονται στη σταθερή αναδίπλωση μιας

πρωτεΐνης. Αντίθετα αλληλουχίες πλούσιες σε αρνητικά φορτισμένα αμινοξέα (D,E) και Προλίνες συνεπάγονται δομική αταξία (άξονας 18, Εικόνα 3.6).

Χρησιμοποιώντας την παραπάνω μέθοδο εκτίμησης της συνολικής ενέργειας αναπτύχθηκε ένα εργαλείο, το IUPred (Dosztanyi et al, 2005a) με σύστημα βαθμολόγησης ικανό να προβλέψει την εσωτερική αταξία των πρωτεϊνών. Πιο αναλυτικά, το συγκεκριμένο βιοπληροφορικό εργαλείο υπολογίζει την πιθανότητα αταξίας σε κάθε θέση της αλληλουχίας εκτιμώντας την συνολική ενέργεια αλληλεπίδρασης των γειτονικών αμινοξέων στη θέση αυτή.



**Εικόνα 3.7 – Μέσο προφίλ δομικής αταξίας (disorder)**

Χρησιμοποιώντας το βιοπληροφορικό εργαλείο IUPred, το οποίο προβλέπει περιοχές αταξίας, υπολογίσαμε την κατά μέσο όρο πιθανότητα αταξίας ανά θέση στην αλληλουχία για τις: πρόδρομες μορφές των εκκρινόμενων πρωτεϊνών (μαύρο), ώριμες μορφές των εκκρινόμενων πρωτεϊνών (γκρί), κυτταροπλασματικές

πρωτεΐνες (κόκκινο) και πρωτεΐνες της εσωτερικής μεμβράνης (πράσινο). Οι τέσσερις αυτές κατηγορίες παρουσιάζουν διαφορετικά προφίλ και συγκεκριμένα κυμαίνονται σε διαφορετικά ενεργειακά επίπεδα. Ξεκινώντας από τις μεμβρανικές πρωτεΐνες οι οποίες είναι πιο σταθερές (μέση πιθανότητα αταξίας ~0.1) ακολουθούν οι κυτταροπλασματικές πρωτεΐνες (~0.23) και τέλος οι ώριμες μορφές των εκκρινόμενων πρωτεϊνών (~0.31). Παρατηρούμε επίσης ότι οι πρόδρομες μορφές των εκκρινόμενων πρωτεϊνών αποτελούνται από δύο περιοχές με διαφορετικά επίπεδα ενέργειας. Η περιοχή 1-23 αντιστοιχεί στο ΣΠ το οποίο φαίνεται να είναι πολύ πιο σταθερό από ότι το υπόλοιπο τμήμα της πρωτεΐνης με πολύ μικρή πιθανότητα αταξίας κοντά σε αυτή των μεμβρανικών πρωτεϊνών. Τέλος απουσία του ΣΠ το πρώιμο ΩΤ (θέσεις 1-100) των εκκρινόμενων περιοχών (γκρι) παρουσιάζει ακόμα μεγαλύτερη πιθανότητα αταξίας (ξεκινώντας από ~0.4).

Χρησιμοποιώντας το εργαλείο IUPred υπολογίσαμε την κατά μέσο όρο πιθανότητα αταξίας ανά θέση στην αλληλουχία για τέσσερις κατηγορίες: 1) πρόδρομες μορφές εκκρινόμενων πρωτεϊνών, 2) ώριμα τμήματα εκκρινόμενων πρωτεϊνών, 3) κυτταροπλασματικές πρωτεΐνες και 4) πρωτεΐνες της εσωτερικής μεμβράνης (Εικόνα 3.7).

Τα μέσο προφίλ αταξίας των πρωτεϊνών μας οδηγεί σε κάποια ενδιαφέροντα συμπεράσματα. Αρχικά παρατηρούμε ότι οι τέσσερις αυτές κατηγορίες παρουσιάζουν διαφορετικά προφίλ και κυμαίνονται σε διαφορετικά ενεργειακά επίπεδα. Ξεκινώντας από τις



μεμβρανικές πρωτεΐνες οι οποίες είναι πιο σταθερές (μέση πιθανότητα αταξίας  $\sim 0.1$ ) ακολουθούν οι κυτταροπλασματικές πρωτεΐνες ( $\sim 0.23$ ) και τέλος οι ώριμες μορφές των εκκρινόμενων πρωτεϊνών ( $\sim 0.31$ ; (Εικόνα 3.7)).

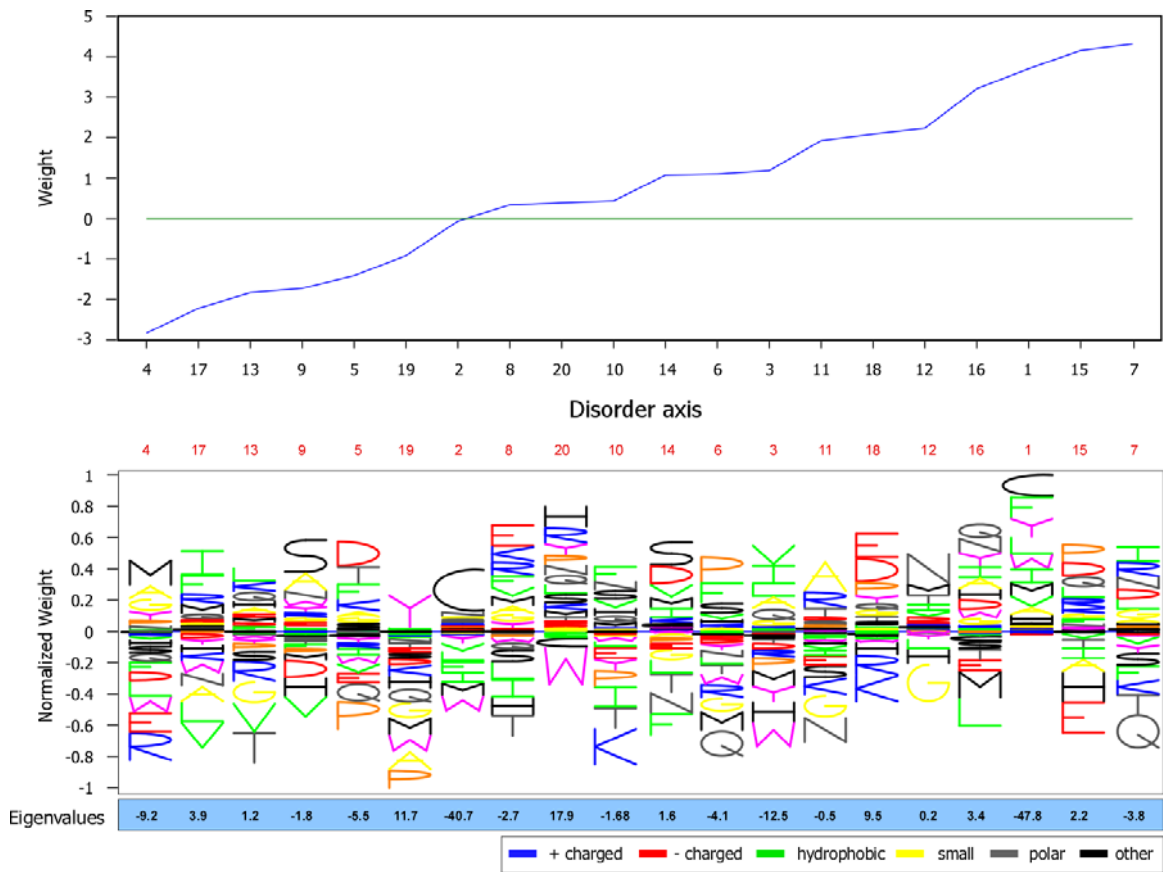
Κατά συνέπεια οι αλληλουχίες των εκκρινόμενων πρωτεϊνών εμπεριέχουν την πληροφορία για μικρότερη πιθανότητα ενέργειας αλληλεπίδρασης. Αυτό μεταφράζεται σε μικρότερη πιθανότητα αναδίπλωσης και συμφωνεί με πειραματικές ενδείξεις που θέλουν τις εκκρινόμενες πρωτεΐνες να παραμένουν «ξεδίπλωτες» για περισσότερο χρονικό διάστημα (Chatzi et al, in preparation).

Παρατηρούμε επίσης ότι οι πρόδρομες μορφές των εκκρινόμενων πρωτεϊνών αποτελούνται από δύο περιοχές με μεγάλη διαφορά στην πιθανότητα αταξίας (μαύρη καμπύλη; Εικόνα 3.7). Η πρώτη αντιστοιχεί στο ΣΠ (θέσεις 1-23) η οποία φαίνεται να είναι ενεργειακά σταθερή ( $\sim 10\%$  πιθανότητα αταξίας). Η ιδιότητα αυτή κατά πάσα πιθανότητα οφείλεται στην υδρόφοβη περιοχή (h-domain) των ΣΠ. Η δεύτερη περιοχή αφορά στο πρώιμο ΩΤ (θέσεις 24-74 ή αλλιώς +1 μέχρι +51 του ΩΤ) και παρουσιάζει πολύ μεγαλύτερη αστάθεια ( $\sim 30\%$  πιθανότητα).

Τέλος είναι σημαντικό να επισημάνουμε ότι όταν το ΣΠ απουσιάζει (γκρί καμπύλη έναντι της μάυρης; Εικόνα 3.7) τότε το πρώιμο ΩΤ αποσταθεροποιείται ακόμα περισσότερο, με την περιοχή του καρβοξυτελικού άκρου να πλησιάζει μέχρι και  $40\%$  πιθανότητα για εσωτερική αταξία. Το αποτέλεσμα αυτό συνεισφέρει στην θεωρία ότι οι εκκρινόμενες πρωτεΐνες πιθανόν να περιέχουν ένα σύνθετο σήμα έκκρισης το οποίο αποτελείται από το ΣΠ και το πρώιμο ΩΤ, μάλιστα φαίνεται το καθένα να έχει διαφορετική πιθανότητα αναδίπλωσης.

### **3.7 Εκπαίδευση μοντέλων διαχωρισμού χρησιμοποιώντας την πληροφορία της πιθανότητας αταξίας**

Η προηγούμενη ανάλυση έδειξε ότι τα ΩΤ των εκκρινόμενων πρωτεϊνών παρουσιάζουν μεγαλύτερη πιθανότητα αταξίας και η ικανότητα αυτή εμπεριέχεται στην αμινοξική τους σύσταση. Η πληροφορία της πιθανότητας ενέργειας αλληλεπίδρασης φαίνεται ότι συνοψίζει τα χαρακτηριστικά των εκκρινόμενων πρωτεϊνών. Αποφασίσαμε να αποδείξουμε την υπόθεση ενσωματώνοντας στην διαδικασία της εκπαίδευσης την πληροφορία της πιθανότητας ενέργειας αλληλεπίδρασης.



**Εικόνα 3.8 – Μοντέλο δομικής αταξίας – επιλεγμένα χαρακτηριστικά**

Εκπαιδεύσαμε ένα μοντέλο διαχωρισμού χρησιμοποιώντας ως μεταβλητές εκπαίδευσης 20 τιμές συνολικής ενέργειας αλληλεπίδρασης (εξίσωση 3.1) που αντιστοιχούν στην ενεργειακή συνεισφορά για κάθε τύπο αμινοξέος και προκύπτουν από τους 20 άξονες αταξίας (ιδιοδιανύσματα του πίνακα  $P$ ). Το μοντέλο αυτό το ονομάσαμε μοντέλο αταξίας. Α) τα επιλεγμένα χαρακτηριστικά του μοντέλου που αντιστοιχούν στους 20 άξονες αταξίας με τα αντίστοιχα βάρη που προκύπτουν από την γραμμική εξίσωση διαχωρισμού. Β) απεικόνιση των 20 αξόνων αταξίας. Λαμβάνοντας υπ' όψιν τα πιο σημαντικά χαρακτηριστικά (πιο θετικά και πιο αρνητικά βάρη) τότε η αμινοξική σύσταση των εκκρινόμενων πρωτεϊνών υπαγορεύεται από τους άξονες αταξίας 7 και 15. Οι άξονες αυτοί μας λένε ότι υπάρχουν δύο τρόποι να συνθέσεις εκκρινόμενες πρωτεΐνες, ο πρώτος χρησιμοποιώντας ως επί το πλείστον σε Γλουταμίνες (Q), Θρεονίνες (T) και Λυσίνες (K) και ένας δεύτερος με Προλίνες (P), Ασπαρτικά οξέα (D) και Γλουταμίνη (Q). Αντίστοιχα με βάση τους άξονες που επιλέγονται με αρνητικά βάρη (4 και 17) η σύσταση των κυτταροπλασματικών πρωτεϊνών βασίζεται είτε στο συνδυασμό υδρόφοβων αμινοξέων (F,I) και Αργινίνης(R), είτε εναλλακτικά σε Ασπαρτικό οξύ (E), Αργινίνη (R) σε συνδυασμό με τα υδρόφοβα αμινοξέα Λευκίνη (L) και Τρυπτοφάνη (W).

Προς την διερεύνηση της αμινοξικής σύστασης που εξηγεί την μεγαλύτερη πιθανότητα αταξίας των εκκρινόμενων πρωτεϊνών υπολογίσαμε την συνολική ενέργεια αλληλεπίδρασης ανά τύπο αμινοξέος:

$$e_i^k (estimated) = sign(\lambda_i) \sum_{j=1}^{20} P_{ij} n_j^k, \quad (3.1)$$

όπου για την πρωτεΐνη  $k$ ,  $e_i^k$  είναι η συνολική ενέργεια όλων των αμινοξέων τύπου  $i$  που αλληλεπιδρούν με τα υπόλοιπα αμινοξέα της αλληλουχίας,  $n_j$  η συχνότητα των αμινοξέων τύπου  $j$  στην αλληλουχία,  $P_{ij}$  η αντίστοιχη τιμή πιθανότητας ενέργειας για το ζεύγος αμινοξέων  $i, j$ ,  $\lambda_i$  η ιδιοτιμή του ιδιοδιανύσματος  $i$  του πίνακα  $P$ .

Έτσι εισάγαμε στην διαδικασία της εκπαίδευσης 20 μεταβλητές που αντιστοιχούν στην ενεργειακή συνεισφορά κάθε τύπου αμινοξέος και εκπαιδεύσαμε ένα μοντέλο διαχωρισμού των ώριμων μορφών από τις κυτταροπλασματικές (μοντέλο αταξίας) χρησιμοποιώντας μόνο αυτές τις τιμές. Συγκρίνοντας τα δύο μοντέλα, το μοντέλο αταξίας έχει απόδοση ~1% καλύτερη από το μοντέλο ώριμης μορφής (~91% έναντι 90%) (ενότητα 3.5.1) ενώ επιλέγονται 20 χαρακτηριστικά (όλες οι μεταβλητές) σε σύγκριση με τα 121 στη δεύτερη περίπτωση (Πίνακας 2.1). Τέλος παρατηρούμε ότι το μοντέλο αταξίας αποδίδει πολύ καλύτερα στα δεδομένα αξιολόγησης αλλά και τα πειραματικά δεδομένα.

Απεικονίσαμε τους άξονες αταξίας με σειρά αυξανόμενου συντελεστή βάρους στην γραμμική εξίσωση διαχωρισμού του μοντέλου (Εικόνα 3.8). Πρώτο συμπέρασμα είναι ότι η διαδικασία επιλογής χαρακτηριστικών κατά την εκπαίδευση επέλεξε όλους τους άξονες αταξίας ως σημαντικούς για τον διαχωρισμό. Με μεγαλύτερα βάρη επιλέχτηκαν οι άξονες 15, 7 (μέγιστα θετικά) 4 και 17 (μέγιστα αρνητικά).

Αν αναλύσουμε την αμινοξική σύσταση όπως συνοψίζεται στους συγκεκριμένους άξονες αταξίας καταλήγουμε σε ενδιαφέροντα συμπεράσματα για την επιλογή των αμινοξέων που υπαγορεύουν την αμινοξική σύσταση κάθε κατηγορίας πρωτεϊνών. Οι άξονες 15 και 7 μας λένε ότι υπάρχουν δύο τρόποι να συνθέσεις μια εκκρινόμενη πρωτεΐνη, ο πρώτος πλούσιος σε Γλουταμίνες (Q), Θρεονίνες (T) και Λυσίνες (K) και ένας δεύτερος με Προλίνες (P), Ασπαρτικά οξέα (D) και Γλουταμίνη (Q). Αντίστοιχα με βάση τους άξονες αταξίας 4 και 17, η σύσταση των κυτταροπλασματικών πρωτεϊνών βασίζεται είτε στο συνδυασμό υδρόφοβων αμινοξέων (F,I) με το

κατάλοιπο της Αργινίνη (R), είτε εναλλακτικά στο συνδυασμό συγκεκριμένων φορτισμένων κατάλοιπων (Ασπαρτικό οξύ (E) και Αργινίνη (R)) με τα υδρόφοβα αμινοξέα Λευκίνη (L) και Τρυπτοφάνη (W).

Πίνακας 3.2 – Σύγκριση απόδοσης μοντέλων

Μοντέλο	Σύνολο Εκπαίδευσης (AUC)	Σύνολο Αξιολόγησης (AUC)	Πειραματικά Δεδομένα (AUC)	Επιλεγμένα Χαρακτηριστικά
<i>Μοντέλο πρόδρομης μορφής</i>	98.57	97.04	72.61	123
<i>Μοντέλο ώριμης μορφής</i>	90.37	74.23	79.58	121
<i>Μοντέλο αταξίας</i>	91.14	87.73	83.17	20(all)

Η επιλογή υδρόφοβων αμινοξέων σε συνδυασμό με αρνητικά και θετικά φορτισμένα αμινοξέα συμφωνεί με τα μέχρι τώρα δεδομένα για κυτταροπλασματικές πρωτεΐνες οι οποίες αποκτούν σφαιρική δομή. Τα υδρόφοβα αμινοξέα είναι γνωστό ότι σχηματίζουν τον υδρόφοβο πυρήνα ενώ τα φορτισμένα βρίσκονται στην επιφάνεια των μορίων και έρχονται σε επαφή με το υδατικό περιβάλλον (Tompa, 2002).

Οι τύποι αμινοξέων που επιλέγονται στην περίπτωση των εκκρινόμενων πρωτεϊνών προτάσει μια πιο υψηλά ενεργειακή κατάσταση και μια μεγαλύτερη πιθανότητα αταξίας. Η συγκεκριμένη παρατήρηση συμφωνεί με τα αντίστοιχα συμπεράσματα στα οποία καταλήξαμε μέσω του μοντέλου ώριμης μορφής αλλά και με τις πειραματικές ενδείξεις ότι το 83% των εκκρινόμενων παραμένουν υδατοδιαλυτές χωρίς την βοήθεια μοριακών οδηγών (Niwa et al, 2009).



---

**Πίνακας 3.3 – Μοντέλα διαχωρισμού των ώριμων μορφών των εκκριτικών πρωτεϊνών από τις κυτταροπλασματικές πρωτεΐνες**

Συγκρίνουμε τα μοντέλα διαχωρισμού των αλληλουχιών των ώριμων τμημάτων (ΩΤ) από τις κυτταροπλασματικές, τα οποία προέκυψαν χρησιμοποιώντας οχτώ διαφορετικές μεταβλητές εκπαίδευσης: αμινοξική αλληλουχία κωδικοποιημένη με απλό, χαλαρό και συμπαγή τρόπο (Πίνακας 6.4 μέχρι 6.6), αμινοξική σύσταση, σύσταση σε διπεπτίδια και τριπεπτίδια, ψευδο-αμινοξική σύσταση (δες 6.2.8) και συνολική ενέργεια αλληλεπίδρασης ανά τύπο αμινοξέος (ή αλλιώς πιθανότητα αταξίας ανά άξονα αταξίας) (δες ενότητα 3.7).

Όλες οι μεταβλητές εκπαίδευσης στην ουσία κωδικοποιούν με διαφορετικό τρόπο την αλληλουχία ή τις φυσικοχημικές ιδιότητες των αμινοξέων που βρίσκονται στις θέσεις +3 μέχρι +102 των ώριμων τμημάτων (+2 μέχρι +101 στην περίπτωση των κυτταροπλασματικών πρωτεϊνών). Οι τρεις πρώτες μεταβλητές αφορούν διαφορετικές αναπαραστάσεις των αμινοξέων. Στην απλή περίπτωση κάθε αμινοξύ απεικονίζεται ξεχωριστά ενώ στην «χαλαρή» και «συμπαγή» απεικόνιση δοκιμάζονται διαφορετικές ομαδοποιήσεις των αμινοξέων με βάση τις φυσικοχημικές τους ιδιότητες. Στη συνέχεια δοκιμάσαμε μεταβλητές όπως η αμινοξική σύσταση αλλά και η σύσταση σε ζεύγη ή τριπλέτες κατάλοιπων. Εναλλακτικός τρόπος αναπαράστασης μιας αμινοξικής αλληλουχίας αποτελεί η ψευδο-αμινοξική σύσταση (δες 6.2.8; (Chou, 2001)) η οποία στην παρούσα υλοποίηση εμπεριέχει την πληροφορία της πολλαπλασιαστικής υδροφοβικότητας γειτονικών αμινοξέων (η γειτονιά ορίζεται από ανά δύο μέχρι ανά 60 θέσεις απόσταση). Τέλος εκπαιδεύσαμε μοντέλα με 20 μεταβλητές ενέργειας που αντιστοιχούν στην συνολική ενεργειακή συνεισφορά κάθε τύπου αμινοξέος δεδομένου της αμινοξικής σύστασης της αλληλουχίας (δες ενότητα 3.7).

Παρατηρούμε ότι όλα τα μοντέλα που προκύπτουν καταφέρνουν να διαχωρίσουν τα ΩΤ από τις κυτταροπλασματικές πρωτεΐνες με πολύ υψηλό ποσοστό επιτυχίας (AUC >81%) παρόλο που τα ΣΠ απουσιάζουν, ενώ σχεδόν όλα τα μοντέλα που προκύπτουν είναι γραμμικά (πρώτου βαθμού πολυώνυμα). Παρόλο που η απόδοση είναι ικανοποιητική σε όλες τις περιπτώσεις, τα μοντέλα παρουσιάζουν διαφορές στις στην απόδοση τους στα πειραματικά δεδομένα αλλά και στον αριθμό των χαρακτηριστικών που επιλέγονται. Είναι αξιοσημείωτο ότι πολύ περισσότερα χαρακτηριστικά χρειάζονται στην περίπτωση που αμινοξική αλληλουχία κωδικοποιείται με απλό, συμπαγή και χαλαρό τρόπο (μοντέλο 1, 121 χαρακτηριστικά) ενώ σε άλλες περιπτώσεις όπως η ενέργεια αλληλεπίδρασης ή η αμινοξική σύσταση πολύ λιγότερα χαρακτηριστικά είναι αρκετά για να ξεχωρίσουν τις κατηγορίες (στα μοντέλα 5 και 8, επιλέγονται μόνο 20 και 17 χαρακτηριστικά αντίστοιχα).

---



A/A	Ομαδοποίηση Αμινοξέων			Συσταση σε:			Ψευδο-αμινιξική Σύσταση	Ενέργεια αλληλεπίδρασης	Απόδοση μοντέλου (AUC))			Τάξη	Αριθμός χαρακτηριστικών
									Σύνολο Εκπαίδευσης	Σύνολο Αξιολόγησης	Πειραματικά Δεδομένα		
	απλή	χαλαρή	συμπαγής	αμινοξέα	διπεπτίδια	τριπεπτίδια							
<b>1</b>	+	+	+						90.37	74.23	79.58	1	121
<b>2</b>			+						81.75	67.62	52.05	1	63
<b>3</b>		+							87.82	81.09	90.29	2	all (1100)
<b>4</b>	+		+						89.59	73.40	71.07	1	122
<b>5</b>							+		91.14	87.73	83.17	1	20(all)
<b>6</b>		+	+				+		90.51	81.47	87.52	1	48
<b>7</b>							+	+	90.99	87.75	88.30	2	all (160)
<b>8</b>							+		91.09	87.94	85.47	2	17
<b>9</b>	+				+				94.11	83.13	78.66	1	165
<b>10</b>				+	+	+			90.54	85.71	77.94	1	40



### 3.8 Μελέτη υδρόφοβων περιοχών στα ώριμα τμήματα

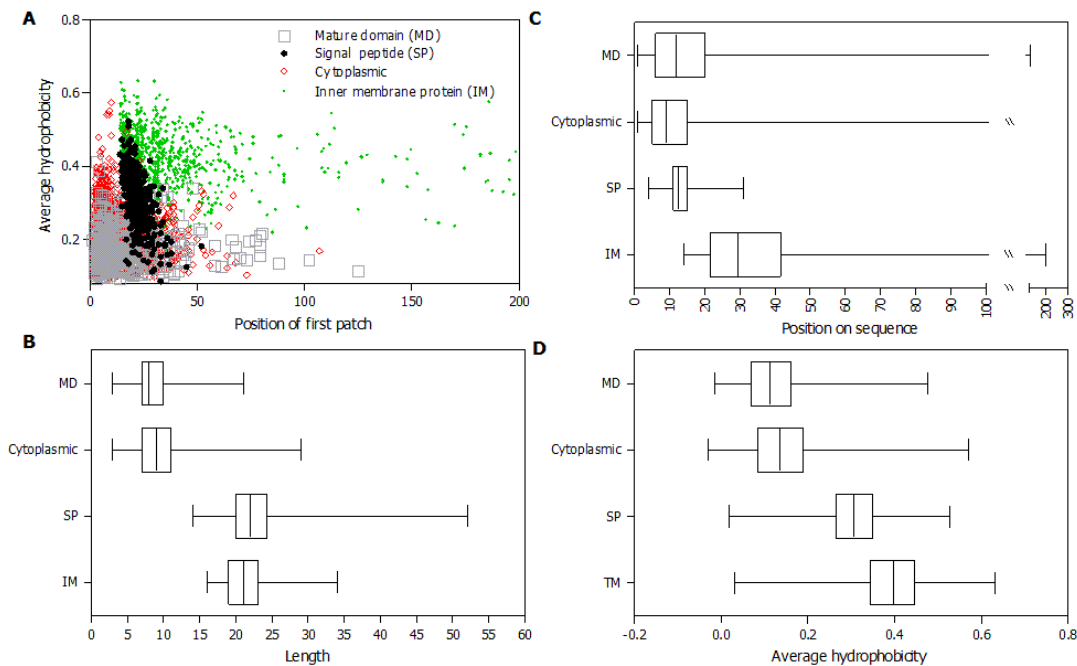
Πρόσφατα αποδείχτηκε ότι το ΣΠ και το ΩΤ των εκκρινόμενων πρωτεϊνών προσδένονται σε διαφορετικές θέσεις πάνω στην πρωτεΐνη κινητήρα του συστήματος Sec, την SecA. Μεταλλάξεις που έγιναν στο ΩΤ της περιπλασμικής πρωτεΐνης PhoA υπέδειξαν δύο υδρόφοβες περιοχές στις θέσεις 67-71 και 93-102 (Chatzi et al, in preparation). Είναι επίσης ενδιαφέρον να αναφέρουμε ότι το ΩΤ της PhoA είναι ικανό να οδηγηθεί σε έκκριση χωρίς την παρουσία του ΣΠ (Chatzi et al, in preparation). Το ίδιο φαίνεται αν συμβαίνει και με άλλες πρωτεΐνες (Hayano et al, 1991) ανιχνεύονται και στο περίπλασμα χωρίς της προφανή παρουσία κάποιου ΣΠ.

Στην ανάλυση με μοντέλα διαχωρισμού που περιγράψαμε παραπάνω, οι αλληλουχίες στοιχήθηκαν στο αμινοτελικό τους άκρο. Τα μοντέλα που εκπαιδεύτηκαν δεν ανέδειξαν κανένα αμινοξικό μοτίβο παρόλο που η διαχωριστική τους ικανότητα ήταν υψηλή. Όμως η αδυναμία επιλογής χαρακτηριστικών σε καθορισμένες θέσεις δεν συνεπάγεται απαραίτητα στην ανυπαρξία αμινοξικών μοτίβων. Ενδέχεται τα πιθανά σήματα να βρίσκονται σε διαφορετικές θέσεις σε κάθε πρωτεΐνη και να «ισοπεδώνονται» λόγω στοίχισης των αλληλουχιών. Μια άλλη εξήγηση θα μπορούσε να είναι ότι τα πιθανά σήματα σχηματίζονται στον τρισδιάστατο χώρο μετά από αναδίπλωση των ΩΤ και δεν είναι κατ' ανάγκη σε συνεχόμενες θέσεις στην αλληλουχία.

Κατά συνέπεια η απεικόνιση της αλληλουχίας με στοιχισμένο τρόπο ίσως δεν εξυπηρετεί στην ανίχνευση πιθανών υδρόφοβων μοτίβων. Διερευνήσαμε το περιεχόμενο των εκκρινόμενων πρωτεϊνών σε υδρόφοβες περιοχές χρησιμοποιώντας την κλίμακα υδροφοβικότητας Kyte και Doolittle (δες ενότητα 6.2.11). Συνεκτιμώντας τα πειραματικά δεδομένα που αφορούν τις υδρόφοβες περιοχές της πρωτεΐνης PhoA, εστίασαμε στην εύρεση περιοχών μήκους από 5 έως 9 αμινοξέα. Προσδιορίσαμε τον αριθμό τη θέση, το μήκος αλλά και την μέγιστη υδροφοβικότητα αυτών των περιοχών και συγκρίναμε τις ιδιότητες αυτές ανάμεσα στις εκκρινόμενες πρωτεΐνες και τις κυτταροπλασματικές (Εικόνα 3.9) έχοντας ως μέτρα αναφοράς την υδρόφοβη περιοχή των ΣΠ και τις ΔΠ των μεμβρανικών πρωτεϊνών.

Προσδιορίσαμε την θέση που βρίσκεται η πρώτη υδρόφοβη περιοχή στις εκκρινόμενες και κυτταροπλασματικές. Φαίνεται ότι υδρόφοβες περιοχές ξεκινάνε νωρίτερα στην αλληλουχία των κυτταροπλασματικών πρωτεϊνών (μέση θέση 9 σε σύγκριση με 12 στα ΩΤ) (Εικόνα 3.9C). Στη συνέχεια υπολογίσαμε το μήκος αλλά την μέγιστη υδροφοβικότητα αυτών των περιοχών. Οι

εκκρινόμενες πρωτεΐνες τείνουν να έχουν μικρότερες τέτοιες περιοχές κατά ένα αμινοξύ (8 έναντι 9 αμινοξέων; Εικόνα 3.9B) και είναι ελάχιστα πιο υδρόφοβες (~0.11 έναντι~0.14; Εικόνα 3.9D).



**Εικόνα 3.9 – Σύνοψη των χαρακτηριστικών των υδρόφοβων περιοχών.**

**A.** Διάγραμμα σημείων που απεικονίζει την θέση της πρώτης υδρόφοβης περιοχής σε συνάρτηση με την αντίστοιχη μέση υδροφοβικότητα της για τις εκκρινόμενες πρωτεΐνες (γκρί τετράγωνο) και κυτταροπλασματικές (κόκκινος κύκλος). Στο ίδιο γράφημα, για λόγους σύγκρισης, απεικονίζονται οι πρώτες ΔΠ των μεμβρανικών πρωτεϊνών (πράσινα σημεία) καθώς επίσης και ο υδρόφοβος πυρήνας των ΣΠ (μαύροι κύκλοι). **B.** Θηκογράφημα (box and whisker plot) που απεικονίζει την θέση της πρώτης υδρόφοβης περιοχής, οι κατηγορίες είναι αντίστοιχες με το προηγούμενο γράφημα **C.** μήκος της πρώτης υδρόφοβης περιοχής **D.** μέση υδροφοβικότητα της πρώτης υδρόφοβης περιοχής. Παρατηρούμε ότι οι υδρόφοβες περιοχές ξεκινάνε νωρίτερα στις κυτταροπλασματικές πρωτεΐνες σε σύγκριση τα ΩΤ (θέσεις 11 και 16 αντίστοιχα; γράφημα C). Επίσης οι υδρόφοβες περιοχές στα ΩΤ παρουσιάζουν μικρές διαφορές ως προς το μήκος και την μέση υδροφοβικότητα σε σύγκριση με αυτές των κυτταροπλασματικών πρωτεϊνών (μήκη ~8 και ~9 αμινοξέα αντίστοιχα ενώ μέση υδροφοβικότητα ~12 και 14 αντίστοιχα; γραφήματα B και D)

### 3.9 Μελέτη των περιοχών που προάγουν τον σχηματισμό συσσωματώσεων

Είναι γνωστό ότι πρωτεΐνες που δεν αναδιπλώνονται σωστά σε ένα κύτταρο αλληλεπιδρούν μεταξύ τους σχηματίζοντας συσσωματώματα τα οποία είναι συνήθως τοξικά για το κύτταρο (Bednarska et al, 2013; Navarro et al, 2014). Οι αλληλουχίες στις οποίες οφείλεται η δημιουργία συσσωματώσεων (αμυλοειδή) έχουν ιδιαίτερα μελετηθεί λόγω της άμεσης συσχέτισης τους με ανθρώπινες ασθένειες όπως το Alzheimer. Μελέτες για τον τρόπο που πολυμερίζονται τα αμυλοειδή και άλλα πρωτεϊνικά πολυμερή όπως τα ινίδια (fibrils και curli) στα βακτήρια (Cherny et al, 2005) συνηγορούν στο συμπέρασμα ότι η αμινοξική σύσταση μιας αλληλουχίας καθορίζει το βαθμό και την προδιάθεση της να δημιουργήσει συσσωματώματα.

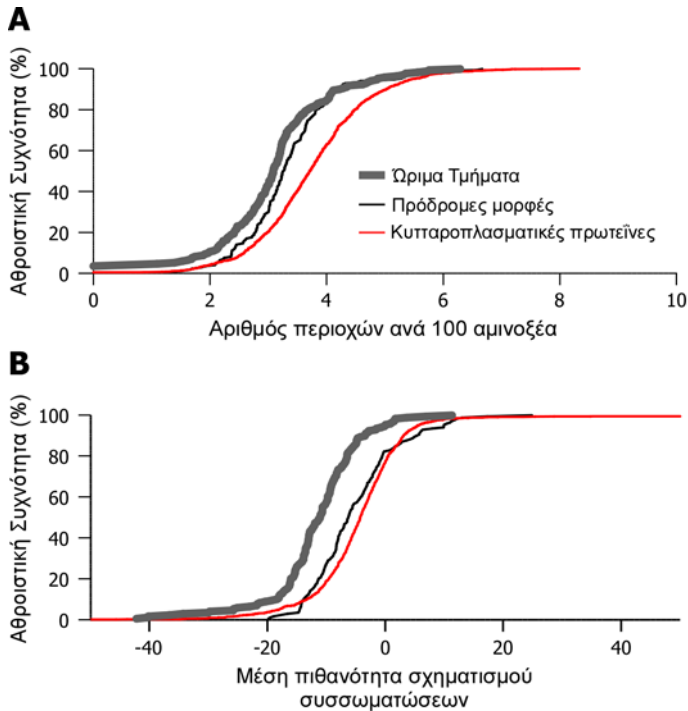
Διάφορα βιοπληροφορικά εργαλεία έχουν αναπτυχθεί για την πρόβλεψη περιοχών που οδηγούν σε σχηματισμό συσσωματώσεων (hot spots) τα οποία βασίζονται σε χαρακτηριστικά όπως η μέση πυκνότητα των αμινοξέων (Fernandez-Escamilla et al, 2004) και η συχνότητα εμφάνισης κάθε είδους αμινοξέος σε γνωστές αλληλουχίες που σχηματίζουν αμυλοειδή (Conchillo-Sole et al, 2007).

Στις προηγούμενες ενότητες συσχέτισαμε την αμινοξική σύσταση με την πιθανότητα αταξίας και το ρυθμό αναδίπλωσης. Αποφασίσαμε να εξετάσουμε εάν η μεγαλύτερη πιθανότητα για εσωτερική αταξία συνεπάγεται αντίστροφα, μικρότερη προδιάθεση για σχηματισμό συσσωματώσεων στις εκκρινόμενες πρωτεΐνες.

Χρησιμοποιώντας το βιοπληροφορικό εργαλείο AGGRESCAN (Conchillo-Sole et al, 2007) προβλέψαμε για τις εκκριτικές και κυτταροπλασματικές αλληλουχίες την ύπαρξη περιοχών με προδιάθεση για σχηματισμό συσσωματώσεων. Το εργαλείο αυτό εφαρμόζει ένα σύστημα βαθμονόμησης των αμινοξέων όπως προκύπτει από την συχνότητα εμφάνισης τους σε γνωστές αλληλουχίες που έχουν την τάση να πολυμερίζονται όπως τα αμυλοειδή.

Συγκρίναμε τα αποτελέσματα της πρόβλεψης για τα ΩΤ, τις πρόδρομες μορφές και τις κυτταροπλασματικές πρωτεΐνες (Εικόνα 3.10). Ο μέσος αριθμός τέτοιων περιοχών είναι μεγαλύτερος για τις κυτταροπλασματικές πρωτεΐνες σε σύγκριση με τις πρόδρομες μορφές (~3.7 έναντι ~3), ενώ υπάρχει μικρή διαφορά ανάμεσα στα ΩΤ και στις πρόδρομες μορφές (~3.3 έναντι ~3). Στην δεύτερη περίπτωση η διαφορά πιθανόν να οφείλεται στον υδρόφοβο πυρήνα (h-domain) του ΣΠ. Τέλος εάν συγκρίνουμε την μέση πιθανότητα σχηματισμού συσσωματώσεων ανά αμινοξύ παρατηρούμε ότι τα ΩΤ είναι συνολικά λιγότερο πιθανό να σχηματίσουν συσσωματώματα από

ότι οι κυτταροπλασματικές αλληλουχίες (μέγιστη τιμή μέσης πιθανότητας σχηματισμού συσσωματώσεων ίση με  $\sim 11.3$  έναντι  $\sim 98.9$ ) ενώ το ΣΠ φαίνεται να συνεισφέρει στη αύξηση της πιθανότητας (μέγιστη μέση πιθανότητα  $\sim 24.9$  για τις πρόδρομες μορφές; μαύρη καμπύλη).



**Εικόνα 3.10 – Σύνοψη περιοχών που προάγουν τον σχηματισμό συσσωματώσεων**

Προβλέψαμε περιοχές με έντονη προδιάθεση για σχηματισμό συσσωματώσεων χρησιμοποιώντας το βιοπληροφορικό εργαλείο AGGRESCAN, το οποίο χρησιμοποιεί ένα σύστημα βαθμολόγησης των αλληλουχιών που βασίζεται σε μια κλίμακα βαθμονόμησης των αμινοξέων ανάλογα με την συχνότητα που συναντώνται σε τέτοιες περιοχές. Α) Αριθμός περιοχών με υψηλή προδιάθεση τάση για σχηματισμό συσσωματώσεων (hot spots) ανά 100 αμινοξέα Β) Μέση πιθανότητα σχηματισμού συσσωματώσεων στην αλληλουχία.

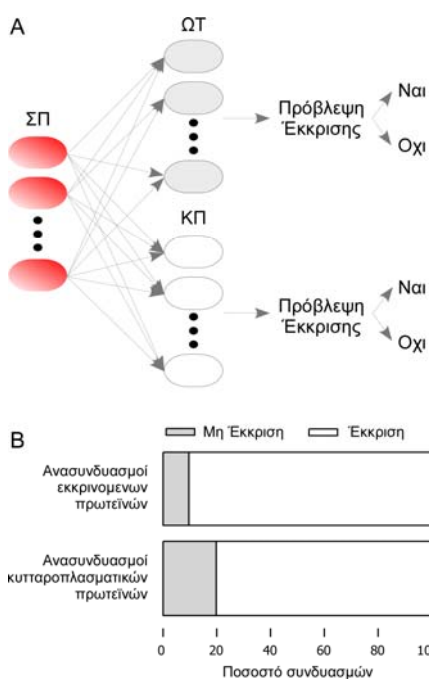
Συγκρίνουμε τρεις κατηγορίες αλληλουχιών: ΩΤ, πρόδρομες μορφές και κυτταροπλασματικές πρωτεΐνες. Ο μέσος αριθμός περιοχών είναι μεγαλύτερος για τις κυτταροπλασματικές πρωτεΐνες σε σύγκριση με τις πρόδρομες μορφές ( $\sim 3.7$  έναντι  $\sim 3$ ), ενώ υπάρχει μικρή διαφορά ανάμεσα στα ΩΤ και στις πρόδρομες μορφές ( $\sim 3.3$  έναντι  $\sim 3$ ). Στην δεύτερη περίπτωση η διαφορά πιθανόν να οφείλεται στο ΣΠ και συγκεκριμένα στον υδρόφοβο πυρήνα του (h-domain). Τέλος εάν συγκρίνουμε την μέση πιθανότητα σχηματισμού συσσωματώσεων ανά αμινοξύ παρατηρούμε ότι τα ΩΤ είναι συνολικά λιγότερο πιθανό να σχηματίσουν συσσωματώματα σε σχέση με τις κυτταροπλασματικές αλληλουχίες (μέγιστη τιμή μέσης πιθανότητας σχηματισμού συσσωματώσεων ίση με  $\sim 11.3$  έναντι  $98.9$ ) ενώ παρουσία του ΣΠ πεπτιδίου αυξάνεται αυτή η πιθανότητα ( $\sim 24.9$ ).

### 3.10 *In silico* συνδυασμοί σηματοδοτικών πεπτιδίων και ώριμων τμημάτων

Όπως ήδη αναφέραμε στην ενότητα 3.6 οι εκκρινόμενες πρωτεΐνες αποτελούνται από δύο περιοχές με διαφορετική ικανότητα αναδίπλωσης, το ΣΠ (θέσεις 1-23) και το πρώιμο ΩΤ (θέσεις 24-74). Η παρατήρηση αυτή εγείρει διάφορα ερωτήματα όπως: ποιος είναι ο ρόλος της κάθε περιοχής στην διαδικασία της έκκρισης, πώς τα δύο αυτά σήματα συνδυάζονται και αλληλεπιδρούν μεταξύ τους για βελτιστοποίηση της έκκρισης και αν υπάρχουν χαρακτηριστικά στην μια περιοχή που σχετίζονται αντίστοιχα με κάποια άλλα τις δεύτερης.

Αποφασίσαμε να διερευνήσουμε την συσχέτιση ανάμεσα στα χαρακτηριστικά των ΣΠ και των αντίστοιχων ΩΤ και για το λόγο αυτό «ανακατέψαμε» τα ΣΠ των πρωτεϊνών. Συνδυάσαμε *in silico* 468 ΣΠ με όλα τα αντίστοιχα ΩΤ και εκτιμήσαμε την ικανότητα έκκρισης αυτών των συνδυασμών χρησιμοποιώντας το μοντέλο πρόδρομης μορφής (Εικόνα 3.4, Πίνακας 3.4) Αντίστοιχοι συνδυασμοί (fusions) έγιναν μεταξύ ΣΠ και κυτταροπλασματικών πρωτεϊνών.

Η ανάλυση έδειξε ότι περίπου 79% των κυτταροπλασματικών πρωτεϊνών πρόκειται να εκκριθούν εάν τους προσθέσουμε ένα ΣΠ. Ένα σημαντικό ποσοστό της τάξης του 20% δεν θα οδηγηθεί σε έκκριση (Εικόνα 3.11).



**Εικόνα 3.11 – *In silico* συνδυασμοί ΣΠ με ΩΤ και κυτταροπλασματικές αλληλουχίες**

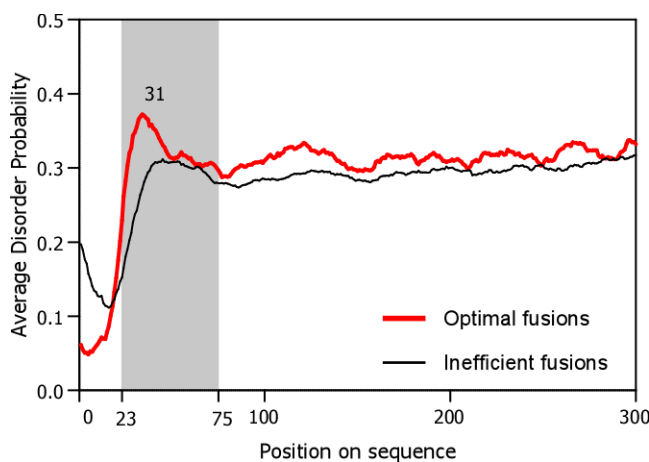
Α) Συνδυάσαμε τα ΣΠ 468 εκκρινόμενων πρωτεϊνών του *E.coli* με όλα τα αντίστοιχα ΩΤ αλλά και τα προσθέσαμε μπροστά από τις αλληλουχίες 1899 κυτταροπλασματικών πρωτεϊνών (ΚΠ). Στη συνέχεια χρησιμοποιήσαμε το μοντέλο πρόδρομης μορφής για να προβλέψουμε εάν οι συνδυασμοί αυτοί πρόκειται να εκκριθούν ή να παραμείνουν στο κυτταρόπλασμα. Β) υπολογίσαμε το ποσοστό των συνδυασμών που πρόκειται να εκκριθούν. Το 9% των συνδυασμών σηματοδοτικών πεπτιδίων με άλλα ώριμα τμήματα (εκτός του δικού τους) προβλέπεται προβληματικό στην έκκριση. Αντίθετα στην αντίστοιχη περίπτωση όπου κυτταροπλασματικές πρωτεΐνες (ΚΠ) συγχωνεύτηκαν με ΣΠ ένα μεγαλύτερο ποσοστό πρωτεϊνών (~20%) προβλέπεται ότι δεν θα καταφέρει να εκκριθεί παρά την παρουσία του ΣΠ.

Στην περίπτωση των εκκρινόμενων πρωτεϊνών όπου τα ΣΠ συνδυάστηκαν με άλλα ΩΤ παρατηρήσαμε επίσης ότι ένα μέρος των συνδυασμών (~9%) προβλέπεται ανίκανο προς έκκριση. Αυτό ενισχύει την θεωρία ότι η ύπαρξη του ΣΠ δεν είναι ικανή συνθήκη για έκκριση και ότι θα πρέπει να υπάρχει κάποια διαδικασία βελτιστοποίησης του συνδυασμού των χαρακτηριστικών του ΣΠ και του πρώιμου ΩΤ.

Στη συνέχεια διερευνήσαμε το συνδυασμό ΣΠ και ΩΤ από την σκοπιά του ενεργειακού περιεχομένου. Επιλέξαμε τους *in silico* συνδυασμούς με τις μεγαλύτερες και μικρότερες βαθμολογίες (scores) με βάση το μοντέλο διαχωρισμού πρόδρομης μορφής. Ονομάσαμε τους

πρώτους βέλτιστους (optimal) και τους δεύτερους χειρότερους (inefficient) συνδυασμούς και υπολογίσαμε το μέσο προφίλ αταξίας για κάθε περίπτωση (Εικόνα 3.12).

Παρατηρούμε ότι οι βέλτιστοι συνδυασμοί παρουσιάζουν μια αυξημένη πιθανότητα αταξίας περίπου στα πρώτα 24 αμινοξέα του ΩΤ (θέση 24 με 47) με το μέγιστο στην θέση 31 (36%; Θέση +8 του ΩΤ) αλλά και πολύ σταθερά ΣΠ (πιθανότητα αταξίας <1%). Αντίθετα στην περίπτωση των χειρότερων συνδυασμών (inefficient) το πρώιμο ΩΤ προβλέπεται να έχει μικρότερη αταξία (μέγιστο στο 31%) ενώ το ΣΠ να είναι πιο ασταθές (πιθανότητα αταξίας από 11% μέχρι 20%).



**Εικόνα 3.12 Προφίλ αταξίας των βέλτιστων και χειρότερων συνδυασμών.**

Συνδυάσαμε *in silico* τα ΣΠ με όλα τα αντίστοιχα ΩΤ και επιλέξαμε τους συνδυασμούς με τις μεγαλύτερες (optimal) και μικρότερες (inefficient) βαθμολογίες (scores) με βάση το μοντέλο διαχωρισμού πρόδρομης μορφής. Στη συνέχεια υπολογίσαμε και συγκρίναμε το μέσο προφίλ αταξίας ανά κατηγορία. Παρατηρούμε ότι οι βέλτιστοι συνδυασμοί παρουσιάζουν μια αυξημένη

αταξία στα πρώτα 24 αμινοξέα του ΩΤ (θέση +24 με +47) με το μέγιστο στην θέση 31 (+8 του ΩΤ) αλλά και πολύ σταθερά ΣΠ. (πιθανότητα αταξίας <10%). Αντίθετα στην περίπτωση των χειρότερων συνδυασμών (inefficient) το πρώιμο ΩΤ προβλέπεται να έχει μικρότερη αταξία ενώ το ΣΠ μεγαλύτερη. Συνεκτιμώντας τις δύο αυτές παρατηρήσεις μπορούμε να υποθέσουμε ότι η ενεργειακή διαφορά ανάμεσα στις δύο περιοχές (ΩΤ και ΣΠ) ενδέχεται να παίζει ρόλο στην διαδικασία της έκκρισης.

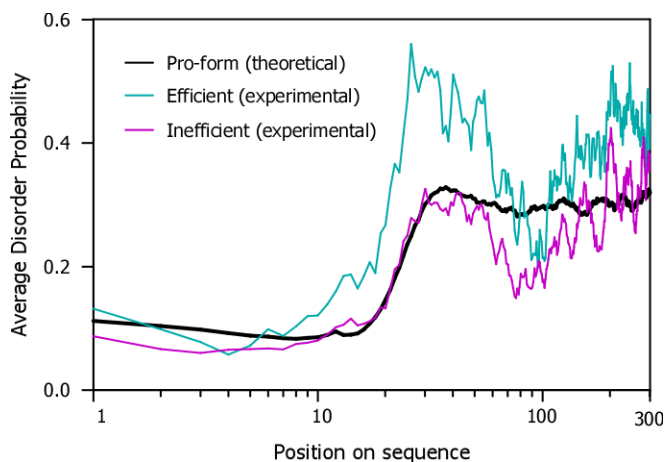
Οι δύο αυτές παρατηρήσεις μπορούν να συνοψιστούν στο ότι η ενεργειακή διαφορά ανάμεσα στο πρώιμο ΩΤ και το ΣΠ είναι μεγαλύτερη στους βέλτιστους προς έκκριση συνδυασμούς. Η παρατήρησή μας οδηγεί στην υπόθεση ότι η διαφορά ενέργειας ανάμεσα στις δύο αυτές περιοχές ενδέχεται να παίζει κάποιο μηχανιστικό ρόλο στην διαδικασία της έκκρισης. Συνδυάζοντας την πειραματική παρατήρηση ότι οι δύο αυτές περιοχές προσδένονται σε ανεξάρτητα σημεία πάνω στην πρωτεΐνη SecA (Sardis et al, in preparation) τότε ενδεχομένως το πρώιμο ΩΤ των εκκρινόμενων περιοχών να λειτουργεί ως ένας ευέλικτος βρόγχος που μεσολαβεί για την βέλτιστη πρόσδεση του ΣΠ αλλά και των υδρόφοβων περιοχών του ΩΤ που ακολουθούν.

### 3.11 Πειραματική επιβεβαίωση των μοντέλων διαχωρισμού

Τα μοντέλα διαχωρισμού που εκπαιδεύτηκαν στο GEMS έχουν πολύ υψηλή απόδοση (AUC) σύμφωνα με τεχνικές ενδοπιστοποίησης (cross-validation) στα δεδομένα εκπαίδευσης αλλά και σε άγνωστα δεδομένα (σύνολο αξιολόγησης) (Πίνακας 3.4).

Για την πειραματική επαλήθευση των μοντέλων αναζητήσαμε στην βιβλιογραφία παλαιότερες δημοσιεύσεις όπου για εκκρινόμενες πρωτεΐνες έχει μετρηθεί η απόδοση της έκκρισης (δηλ. ποσοστό της ποσότητας της πρωτεΐνης που καταφέρνει να εκκριθεί).

Συλλέξαμε πειραματικές μετρήσεις για 120 παράγωγα αγρίου τύπου εκκρινόμενων πρωτεϊνών οι οποίες είχαν υποστεί μεταλλάξεις στην περιοχή του ΣΠ η/και του πρώιμου ΩΤ (Πίνακας 6.8). Πιο αναλυτικά υπάρχουν 37 παράγωγα με μεταλλάξεις στο ΣΠ, 73 με μεταλλάξεις στο ΩΤ και 10 με μεταλλάξεις και στις δύο περιοχές. Οι τύποι μεταλλάξεων ποικίλουν από διαγραφές, αντικαταστάσεις και εισαγωγές ενώ η διαθεσιμότητα των δεδομένων ήταν διαφορετική από έτοιμα γραφήματα έως ακατέργαστα δεδομένα (δες ενότητα 6.2.9).



**Εικόνα 3.13 – Πειραματικά δεδομένα, μέση πιθανότητα αταξίας**

Έπειτα από βιβλιογραφική αναζήτηση συγκεντρώσαμε 120 παράγωγα αγρίου τύπου εκκρινόμενων πρωτεϊνών που έχουν υποστεί μεταλλάξεις στην περιοχή του ΣΠ ή/και του ΩΤ και για τα οποία έχει μετρηθεί η ποσότητα έκκρισης. Χωρίσαμε τα παράγωγα σε δύο κατηγορίες αυτά με χαμηλό ποσοστό έκκρισης (inefficient; μικρότερο από 10% επί της αγρίου τύπου πρωτεΐνης) και αυτά μη υψηλό ποσοστό έκκρισης (efficient) και υπολογίσαμε το μέσο προφίλ αταξίας. Παρατηρούμε ότι οι πρωτεΐνες με προβληματική έκκριση έχουν το ίδιο σταθερό ΣΠ σε σχέση με αυτές με μεγαλύτερη έκκριση, έχουν όμως χαμηλής αταξίας πρώιμο ΩΤ.

Για τα συγκεκριμένα παράγωγα έχει μετρηθεί πειραματικά η ποσότητα της εκκρινόμενης πρωτεΐνης ως ποσοστό επί της ποσότητας έκκρισης των αγρίου τύπου πρωτεϊνών (secretion efficiency). Τα ποσοστά κυμαίνονται από 0% έως και 105% (μεγαλύτερη έκκριση από την αγρίου τύπου πρωτεΐνη). Θεωρήσαμε ως ικανά προς έκκριση τα παράγωγα με ποσοστό μεγαλύτερο από 10% επί του ποσού έκκρισης της αρχικής πρωτεΐνης. Στη συνέχεια υπολογίσαμε την μέση

πιθανότητα αταξίας για αυτές της δύο κατηγορίες. Τα αποτελέσματα δείχνουν ότι τα παράγωγα που εκκρίνονται σε ικανοποιητικές ποσότητες φαίνεται να έχουν μεγαλύτερη πιθανότητα αταξίας στην περιοχή του πρώιμου ΩΤ.

### 3.12 Σύγκριση απόδοσης με άλλα βιοπληροφορικά εργαλεία

Συγκρίναμε την απόδοση των μοντέλων ώριμης και πρόδρομης μορφής και το μοντέλο αταξίας με τρία αντίστοιχα βιοπληροφορικά εργαλεία: SignalP 4.1 (Petersen et al, 2011), LipoP (Juncker et al, 2003) και Phobius (Juncker et al, 2003; Kall et al, 2004). Τα δεδομένα που χρησιμοποιήθηκαν για τον υπολογισμό της απόδοσης ήταν τα αντίστοιχα σύνολα εκπαίδευσης και αξιολόγησης (train and test sets) που χρησιμοποιήσαμε στην δική μας ανάλυση, όπως επίσης και τις πειραματικές μετρήσεις που συγκεντρώσαμε από την βιβλιογραφία (δες ενότητα 6.2.10).

**Πίνακας 3.4 – Σύγκριση των μοντέλων διαχωρισμού με άλλα βιοπληροφορικά εργαλεία**

Στον πίνακα αυτό συνοψίζουμε την απόδοση των τριών βασικών μοντέλων (μοντέλα ώριμης/πρόδρομης μορφής και μοντέλο αταξίας) και αντιπαραθέτουμε με την απόδοση τριών βιοπληροφορικών εργαλείων: SignalP 4.1, προβλέπει ύπαρξη τύπου πεπτιδίου σήματος, το LipoP το οποίο προβλέπει τύπου I και II πεπτίδια σήματα και το Phobius το οποίο εκτός από διαμεμβρανικές έλικες προβλέπει και ΣΠ. Δεδομένα: Σύνολο εκπαίδευσης, (80% του συνόλου των δειγμάτων), Σύνολο αξιολόγησης (το υπόλοιπο 20% του συνόλου των δειγμάτων), Πειραματικά δεδομένα: 120 παράγωγα εκκρινόμενων πρωτεϊνών για τα οποία έχει μετρηθεί το ποσό έκκρισης (Πίνακας 6.8).

Βιοπληροφορικό Εργαλείο/ Μοντέλο	Σύνολο Εκπαίδευσης	Σύνολο Αξιολόγησης	Πειραματικά Δεδομένα
<i>SignalP 4.1</i>	99.04	98.04	51.64
<i>LipoP</i>	99.97	99.82	61.39
<i>Phobius</i>	99.34	99.54	72.08
<i>Μοντέλο πρόδρομης μορφής</i>	98.57	97.04	72.61
<i>Μοντέλο ώριμης μορφής</i>	90.37	74.23	79.58

Τα βιβλιογραφικά δεδομένα θεωρούμε ότι έχουν μεγαλύτερη βαρύτητα στην συγκεκριμένη σύγκριση καθώς είναι πραγματικές τιμές της ποσότητας έκκρισης μετρημένες πειραματικά με διαφορετικές τεχνικές, για αγρίου τύπου αλλά και μεταλλαγμένες πρωτεΐνες. Το ποσό της έκκρισης



το οποίο εκφράζεται ποσοστιαία ως προς την αγρίου τύπου πρωτεΐνη παίρνει τιμές από 0% έως και πάνω από το 105% (πάνω από την έκκριση της αγρίου τύπου πρωτεΐνης). Θεωρήσαμε ως μη αποτελεσματικά εκκρινόμενες πρωτεΐνες (δηλ. κυτταροπλασματικές) όσες έχουν μετρηθεί με ποσό έκκρισης λιγότερο από 10% σε σχέση με την αγρίου τύπου πρωτεΐνη.

### 3.13 Πρόβλεψη εκκριτικών πρωτεϊνών σε άλλα Gram<sup>-</sup> και Gram<sup>+</sup> βακτήρια

Τα μοντέλα ώριμης και πρόδρομης μορφής εκπαιδεύτηκαν χρησιμοποιώντας εμπειριστικά σύνολα εκκρινόμενων πρωτεϊνών του συστήματος Sec και κυτταροπλασματικών πρωτεϊνών του βακτηρίου *E.coli* (δες ενότητα 6.2.6) και δεν επεκτείνονται σε πρωτεΐνες από άλλα Gram<sup>-</sup> βακτήρια. Στατιστικές αναλύσεις έχουν δείξει ότι το πρώιμο ΩΤ των εκκρινόμενων πρωτεϊνών στα Gram<sup>+</sup> βακτήρια είναι θετικά φορτισμένο ενώ στα Gram<sup>-</sup> αρνητικά φορτισμένο (Kajava et al, 2000). Ίσως αυτός είναι και ο λόγος που σε κάποια BE η πρόβλεψη των Gram<sup>+</sup> και Gram<sup>-</sup> εκκρινόμενων πρωτεϊνών γίνεται με ξεχωριστά εκπαιδευμένους αλγόριθμους (Gardy et al, 2003; Petersen et al, 2011).

**Πίνακας 3.5 – Πρόβλεψη σε άλλα βακτήρια**

Συλλέξαμε τα πρωτεϊνώματα 10 Gram<sup>+</sup> και 25 Gram<sup>-</sup> βακτηρίων από το Uniprot (Dimmer et al, 2012) και χρησιμοποιώντας τρία BE προβλέψαμε της εκκρινόμενες πρωτεΐνες τύπου Sec.

Μοντέλο	Gram <sup>-</sup>	Gram <sup>+</sup>
Πρόδρομης μορφής	<b>96.95%</b>	<b>98.16%</b>
Ωριμης μορφής	<b>60.15%</b>	<b>66.89%</b>
Αταξίας	<b>80.86%</b>	<b>89.57%</b>

Για να αποδείξουμε την καθολικότητα των μοντέλων ώριμης και πρόδρομης μορφής, αποφασίσαμε να συγκεντρώσουμε εκκρινόμενες πρωτεΐνες άλλων Gram<sup>-</sup> αλλά και Gram<sup>+</sup> βακτηρίων και να εκτιμήσουμε την απόδοση των μοντέλων σε αυτά στελέχη. Εν τέλει επιλέξαμε 25 Gram<sup>-</sup> και 10 Gram<sup>+</sup> βακτήρια (Πίνακας 6.9) και προσδιορίσαμε τα αντίστοιχα σύνολα των εκκρινόμενων πρωτεϊνών τύπου Sec χρησιμοποιώντας τρία BE (Bagos et al, 2010; Juncker et al, 2003; Petersen et al, 2011) (δες ενότητα 6.2.16). Καταλήξαμε σε ένα σύνολο από 6952 και 1361 εκκρινόμενες πρωτεΐνες από τα Gram<sup>-</sup> και Gram<sup>+</sup> βακτήρια αντίστοιχα.

Από την ανάλυση προέκυψε ότι το μοντέλο πρόδρομης μορφής αποδίδει πολύ καλά και στις δύο περιπτώσεις (Gram<sup>+</sup> και Gram<sup>-</sup>) γεγονός που μπορεί να σημαίνει ότι τα χαρακτηριστικά των ΣΠ του *E.coli* μπορούν να γενικευτούν και σε άλλα βακτήρια ακόμα και Gram<sup>+</sup>. Το μοντέλο

---

ώριμης μορφής αποτυγχάνει να προβλέψει το ίδιο επιτυχημένα τα ώριμα τμήματα των πρωτεϊνών (~60% και ~66%). Αυτό ίσως οφείλεται στο γεγονός ότι τα χαρακτηριστικά των ΩΤ στο βακτήριο *E.coli* ενδέχεται να μην γενικεύονται σε άλλα βακτήρια. Τέλος το μοντέλο αταξίας μπορεί να προβλέψει πολύ πιο σωστά τα ΩΤ των εκκρινόμενων πρωτεϊνών (~80% και ~89% για Gram<sup>-</sup> και Gram<sup>+</sup> βακτήρια αντίστοιχα).

### 3.14 Συζήτηση

Χρησιμοποιώντας μεθόδους μηχανικής μάθησης και την πληροφορία των αμινοξικών αλληλουχιών εκπαιδεύσαμε δύο βασικά μοντέλα, το μοντέλο πρόδρομης και ώριμης μορφής τα οποία διαχωρίζουν με υψηλή απόδοση (~98% και 90%) τις εκκρινόμενες από τις κυτταροπλασματικές αλληλουχίες. Τα μοντέλα προέκυψαν γραμμικά γεγονός που υποδεικνύει ότι τα χαρακτηριστικά που επιλέγονται είναι ανεξάρτητα μεταξύ τους. Το συμπέρασμα παραμένει να εξεταστεί περαιτέρω με αλγόριθμους αιτιότητας (causality).

Το μοντέλο πρόδρομης μορφής εκπαιδεύτηκε χρησιμοποιώντας τα εμπειριστατωμένα σύνολα τύπου I (περιπλασμικές πρωτεΐνες και πρωτεΐνες της εξωτερικής μεμβράνης) και τύπου II (λιποπρωτεΐνες) εκκρινόμενων πρωτεϊνών και προβλέπει. Μπορεί να προβλέψει και τους δύο τύπους εκκριτικών πρωτεϊνών με αντίστοιχα υψηλή απόδοση με αυτή που έχουν πιο εξειδικευμένα BE (~99% LipoP και SignalP 4.0; Πίνακας 3.4).

Το μοντέλο ώριμης μορφής προβλέπει τα ΩΤ των εκκρινόμενων πρωτεϊνών σε ποσοστό ~90% παρόλο που το βασικό εκκριτικό σήμα, το ΣΠ, απουσιάζει. Τα χαρακτηριστικά που επιλέγονται στα ΩΤ είναι διάσπαρτα όμως ορισμένα από αυτά μπορούν να συνοψιστούν σε απλούς κανόνες όπως αποφυγή Αργινίνης και υδρόφοβων αμινοξέων (LIV) στις θέσεις +1 με +23, Επίσης ο συνδυασμός Λυσίνης (K) με Ασπαρτικό οξύ (D) φαίνεται ότι επιλέγεται έναντι Αργινίνης (R) σε συνδυασμό με Γλουταμινικό οξύ (E) στις εκκρινόμενες πρωτεΐνες καθώς επίσης και Προλίνες στις θέσεις +14 μέχρι +18 και πολικά κατάλοιπα (Q,S,T) από την θέση +10 και μετά. Η αριθμητική αναπαράσταση και ο το τρόπος απεικόνισης επιβάλλει την στοίχιση των αλληλουχιών στο αμινοτελικό άκρο (σημείο αποκοπής). Η αποτυχία εντοπισμού αμινοξικών μοτίβων ίσως οφείλεται στο γεγονός ότι «ισοπεδώνονται» λόγω λανθασμένης στοίχισης. Σωστή στοίχιση των αλληλουχιών ενδέχεται να μην υπάρχει καθώς τα αμινοξικά μοτίβα μπορεί να είναι σε διαφορετικές θέσεις σε κάθε αλληλουχία.

Διερευνήσαμε την αμινοξική σύσταση των εκκρινόμενων πρωτεϊνών και πως αυτή καθορίζει την ικανότητα αναδίπλωσης τους. Τα ΩΤ των εκκρινόμενων πρωτεϊνών παρουσιάζουν μεγαλύτερη πιθανότητα αταξίας και αυτό αντανακλάται από τη αμινοξική τους σύσταση. Το συμπέρασμα αυτό ταιριάζει με την παρατήρηση για τα διάσπαρτα επιλεγμένα χαρακτηριστικά με βάση το μοντέλο ώριμης μορφής (Εικόνα 3.4B) τα οποία είχαμε προσπαθήσει να συνοψίσουμε σε πολικά κατάλοιπα, Ασπαρτικό οξύ, Λυσίνη και Προλίνες. Το ΣΠ παρουσιάζει πολύ χαμηλή

πιθανότητα αταξίας (~10%) σε σχέση με το πρώιμο ΩΤ (~31%). Η παρουσία του ΣΠ στις πρόδρομες μορφές φαίνεται να σταθεροποιεί το πρώιμο ΩΤ καθώς όταν το ΣΠ απουσιάζει η πιθανότητα αταξίας πλησιάζει μέχρι και το 40%. Τα συμπεράσματα αυτά συνηγορούν σε ένα μηχανιστικό ρόλο της περιοχής του πρώιμου ΩΤ στην διαδικασία της έκκρισης. Το ΣΠ είναι γνωστό ότι προσδένεται σε διαφορετική θέση από αυτήν που προσδένονται τα σήματα στην περιοχή του ΩΤ. Ενδεχομένως η περιοχή του πρώιμου ΩΤ (θέσεις +1 μέχρι +30) να τελούν το ρόλο ενός «συνδετήρα» ανάμεσα σε αυτά τα δύο σήματα που καθορίζει την βέλτιστη τοποθέτηση τους πάνω στην SecA. Αυτό θα εξηγούσε γιατί δεν επιτυγχάνεται πάντα έκκριση όταν συνδυάζονται διαφορετικά ΣΠ και ΩΤ. Η υπόθεση αυτή μένει να επιβεβαιωθεί με πειραματικές μεθόδους.

Στη συνέχεια εκπαιδεύσαμε το μοντέλο αταξίας που βασίστηκε στην πληροφορία της πιθανότητας αταξίας, εισάγοντας στην εκπαίδευση 20 μεταβλητές που αντιστοιχούν στην συνολική ενεργειακή συνεισφορά των 20 αμινοξέων. Το μοντέλο αταξίας έχει μεγαλύτερη απόδοση στα πειραματικά δεδομένα ίσως αυτό εξηγείται επειδή οι μεταλλάξεις που έχουν γίνει στην περιοχή του ΩΤ (εισαγωγή θετικών φορτίων) έχει ως αποτέλεσμα την διατάραξη του ενεργειακού περιεχομένου της περιοχής αυτής (Εικόνα 3.12). Επίσης το μοντέλο αυτό εξηγεί την αμινοξική σύσταση που επιτάσσει η εξέλιξη για τις εκκρινόμενες πρωτεΐνες. Δύο εναλλακτικοί συνδυασμοί αμινοξέων που προάγουν την εσωτερική αταξία επιλέγονται. Ο πρώτος με συνδυασμό κυρίως Γλουταμίνης (Q), Θρεονίνης (T) και Λυσίνης (K) και ένας δεύτερος με Προλίνη (P), Ασπαρτικό οξύ (D) και Γλουταμίνη (Q).

Στη συνέχεια πραγματοποιήσαμε *in silico* συνδυασμούς ΣΠ με ΩΤ και προβλέψαμε αν θα εκκριθούν χρησιμοποιώντας το μοντέλο πρόδρομης μορφής. Ένα ποσοστό των συνδυασμών ΣΠ με άλλα ΩΤ (~ 9%) προβλέφθηκε ότι δεν πρόκειται να εκκριθεί και αυτό συμφωνεί με τις μέχρι τώρα πειραματικές ενδείξεις όπου το ΩΤ της πρωτεΐνης PhoA εάν συνδυαστεί με άλλα ΣΠ εκκρίνεται με διαφορετική απόδοση από καθόλου έως και πολύ περισσότερο από την αγρίου τύπου πρωτεΐνη (Orfanoudaki et al, in preparation).

Έπειτα διαχωρίσαμε τους *in silico* συνδυασμούς με βάση την βαθμολογία από το μοντέλο πρόδρομης μορφής, σε βέλτιστες και μη εκκρινόμενες αλληλουχίες και υπολογίσαμε το μέσο προφίλ αταξίας ανά κατηγορία. Η ανάλυση μας οδήγησε στο συμπέρασμα ότι στους βέλτιστους συνδυασμούς το ΣΠ παρουσιάζει μεγάλη ευστάθεια (μικρή πιθανότητα αταξίας, <~1%) ενώ το

---

πρώιμο ΩΤ μεγαλύτερη (μέχρι και 36%). Η παρατήρηση μας οδηγεί στο συμπέρασμα ότι η διαφορά ενέργειας ανάμεσα στις δύο αυτές περιοχές ενδέχεται να παίζει κάποιο μηχανιστικό ρόλο στην διαδικασία της έκκρισης.

Τέλος συγκεντρώσαμε εκκριτικές πρωτεΐνες από 34 βακτήρια και υπολογίσαμε την απόδοση των τριών βασικών μοντέλων στην πρόβλεψη εκκρινόμενων πρωτεϊνών τύπου Sec Σε άλλα βακτηριακά στελέχη. Το μοντέλο ώριμης μορφής αποτυγχάνει να προβλέψει το ίδιο καλά τα ΩΤ των πρωτεϊνών (~60% και ~66%). Αυτό ίσως οφείλεται στο γεγονός ότι τα χαρακτηριστικά των ΩΤ στο βακτήριο *E.coli* ενδέχεται να μην γενικεύονται σε άλλα βακτήρια. Όμως το μοντέλο αταξίας μπορεί να προβλέψει σε μεγαλύτερο ποσοστό τα ΩΤ των εκκρινόμενων πρωτεϊνών (~80% και ~89% για Gram-+ και Gram+ βακτήρια αντίστοιχα). Η βελτιωμένη απόδοση του μοντέλου αταξίας ίσως μπορεί να αποδοθεί στο ότι διαχωρίζει τις πρωτεΐνες με βάση το ενεργειακό τους περιεχόμενο και όχι τα συγκεκριμένα αμινοξέα που επιλέγονται. Σύμφωνα με αυτήν την υπόθεση η εσωτερική αταξία των εκκρινόμενων πρωτεϊνών θα μπορούσε να είναι κοινό χαρακτηριστικό ανάμεσα στα βακτήρια ενώ εκάστοτε επιλογή αμινοξέων να διαφέρει.

Συνοψίζοντας, για πρώτη φορά παρουσιάσαμε δεδομένα που αποδεικνύουν ότι η αμινοξική σύσταση των εκκρινόμενων πρωτεϊνών διαφέρει από αυτή των κυτταροπλασματικών πρωτεϊνών και καθορίσαμε τους κανόνες που την υπαγορεύουν. Στις εκκρινόμενες πρωτεΐνες επιλέγονται αμινοξέα που προάγουν την εσωτερική αταξία ενώ στις κυτταροπλασματικές πρωτεΐνες επιλέγονται υδρόφοβα αμινοξέα τα οποία σχηματίζουν το υδρόφοβο πυρήνα των πρωτεϊνών σε συνδυασμό με αρνητικά και θετικά φορτισμένα κατάλοιπα τα οποία είναι υδρόφιλα και βρίσκονται στην επιφάνεια των μορίων.

Καταλήγουμε στο συμπέρασμα ότι οι εκκρινόμενες πρωτεΐνες στην στατιστική πλειοψηφία τους ανήκουν στην κατηγορία των πρωτεϊνών με εσωτερική αταξία (IDPs: Intrinsically disordered proteins). Ενδεικτικό παράδειγμα αποτελούν οι παράλογες πρωτεΐνες PpiA και PpiB, δύο ένζυμα όπου το πρώτο εκκρίνεται στο περίπλασμα ενώ το δεύτερο παραμένει στο κυτταρόπλασμα. Πειράματα αποδεικνύουν ότι ο ρυθμός αναδίπλωσης της περιπλασματικής (PpiA) πρωτεΐνης είναι σημαντικά πιο αργός (Chatzi et al, in preparation).

Η δομική αταξία σε συνδυασμό με την ταυτόχρονη ύπαρξη σύντομων υδρόφοβων περιοχών οι οποίες έχουν ρόλο στόχευσης στην πρωτεΐνη κινητήρα SecA και τα οποία είναι εναλλάξιμα (Chatzi et al, in preparation). Η ταυτόχρονη ικανοποίηση των δύο χαρακτηριστικών

---

μπορεί να ισχύει αν υποθέσουμε ότι μικρές περιοχές με υδρόφοβο χαρακτήρα εναλλάσσονται με περιοχές μεγάλης αταξίας. Η υπόθεση αυτή μένει να εξεταστεί με υπολογιστικές και μαθηματικές μεθόδους.

Παρουσιάσαμε μια ανάλυση η οποία ανέδειξε για πρώτη φορά τα αμινοξικά χαρακτηριστικά που εμπεριέχονται στα πρώιμα ΩΤ των εκκρινόμενων πρωτεϊνών καθώς και τις δομικές ιδιότητες που απορρέουν από αυτά. Το γεγονός ότι η πρώιμη περιοχή του ΩΤ ενδέχεται να καθορίζει την έκκριση σε συνδυασμό με τα χαρακτηριστικά του ΣΠ έχει σημειωθεί σε παλαιότερες αναλύσεις με πρώτη βασική παρατήρηση ότι το ΣΠ δεν είναι ικανό από μόνο του να οδηγήσει σε έκκριση οποιαδήποτε πρωτεΐνη (Boyd et al, 1990; Kadonaga et al, 1984; Moreno et al, 1980). Στη συνέχεια έγιναν προσπάθειες να διευκρινιστούν τα συγκεκριμένα αμινοξέα αλλά και οι θέσεις στην περιοχή του ΩΤ που βελτιώνουν ή χειροτερεύουν το τελικό ποσό της έκκρισης (Kajava et al, 2000; Kim et al, 2000; Nesmeyanova et al, 1997; Summers et al, 1989). Στατιστική ανάλυση του φορτίου του πρώιμου ΩΤ κατέληξε στο συμπέρασμα ότι υπάρχει μια προδιάθεση για αρνητικά φορτισμένα κατάλοιπα στις πρώτες θέσεις (Kajava et al, 2000) και συσχέτισε αυτά τα χαρακτηριστικά με την εκκριτική διαδικασία. Στη παρούσα ανάλυση προτείνουμε ότι το αρνητικό φορτίο του πρώιμου ΩΤ οφείλεται κατά βάση στο σύνολο των λιποπρωτεϊνών και αφορά όξινα κατάλοιπα που σχετίζονται με τον μηχανισμό της μετέπειτα διαλογής τους (δες ενότητα 3.2). Οι μετέπειτα πειραματικές προσεγγίσεις αν και ανέδειξαν κάποια προβληματικά αμινοξέα στην περιοχή των πρώτων 1-18 αμινοξέων (όπως βασικά και υδρόφοβα κατάλοιπα), στην ουσία απέτυχαν να περιγράψουν με σαφή τρόπο την φύση της συνεργατικότητας μεταξύ των δύο αυτών σημάτων. Συλλογικά η ανάλυση μας κατέληξε για πρώτη φορά σε ένα μοντέλο που εξηγεί την συσχέτιση ανάμεσα στα χαρακτηριστικά του ΣΠ και του ΩΤ. Το μοντέλο αυτό προτείνει ότι το ΣΠ και το ΩΤ κυμαίνονται σε διαφορετικά επίπεδα αταξίας τα οποία θα πρέπει να συνδυαστούν μεταξύ τους με βέλτιστο τρόπο.

Είναι σημαντικό επίσης να αναφέρουμε ότι τα μοντέλα που αναπτύχθηκαν στην παρούσα μελέτη μπορούν να επεκταθούν και σε άλλα βακτήρια (δες ενότητα 3.13) ενώ συγκεκριμένα το μοντέλο αταξίας μπορεί να προβλέψει με ικανοποιητική απόδοση τα ΩΤ των εκκρινόμενων πρωτεϊνών στα Gram<sup>+</sup> βακτήρια γεγονός που μας οδηγεί στο συμπέρασμα ότι παρόλο που επιλέγονται διαφορετικού είδους αμινοξέα ο μηχανισμός που απορρέει από αυτά φαίνεται να είναι παρόμοιος. Κατά συνέπεια η παρούσα ανάλυση περιγράφει με σαφή και συνεπυγμένο τρόπο τα χαρακτηριστικά των εκκρινόμενων πρωτεϊνών και αναμένεται να χρησιμοποιηθεί σαν εργαλείο για

---

την ανάπτυξη μοντέλων παλινδρόμησης (regression models) τα οποία να προβλέπουν με ακρίβεια το ποσό της έκκρισης. Τα μοντέλα αυτά αναμένεται να έχουν σημαντικές εφαρμογές στο χώρο της βιομηχανίας και συγκεκριμένα στην παραγωγή ετερόλογων πρωτεϊνών. Προς αυτήν την κατεύθυνση είναι σημαντικό να αναπτυχθούν πιο εξειδικευμένα εργαλεία τα οποία: επιλέγουν το βέλτιστο ΣΠ ανά δεδομένο ΩΤ, δοκιμάζοντας *in silico*, πιθανές σημειακές μεταλλάξεις που αυξάνουν το προβλεπόμενο ποσό έκκρισης των ανασυνδυασμένων πρωτεϊνών.

Τέλος, εκμεταλλευόμενοι την εμπειριστατωμένη ταξινόμηση του πρωτεϊνώματος του *E.coli* (ΚΕΦΑΛΑΙΟ 3, (Orfanoudaki et al, 2014)) είναι ενδιαφέρον να διερευνήσουμε κατά πόσο μπορεί να εφαρμοστεί η μέθοδος υπολογισμού της ενέργειας αλληλεπίδρασης και στις υπόλοιπες υποκατηγορίες των εκκρινόμενων πρωτεϊνών. Ένα υποσύνολο πρωτεϊνών με ιδιαίτερες ιδιότητες, αποτελούν οι περιφερικές πρωτεΐνες (ΚΕΦΑΛΑΙΟ 4) για τις οποίες δεν έχουν βρεθεί μέχρι σήμερα κοινά χαρακτηριστικά και κατά συνέπεια δεν έχουν ακόμα αναπτυχθεί αντίστοιχα βιοπληροφορικά εργαλεία που να μπορούν να τις προβλέψουν.





---

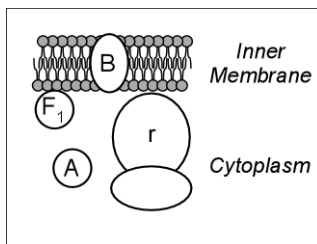
## **ΚΕΦΑΛΑΙΟ 4 Αποτελέσματα – Μελέτη των περιφερικών πρωτεϊνών του *E.coli* με στόχο την χαρτογράφηση των πρωτεϊνικών αλληλεπιδράσεων και των κυτταρικών λειτουργιών στις οποίες συμμετέχουν.**

Τα κύτταρα είναι πολύπλοκες βιολογικές «μηχανές» με πολυεπίπεδη οργάνωση διαφορετικού τύπου βιολογικών μορίων τα οποία «συνεργάζονται» μεταξύ τους για την ολοκλήρωση των κυτταρικών διεργασιών. Οι βασικοί «εργάτες» των κυττάρων είναι οι πρωτεΐνες οι οποίες αναλαμβάνουν διεργασίες όπως: η βιοσύνθεση των μεμβρανών, η διατήρηση του κυτταρικού σκελετού, η κυτταρική διαίρεση, η διακίνηση οργανικών και ανόργανων μορίων, η διακυτταρική επικοινωνία και η παθογένεια.

Οι βιολογικές μεμβράνες συντελούν στην απομόνωση και εξειδίκευση των κυτταρικών λειτουργιών. Στο *E.coli* όπως και σε όλα τα Gram<sup>-</sup> βακτήρια το κυτταρόπλασμα περικλείεται από ένα πολύ-επίπεδο σχηματισμό τον κυτταρικό φάκελο (ΚΦ). Ο ΚΦ αποτελείται από δύο μεμβράνες την πλασματική (ΠΜ) και την εξωτερική (ΕΜ) ανάμεσα εκ των οποίων βρίσκεται το περίπλασμα. Η ΠΜ μεμβράνη απομονώνει τον κυτταροπλασματικό χώρο ενώ ταυτόχρονα αποτελεί μια δυναμική δομή που φιλοξενεί ποικιλία πρωτεϊνικών μορίων συμμετέχοντας σε ζωτικής σημασίας κυτταρικές διεργασίες όπως: η διακίνηση ιόντων, μορίων και μακρομορίων, βιοσύνθεση πολυσακχαριτών, βιοσύνθεση πεπτιδογλυκάνης, μεταβολικά μονοπάτια, μεταβίβαση σημάτων για περιβαλλοντολογικές αλλαγές.

Απαραίτητο βήμα προς την κατανόηση της λειτουργίας ενός κυττάρου αποτελεί η ταξινόμηση των πρωτεϊνών στα διάφορα υποκυτταρικά διαμερίσματα καθώς επίσης και ο καθορισμός των πιθανών αλληλεπιδράσεων μεταξύ τους αλλά και με την ΠΜ.

Οι πρωτεΐνες μπορούν: α) να βρίσκονται ενσωματωμένες στην ΠΜ (διαμεμβρανικές) μέσω διαμεμβρανικών περιοχών (ΔΠ) και β) να αλληλεπιδρούν με την επιφάνεια της ΠΜ (περιφερικές) είτε με άμεσο (αλληλεπίδραση με φωσφολιπίδια) είτε με έμμεσο τρόπο (αλληλεπιδράση με άλλες μεμβρανικές πρωτεΐνες) (Εικόνα 4.1). Περιφερικές πρωτεΐνες μπορούν να υπάρχουν και στις δύο επιφάνειες της ΠΜ και να σχηματίζουν σύμπλοκα πάνω σε αυτή ανάλογα με τις απαιτήσεις του κυττάρου (Dowhan et al, 2008).



**Εικόνα 4.1 – Πρωτεΐνες που σχετίζονται με την πλασματική μεμβράνη**

Σχηματική αναπαράσταση της υποκυτταρικής τοποθέτησης των πρωτεϊνών που βρίσκονται ενσωματωμένες ή αλληλεπιδρούν επιφανειακά με την πλασματική μεμβράνη. Η ονοματολογία ακολουθεί την ορολογία της βάσης δεδομένων STEPdb : A, κυτταροπλασματική; B μεμβρανική πρωτεΐνη; F1: περιφερική πρωτεΐνη της ΠΜ από την πλευρά του κυτταροπλάσματος; r, ριβοσωμική πρωτεΐνη

Οι περιφερικές πρωτεΐνες από την μεριά του κυτταροπλάσματος αποτελούν ένα υποσύνολο μεγάλου ενδιαφέροντος λόγω της αλληλεπίδρασης τους με το κυτταροπλασματικό πρωτεϊνώμα και το πυρηνοειδές αλλά και επειδή εμπλέκονται στα περισσότερα μεταβολικά μονοπάτια του κυττάρου. Αποτελούν διαλυτά μόρια τα οποία προσκολλώνται στην ΠΜ κυρίως μέσω ηλεκτροστατικών αλληλεπιδράσεων και μπορούν να αποκολληθούν από αυτή χρησιμοποιώντας χημικούς παράγοντες όπως το αλάτι, το υψηλό pH και χαστροπικούς παράγοντες (Adelman et al, 1973; Fujiki et al, 1982; Kreibich et al, 1974; Ohlendieck, 2003). Αντίθετα οι μεμβρανικές πρωτεΐνες απαιτούν χρήση αμφίφιλων απορρυπαντικών τα οποία εκτοπίζουν τα φωσφολιπίδια της μεμβράνης και καθιστούν τις πρωτεΐνες διαλυτές σε υδατικό περιβάλλον (Speers et al, 2007).

Σε αντίθεση με το κυτταροπλασματικό πρωτεϊνώμα το οποίο έχει μελετηθεί εκτενώς (Han et al, 2006), το μεμβρανικό και περιφερικό πρωτεϊνώμα είναι ελλιπώς χαρακτηρισμένο. Στην περίπτωση το περιφερικών πρωτεϊνών μάλιστα, δεν υπάρχουν αντίστοιχα BE που να μπορούν να τις προβλέψουν. Αυτό συμβαίνει επειδή είτε δεν αποτελούν αμιγές σύνολο σε σχέση με τα χαρακτηριστικά τους είτε δομικά χαρακτηριστικά, όπως οι αμφίφιλες έλικες, δεν είναι εύκολο να προβλεφθούν (Sapay et al, 2006). Πρόσφατη μελέτη ανιχνεύει με συστηματικό τρόπο τις περιφερικές πρωτεΐνες του βακτηρίου *E.coli* (Papanastasiou et al, 2013). Παλαιότερες αναλύσεις έχουν μόνο περιστασιακά ταυτοποιήσει περιφερικές μονάδες μεμβρανικών συμπλόκων (Huang et al, 2006; Lasserre et al, 2006; Li et al, 2012; Maddalo et al, 2011; Pan et al, 2010; Spelbrink et al, 2005; Stenberg et al, 2005)

Στην παρούσα ανάλυση εστιάζουμε στις περιφερικές πρωτεΐνες της πλασματικής μεμβράνης που βρίσκονται από την πλευρά του κυτταροπλάσματος (F1; Εικόνα 4.1). Βασιζόμενοι στην πρόσφατη ταξινόμηση του *E.coli* (Orfanoudaki et al, 2014)(δες ενότητα 2.10) προχωρήσαμε στην χαρτογράφηση των πρωτεϊνικών αλληλεπιδράσεων και συμπλόκων καθώς και των

---

κυτταρικών λειτουργιών, στις οποίες αυτές συμμετέχουν. Για την χαρτογράφηση αυτή βασιστήκαμε σε επιβεβαιωμένες πρωτεϊνικές αλληλεπιδράσεις που έχουν χαρακτηριστεί σε πρωτεομικές αναλύσεις (Arifuzzaman et al, 2006) (π.χ. πειράματα pull-down, και tandem affinity purification) αλλά και σε βάσεις δεδομένων πρωτεϊνικών αλληλεπιδράσεων (Kerrien et al, 2012). Τέλος υιοθετήσαμε την ταξινόμηση των πρωτεϊνών σε κυτταρικές διεργασίες, Multifun (Karp et al, 2007) και οργάνωση τις περιφερικές πρωτεΐνες σε εννέα κατηγορίες (π.χ. μεταβολισμός και κυτταρική διαίρεση).

Πολλές από αυτές τις πρωτεΐνες είναι οργανωμένες σε λειτουργικές μονάδες, ή αλλιώς πρωτεϊνικά σύμπλοκα (ολιγομερή ή πολυμερή), οι οποίες προσκολλώνται στην ΠΜ με πολλαπλές αλληλεπιδράσεις. Οι περιφερικές πρωτεΐνες καλύπτουν όλο το φάσμα των βιολογικών δραστηριοτήτων ενώ πάνω από τις μισές είναι απαραίτητες για την βιωσιμότητα του κυττάρου. Τα δεδομένα μας υποδηλώνουν ότι ένα αξιοσημείωτο υποσύνολο του κυτταροπλασματικού πρωτεϊνώματος πραγματοποιεί απίστευτα δυναμικές και εκτεταμένες αλληλεπιδράσεις με την ΠΜ οι οποίες μένει να μελετηθούν στο μέλλον και πειραματικά.

#### **4.1 Μελέτη της κυτταρικής λειτουργίας των περιφερικών πρωτεϊνών**

Οι περιφερικές πρωτεΐνες εμπλέκονται σε μεγάλη ποικιλία κυτταρικών αλληλεπιδράσεων που καλύπτουν όλες τις βασικές διεργασίες του κυττάρου (Εικόνα 4.3). Ένας μικρός αριθμός πρωτεϊνών εξακολουθούν να έχουν άγνωστες λειτουργίες (π.χ. YifE και YjgR). Η αντιστοίχιση των πρωτεϊνών σε λειτουργίες δείχνει ότι η πλειοψηφία των περιφερικών πρωτεϊνών (30%) συμμετέχουν σε μεταβολικά μονοπάτια που σχετίζονται με κυτταρική διαίρεση, βιοσύνθεση του κυτταρικού φακέλου, διακίνηση και επεξεργασία πρωτεϊνών, ενεργειακές μετατροπές, μετατόπιση δια μέσου της μεμβράνης, μεταβολισμός μικρών μορίων, όπως επίσης και διεργασίες που σχετίζονται με τα νουκλεϊκά οξέα όπως η αντιγραφή του DNA και η μεταγραφή και μετάφραση του RNA (Εικόνα 4.2 και Εικόνα 4.3).

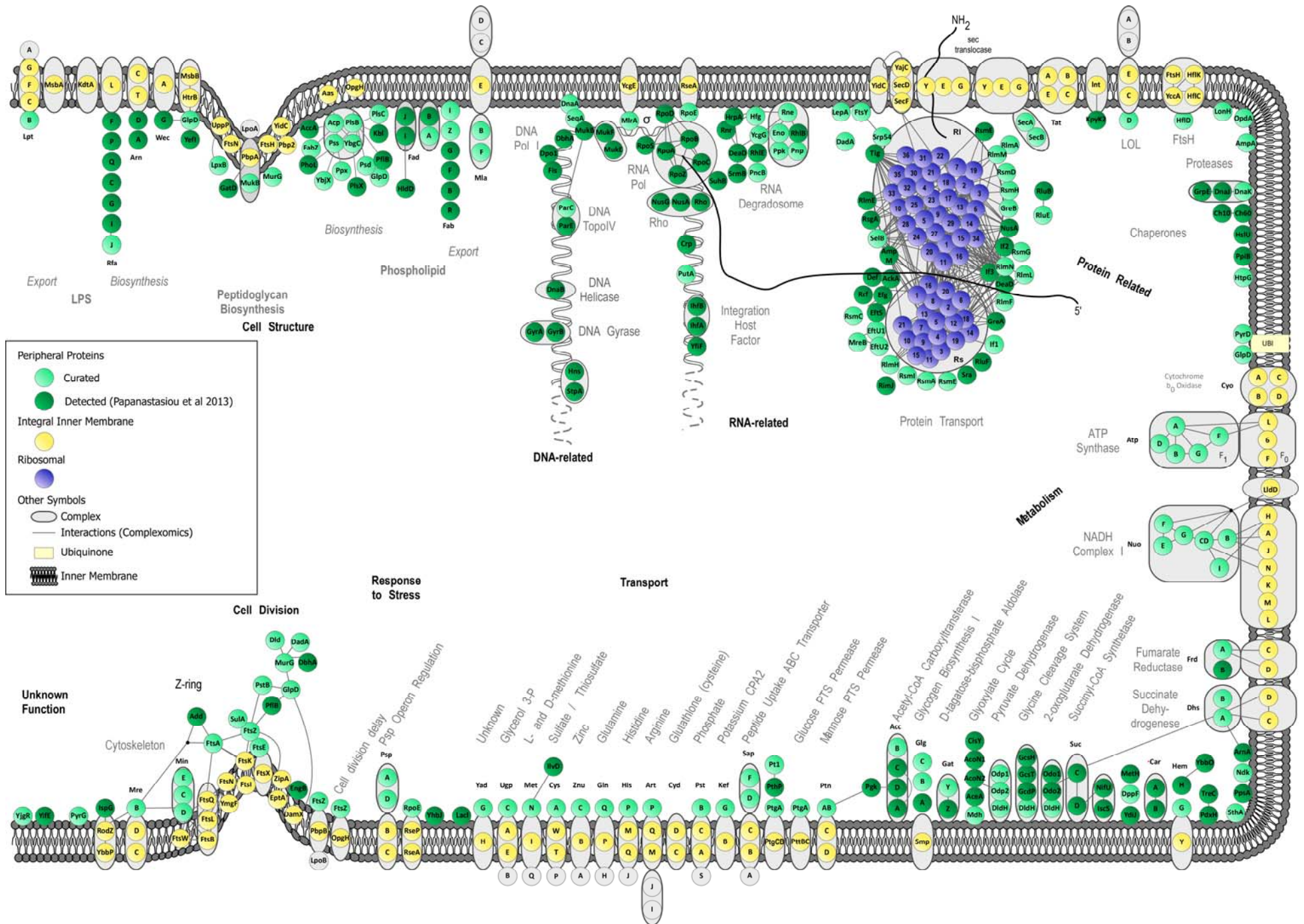
Αποφασίσαμε να διερευνήσουμε κατά πόσο οι περιφερικές πρωτεΐνες είναι αναγκαίες για την βιωσιμότητα του κυττάρου. Βασιστήκαμε στις πρωτεϊνικές αλληλουχίες απαραίτητες για την βιωσιμότητα, που είναι καταχωρημένες στην βάση δεδομένων DEG (Zhang et al, 2004) και εφαρμόσαμε αναζητήσεις ομοιότητας και πολλαπλής στοίχισης ακολουθιών (Johnson et al, 2008) επιλέγοντας τις πρωτεΐνες με ποσοστό ταυτοποίησης μεγαλύτερο από 20% (identity) και e-value

---

μικρότερο από  $10^{-3}$ . Τα αποτελέσματα δείχνουν ότι το 44% των περιφερικών πρωτεϊνών παρουσιάζουν ομολογία με τουλάχιστον μία καταχωρημένη πρωτεΐνη στην βάση δεδομένων DEG.

Τέλος αναζητήσαμε το ποσοστό των περιφερικών πρωτεϊνών που είναι συντηρημένες ανάμεσα στα βακτήρια και πιο ειδικά στα παθογόνα βακτήρια. Τα πλήρη πρωτεϊνώματα 25 μη παθογόνων και 22 παθογόνων βακτηρίων συλλέχθηκαν από την βάση δεδομένων UniProt και για το σύνολο των περιφερικών πρωτεϊνών του *E.coli* αναζητήσαμε ομόλογες πρωτεΐνες στα συγκεκριμένα βακτήρια. Ορίσαμε ως συντηρημένες περιφερικές πρωτεΐνες στα παθογόνα και μη βακτήρια αυτές για τις οποίες βρέθηκαν ομόλογες αλληλουχίες σε τουλάχιστον 16 στελέχη. Τα αποτελέσματα της αναζήτησης έδειξαν ότι >60% των περιφερικών πρωτεϊνών υπάρχουν και σε άλλα βακτήρια, συμπεριλαμβανομένου και παθογόνα στελέχη. Για τις μισές περίπου υπάρχουν ομόλογες πρωτεΐνες σε όλα τα στελέχη που μελετήσαμε.





---

**Εικόνα 4.2 – «Πανοραμική» εικόνα των περιφερικών πρωτεϊνών του *E.coli***


---

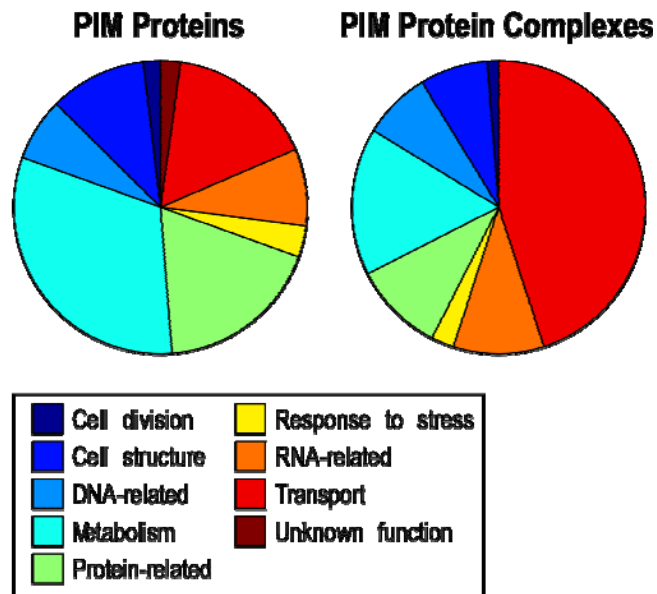
Απεικονίζεται η διαμήκης τομή ενός βακτηρίου σε φάση διαίρεσης. Παρουσιάζεται μόνο η μια πλευρά της ΠΜ ενώ ο υπόλοιπος κυτταρικός φάκελος έχει παραληφθεί για λόγους απλοποίησης. Βασιζόμενοι στην ταξινόμηση των πρωτεϊνών κατά MultiFun οι περιφερικές πρωτεΐνες (πράσινοι κύκλοι) συμμετέχουν σε εννέα διαφορετικές κυτταρικές διεργασίες (σημειώνονται με έντονη γραμματοσειρά): cell structure (κυτταροσκελετός), πρωτεΐνες που αλληλεπιδρούν με μόρια DNA και RNA (DNA-, RNA- and protein related), πρωτεΐνες του μεταβολισμού (metabolism), πρωτεΐνες διακίνησης (transport), πρωτεΐνες στρες (response to stress) και κυτταρική διαίρεση (cell division). Οι αντίστοιχες υποκατηγορίες που εμπεριέχονται σε αυτές, σημειώνονται με γκρι γράμματα.

Οι πρωτεΐνες έχουν ομαδοποιηθεί με βάση τα πρωτεϊνικά σύμπλοκα (εμπεριέχονται σε γκρι περιγράμματα) και τα μεταβολικά/σηματοδοτικά μονοπάτια στα οποία είναι γνωστό ότι συμμετέχουν σε εγγύτητα με τους μεμβρανικούς «συνεργάτες» (κίτρινοι κύκλοι), όπου αυτό είναι δυνατόν (Keseler et al, 2011). Επίσης συμπεριλαμβάνονται πρωτεϊνικές αλληλεπιδράσεις που έχουν διαπιστωθεί με πειράματα διαδοχικής απομόνωσης συγγένειας (tandem-affinity purification experiments) (Aranda et al, 2010) και συμβολίζονται ως συνδετικές γραμμές μεταξύ των μορίων (όπου είναι δυνατόν). Για λόγους διευκόλυνσης της απεικόνισης οι αλληλεπιδράσεις μεταξύ των πρωτεϊνών ενσωματώθηκαν και απεικονίστηκαν στο Cytoscape (Keseler et al, 2011). Οι αλληλεπιδράσεις που αναπαριστώνται στο συγκεκριμένο σχήμα δεν έχουν προκύψει απαραίτητα από την ίδια πειραματική διαδικασία και ενδέχεται να μην ισχύουν ταυτόχρονα, δηλαδή να αποτελούν εναλλακτικά υποσύνολα ενός μεγαλύτερου συμπλόκου που έχουν ανιχνευθεί σε διαφορετικά πειράματα.

Οι πρωτεΐνες που σχετίζονται με την σύνθεση του ΠΠ και των φωσφολιπιδίων έχουν βασιστεί σε γνωστές αλληλεπιδράσεις που βρίσκονται καταχωρημένες στην βάση δεδομένων IntAct (Aranda et al, 2010). Οι πρωτεΐνες του ριβοσώματος (μωβ κύκλοι) απεικονίζονται στην περιοχή της ΠΜ ομαδοποιημένες στις αντίστοιχες υπομονάδες (μεγάλη (50S) και μικρή (30S)). Έπειτα από την μετάφραση τους στα ριβοσώματα οι πολυπεπτιδικές αλυσίδες (συμβολίζονται με χοντρή μαύρη γραμμή) αναγνωρίζονται από το εκκριτικό μονοπάτι Sec είτε μεταμεταφραστικά (μέσω της καθοδήγησης των πρωτεϊνών SecA και SecB) είτε συνμεταφραστικά (υπό την καθοδήγηση της πρωτεΐνης Srp η οποία προσδένεται στο ριβόσωμα) (Chatzi et al, 2013).

Ένα σύνολο από κυτταροπλασματικές πρωτεΐνες βρέθηκαν σε γεινίαση με την ΠΜ (Papanastasiou et al, 2013) και πλέον αποκαλούνται περιφερικές πρωτεΐνες (PIM proteins). Αυτές τις πρωτεΐνες τις συμβολίζουμε με σκούρους πράσινους κύκλους ενώ όσες υπήρξαν γνωστές περιφερικές πρωτεΐνες και τις συλλέξαμε από την βιβλιογραφία τις συμβολίζουμε με ανοιχτό πράσινο. Για πιο πλήρη αναπαράσταση των συμπλόκων απεικονίζουμε και ορισμένες περιπλασματικές πρωτεΐνες στην εξωτερική επιφάνεια της ΠΜ. Πρωτεΐνες με καμία επιβεβαιωμένη λειτουργία σημειώνονται ως άγνωστες (Unknown Function).

---



Εικόνα 4.3 – Ταξινόμηση των περιφερικών πρωτεϊνών με βάση τις κυτταρικές λειτουργίες στις οποίες συμμετέχουν.

Η κατηγοριοποίηση των περιφερικών πρωτεϊνών βασίστηκε στο σύστημα οργάνωσης των κυτταρικών λειτουργιών MultiFun (Riley et al, 2005). Η αντιστοίχιση των πρωτεϊνών του *E.coli* στις γονιδιακές οντολογίες GO (gene ontology) (GOConsortium, 2012) είναι διαθέσιμη στις βάσεις δεδομένων EcoCyc and Uniprot (Consortium, 2012; Keseler et al, 2011). Το μεγαλύτερο μέρος των περιφερικών πρωτεϊνών εμπλέκονται σε μεταβολικά μονοπάτια (π.χ. βιοσύνθεση αμινοξέων, μεταφορά ηλεκτρονίων) αλλά και στην «μετάδοση» της γονιδιακής πληροφορίας (πρωτεΐνες που αλληλεπιδρούν με DNA και RNA). Από την σκοπιά των αντίστοιχων πρωτεϊνικών συμπλόκων στα οποία συμμετέχουν οι περιφερικές πρωτεΐνες τα περισσότερα σχετίζονται με την μεταφορά βιολογικών και ανόργανων μορίων.



---

## 4.2 Συζήτηση

Παρουσιάσαμε μια συστηματική ανάλυση για την χαρτογράφηση του δικτύου πρωτεϊνικών αλληλεπιδράσεων και κυτταρικών διεργασιών στις οποίες εμπλέκονται οι περιφερικές πρωτεΐνες, ένα υποσύνολο που ήταν μέχρι πρόσφατα ελλιπώς χαρακτηρισμένο. Το περιφερικό πρωτεϊνώμα του *E. coli* BL21-DE3 εκτιμάται ότι αποτελεί το 12% του θεωρητικού πρωτεϊνώματος (589 από τις 4842 πρωτεΐνες) ενώ είναι το 14% του βασικού πρωτεϊνώματος και 22% του πρωτεϊνώματος που αναμένεται να εκφράζεται σε πλούσιο θρεπτικό μέσο.

Οι αλληλεπιδράσεις των περιφερικών πρωτεϊνών με την μεμβράνη φαίνεται να είναι πολλαπλές και σύνθετες. Σχετίζονται μη-ομοιοπολικά με τα λιπίδια των μεμβρανών χωρίς να εισχωρούν στο υδρόφοβο πυρήνα της διπλοστοιβάδας (Dowhan et al, 2008). Επιπλέον διαλυτές πρωτεΐνες ενδέχεται να σχετίζονται με την μεμβράνη περιστασιακά και με έμμεσο τρόπο αλληλεπιδρώντας με άλλες περιφερικές πρωτεΐνες. Παρόλο που σε αυτήν την περίπτωση δεν υπάρχει άμεση αλληλεπίδραση με την μεμβράνη προτείνουμε ότι και αυτές οι πρωτεΐνες να θεωρούνται περιφερικές.

Για παράδειγμα η FtsE εμπλέκεται στην διαδικασία της κυτταρικής διαίρεσης αλληλεπιδρώντας είτε άμεσα με την μεμβρανική πρωτεΐνη FtsX είτε έμμεσα μέσω της περιφερικής πρωτεΐνης FtsZ που με την σειρά της «συγκεντρώνει» άλλες πρωτεΐνες που σχετίζονται με την διαδικασία της κυτταρικής διαίρεσης και δεν αλληλεπιδρούν άμεσα με την μεμβράνη (Εικόνα 4.2)(De Leeuw et al, 1999; Lutkenhaus et al, 1997).

Μερικές περιφερικές πρωτεΐνες αλληλεπιδρούν ασθενώς με την μεμβράνη. Από αυτές κάποιες αλληλεπιδρούν με την ΠΜ μέσω ασθενών ηλεκτροστατικών αλληλεπιδράσεων με τις υδρόφιλες κεφαλές των λιπιδίων (π.χ. PspA) ενώ άλλες αλληλεπιδρούν μέσω περιφερικών πρωτεϊνών (π.χ. SecB, HflD και NuoEFG). Άλλα παραδείγματα είναι οι AccA and AccD, δύο υπομονάδες του ενζύμου acetyl-CoA carboxylase, μεταβολικό σύμπλοκο που καταλύει την σύνθεση των λιπαρών οξέων. Στην περίπτωση του *E.coli* το συγκεκριμένο ένζυμο έχει καταχωρηθεί ως κυτταροπλασματικό (Consortium, 2012) ενώ σε ένα συγγενικό βακτήριο, το *B. subtilis* υπάρχει πειραματική ένδειξη της γεινίασης με την ΠΜ (Meile et al, 2006).

Σε πρόσφατη πειραματική ανάλυση (Papanastasiou et al, 2013) ταυτοποιήθηκαν οι πρωτεΐνες AccA, AccC και AccD ως περιφερικές που δημιουργούν ισχυρές ηλεκτροστατικές

---

---

αλληλεπιδράσεις με την μεμβράνη και μπορούν να αποκολληθούν μέσω ιοντικής ισχύος. Σε παλαιότερη μελέτη έχει βρεθεί ότι το συγκεκριμένο σύμπλοκο είναι ιδιαίτερα ασταθές (Choi-Rhee et al, 2003). Είναι σημαντικό να αναφερθεί ότι η παρουσία των συγκεκριμένων μορίων στην ΠΜ ίσως σχετίζεται με κάποια άγνωστη μέχρι σήμερα λειτουργία τους. Μια πιθανή λειτουργία είναι ότι αυτά τα σύμπλοκα εξυπηρετούν την διοχέτευση μικρών μορίων διαμέσου της ΠΜ (Huthmacher et al, 2008; Perez-Bercoff et al, 2011)

Σε άλλες περιπτώσεις οι αλληλεπιδράσεις με την επιφάνεια της ΠΜ είναι ιδιαίτερα σταθερές και παρουσιάζουν αντίσταση στην επεξεργασία με χημικούς παράγοντες. Υπάρχει μεγάλη ποικιλία από τέτοια παραδείγματα όπου συχνά υπάρχει σχεδόν αποδεδειγμένα αντίστοιχος μεμβρανικός υποδοχέας (π.χ. SecA πάνω στο μεμβρανικό σύμπλοκο SecYEG; SRP πάνω στην FtsY; SeqA πάνω στην DnaA). Κάποια από αυτά τα σύμπλοκα μπορούν να διαταραχθούν με μη ιονικούς παράγοντες (π.χ. MinD, MlaB, (Paranastasiou et al, 2013)) γεγονός που αποδεικνύει ότι υπάρχουν ισχυρές υδροφοβικές αλληλεπιδράσεις. Συνεπώς οι παραδοσιακές βιοχημικές μέθοδοι που έχουν κατά καιρούς χρησιμοποιηθεί γιατί τον προσδιορισμό των περιφερικών πρωτεϊνών (Adelman et al, 1973; Fujiki et al, 1982; Kreibich et al, 1974; Ohlendieck, 2003; Steck, 1974) αδυνατούν να ανιχνεύσουν τέτοιου είδους αλληλεπιδράσεις.

Οι διάφορες πρωτομικές αναλύσεις τείνουν να θεωρούν τις κυτταροπλασματικές πρωτεΐνες που απομονώνονται μαζί με τις μεμβράνες, ως θόρυβο (Aivaliotis et al, 2006; Aivaliotis et al, 2007; Alexandersson et al, 2004; Klein et al, 2005; Pieper et al, 2009) καθώς είναι δύσκολο να διευκρινιστεί αν αποτελούν πραγματικές ή τυχαίες αλληλεπιδράσεις.

Επιπλέον η δυναμική αλληλεπίδραση με την μεμβράνη αποτελεί άλλη μια ενδιαφέρουσα πτυχή του περιφερικού πρωτεϊνώματος. Υπό αυτήν σκοπιά μπορούμε να φανταστούμε ότι η ΠΜ λειτουργεί ως μια προσωρινή «αποθήκη» που είτε συγκράτα μόρια ρυθμίζοντας κυτταροπλασματικές συγκεντρώσεις είτε τα απελευθερώνει όταν υπάρχει εξωτερική διέγερση. Τέτοιες περιπτώσεις αποτελούν οι μεταγραφικοί παράγοντες NadR (Raffaelli et al, 1999), RpoE , PutA (Ostrovsky de Spicer et al, 1993) και BglG (Lorian et al, 2003). Ενδιαφέρον παρουσιάζει η κυτταροπλασματική πρωτεΐνη LacI η οποία έχει επιβεβαιωθεί ότι δημιουργεί ισχυρά τετραμερή σύμπλοκα στην μεμβράνη (Paranastasiou et al, 2013). Το πιο πιθανό σενάριο για την περίπτωση των μεταγραφικών παραγόντων είναι ότι παραμένουν δεσμευμένοι στην μεμβράνη

---

μέχρι τη στιγμή που θα υπάρξει κάποια εξωγενής διέγερση (π.χ. πρόσληψη κάποιου μορίου από το κύτταρο).

Παρομοίως, η πρωτεΐνη Rne, υπομονάδα της συμπλόκου της αποικοδόμησης του RNA (RNA degradosome) παραμένει προσδεμένη στην ΠΜ κατά την φάση της κυτταρικής ανάπτυξης ενώ μόνο στο τελικό στάδιο της στατικής φάσης απελευθερώνεται για να γίνει ενζυματικά ενεργή (Khemici et al, 2008; Lopez-Campistrous et al, 2005). Ένα άλλο παράδειγμα δυναμικής αλληλεπίδρασης με την ΠΜ αποτελεί η PspA που σχηματίζει πιθανές ολιγομοριακές δομές στην επιφάνεια της ΠΜ οι οποίες πιστεύεται ότι διοχετεύουν ιόντα (Kobayashi et al, 2007) ή αλληλεπιδρά μέσω της πρωτεΐνης Psp (Adams et al, 2003).

Βασίσαμε την μελέτη μας σε σταθερές και σίγουρες πρωτεϊνικές αλληλεπιδράσεις και καταλήξαμε σε μια καλύτερη εκτίμηση του πρωτεϊνώματος των περιφερικών πρωτεϊνών. Ως αποτέλεσμα εκτιμήσαμε ότι το δίκτυο αλληλεπιδράσεων των περιφερικών πρωτεϊνών είναι πιο δυναμικό και πιο πολύπλοκο από πιστεύαμε μέχρι σήμερα. Πολλές κυτταροπλασματικές πρωτεΐνες αναμένεται να σχηματίζουν φευγαλές αλληλεπιδράσεις με την μεμβράνη ή με περιφερικές πρωτεΐνες που βρίσκονται προσδεμένες στην μεμβράνη. Οι αλληλεπιδράσεις στις οποίες συμμετέχουν οι περιφερικές πρωτεΐνες δεν θα πρέπει να αντιμετωπίζονται ως σταθερές και μεταξύ μόνο συγκεκριμένων ζευγαριών αλλά ως δυναμικές και πολυδιάστατες συμπεριλαμβάνοντας ακόμα και πρωτεΐνες με πολλαπλούς ρόλους (“moonlighting” proteins; (Jeffery, 1999)). Κατά συνέπεια μια περιφερική πρωτεΐνη μπορεί να συμμετέχει σε παραπάνω από ένα σύμπλοκα όπως συμβαίνει στην περίπτωση των: **MukB** (Li et al, 2010; Petrusenko et al, 2006), **MurG** (de Boer, 2010), **RpoE** (Bordes et al, 2011; Hayden et al, 2008), **MreB** (Bendezu et al, 2009a; Kruse et al, 2006; Salje et al, 2011; van den Ent et al, 2010), **PyrG** (Ingerson-Mahar et al, 2010) και **GlpD** (Cozzarelli et al, 1965; Pan et al, 2010; Schryvers et al, 1978). Τα πράγματα φαίνεται επίσης να γίνονται ακόμα περισσότερο πολύπλοκα για ορισμένες πρωτεΐνες οι οποίες μπορούν να έχουν εντυπωσιακά διαφορετικές δομικές λειτουργίες όπως η πρωτεΐνη MreB, ανάλογο της ακτίνης στα βακτήρια η οποία φαίνεται να καθορίζει την πολικότητα των κυττάρων (Gitai et al, 2004), να λειτουργεί ως δομικό στοιχείο, να αλληλεπιδρά με τον παράγοντα επιμήκυνσης EF-Tu και να καθορίζει την μορφολογία του κυττάρου (Soufo et al, 2010) αλλά και να σε αλληλεπίδραση με την πολυμεράση του RNA να καθορίζει την εκκίνηση του διαχωρισμού των χρωμοσωμάτων (Kruse et al, 2006).

---

---

Εν κατακλείδι, ένας αξιοσημείωτος αριθμός από κυτταροπλασματικές πρωτεΐνες αλληλεπιδρούν με την ΠΜ. Το πρωτεΐνωμα των περιφερικών πρωτεϊνών αποτελεί ένα δυναμικό σύνδεσμο της ΠΜ με τις περισσότερες κυτταρικές διεργασίες. Το παρόν δίκτυο αλληλεπιδράσεων αναμένεται να επιτελέσει εργαλείο για την αναγνώριση περιφερικών πρωτεϊνών και σε άλλους οργανισμούς.





---

## **ΚΕΦΑΛΑΙΟ 5 Αποτελέσματα – Μελέτη των φυσικοχημικών ιδιοτήτων των πρωτεϊνών του ΚΦ με στόχο την βέλτιστη ανίχνευση με τεχνικές φασματομετρίας μάζας**

Οι βιολογικές μεμβράνες οριοθετούν τα υποκυτταρικά διαμερίσματα και είναι απαραίτητες για την διπλής κατεύθυνσης ροή χημικών μορίων και σημάτων. Οι ιδιότητες των βιολογικών μεμβρανών καθορίζονται από τις πρωτεΐνες που βρίσκονται ενσωματωμένες σε αυτές.

Όλα τα βακτήρια έχουν μια τουλάχιστον μεμβράνη την πλασματική (ΠΜ). Στα Gram-βακτήρια υπάρχει μια δεύτερη βιολογική μεμβράνη που περικλείει την πρώτη και ονομάζεται εξωτερική (ΕΜ). Ανάμεσα στις δύο μεμβράνες βρίσκεται το περίπλασμα, το οποίο εκτός από πρωτεΐνες περιέχει και άλλα βιολογικά μόρια όπως το πλέγμα πεπτιδογλυκάνης (ΠΠ) το οποίο είναι συνδεδεμένο στην εσωτερική στοιβάδα λιπιδίων της ΕΜ (δες ενότητα 1.2.2). Η ΠΜ το περίπλασμα, η ΕΜ και οι πρωτεΐνες που εμπεριέχονται αποτελούν τον λεγόμενο κυτταρικό φάκελο (ΚΦ).

Η γνώση της κατανομής των πρωτεϊνών στον ΚΦ είναι απαραίτητη για την κατανόηση των πρωτεϊνικών αλληλεπιδράσεων, των κυτταρικών λειτουργιών, της προσαρμογής σε αλλαγές του περιβάλλοντος, της παθογένειας. Πλέον, η μεγάλη διαθεσιμότητα ολοκληρωμένων πρωτεϊνωμάτων αλλά και οι τεχνολογικές εξελίξεις έχουν επιτρέψει την μελέτη των πρωτεϊνών σε ευρεία κλίμακα γνωστή πλέον και ως κλάδος της πρωτεομικής (proteomics) που περιλαμβάνει τεχνικές όπως η φασματομετρία των μαζών (MS: mass spectrometry).

Παρά τις όλο και αυξανόμενες τεχνολογικές καινοτομίες στον κλάδο της πρωτεομικής, η πειραματική ανίχνευση των μεμβρανικών πρωτεϊνών παραμένει κριτικής σημασίας για την πλήρη ανίχνευση του πρωτεϊνώματος του ΚΦ. Το μεμβρανικό πρωτεϊνώμα με βάση την εμπειριστατωμένη υποκυτταρική ταξινόμηση της βάσης δεδομένων STEPdb (δες 0) αποτελεί το ~48% του πρωτεϊνώματος του ΚΦ (Orfanoudaki et al, 2014). Τα χαμηλά επίπεδα έκφρασης των μεμβρανικών πρωτεϊνών (Taniguchi et al, 2010; Yoon et al, 2012) αλλά και ο υδρόφοβος χαρακτήρας τους απαιτούν την χρήση μη συμβατικών μεθόδων επεξεργασίας των δειγμάτων και συχνά τον συνδυασμό υποκυτταρικής κλασματοποίησης με διαδοχικά βήματα διαλυτοποίησης και αύξησης της περιεκτικότητας των μεμβρανικών πρωτεϊνών (Speers et al, 2007; Wagner et al, 2006).

---

Επιπλέον οι τυπικές πρωτεάσες που χρησιμοποιούνται στην προσέγγιση «από κάτω προς τα πάνω» (bottom-up) της πρωτεομικής ανάλυσης, στοχεύουν κατάλοιπα Λυσίνης και Αργινίνης (π.χ. Τρυψίνη, Lys-C) και στην περίπτωση των μεμβρανικών πρωτεϊνών παράγουν μεγάλα πεπτίδια τα οποία 1) απομονώνονται δύσκολα από πηκτώματα πολυακρυλαμίδης (polyacrylamide gels), 2) είναι υδρόφοβα και 3) έχουν λιγότερες πιθανότητες να πάρουν φορτία. Τα δύο τελευταία αφορούν πεπτίδια με ιδιαίτερα μεγάλο λόγο μάζα προς φορτίο (m/z) που βρίσκεται εκτός τεχνικών ορίων ανίχνευσης.

Βελτίωση των μεθόδων προετοιμασίας του δείγματος όπως η αυξημένη θερμοκρασία κατά το διαχωρισμό υγρής χρωματογραφίας (LC separation), πρόσθετες χημικές ενώσεις για την διαλυτοποίηση των πρωτεϊνών αλλά και αναδυόμενες τεχνολογικές εξελίξεις όπως ηλεκτροφόρηση απουσία πηκτώματος (OFFGEL) (Michel et al, 2003) και προετοιμασία δείγματος με διήθηση (FASP: filter aided sample preparation) (Wisniewski et al, 2009b) έχουν ενισχύσει την ανίχνευση των μεμβρανικών πρωτεϊνών (Manadas et al, 2009; Wisniewski et al, 2009a).

Στην συγκεκριμένη ανάλυση παρουσιάζουμε μια εκτεταμένη θεωρητική μελέτη των χαρακτηριστικών που παρεμποδίζουν την ανίχνευση των πρωτεϊνών του ΚΦ στο βακτήριο *E.coli* BL21-DE3 με ιδιαίτερη έμφαση στο μεμβρανικό πρωτεϊνωμα (ΜΠ) το οποίο διαφέρει σε φυσικοχημικές ιδιότητες από τις υπόλοιπες πρωτεΐνες του ΚΦ. Με βάση αυτήν την ανάλυση προσδιορίζουμε το ποσοστό των πρωτεϊνών που είναι ανιχνεύσιμα με την μέθοδο της φασματομετρίας μάζας.

Στη συνέχεια ξεκινώντας από το εμπειριστατωμένο σύνολο των πρωτεϊνών του ΚΦ (Orfanoudaki et al, 2014) και συνδυάζοντας πρωτεϊνωματικά και γονιδιωματικά δεδομένα προσδιορίζουμε το υποσύνολο του πρωτεϊνώματος που αναμένεται να εκφραστεί από το βακτήριο *E.coli* BL21-DE3 όταν αυτό αναπτύσσεται σε πλήρες θρεπτικό μέσο (Luria-Bertani medium).

Έπειτα θέτουμε ως στόχο να καθορίσουμε πως η διαδικασία προετοιμασίας των δειγμάτων επηρεάζει τα τρία βασικά σημεία της πρωτεομικής ανάλυσης: την αναγνώριση πρωτεϊνών, την μέγιστη κάλυψη της αλληλουχίας τους (δηλ. ποσοστό πρωτεΐνης που ανιχνεύεται μέσω αντίστοιχων πεπτιδίων) και τέλος την ποσοτικοποίηση τους. Συγκρίνουμε τέσσερις μεθόδους προετοιμασίας δείγματος, αποδιατακτική ηλεκτροφόρηση (SDS-PAGE), πρωτεόλυση στην επιφάνεια ανεστραμμένων μεμβρανικών κυστιδίων (surface proteolysis), ηλεκτροφόρηση



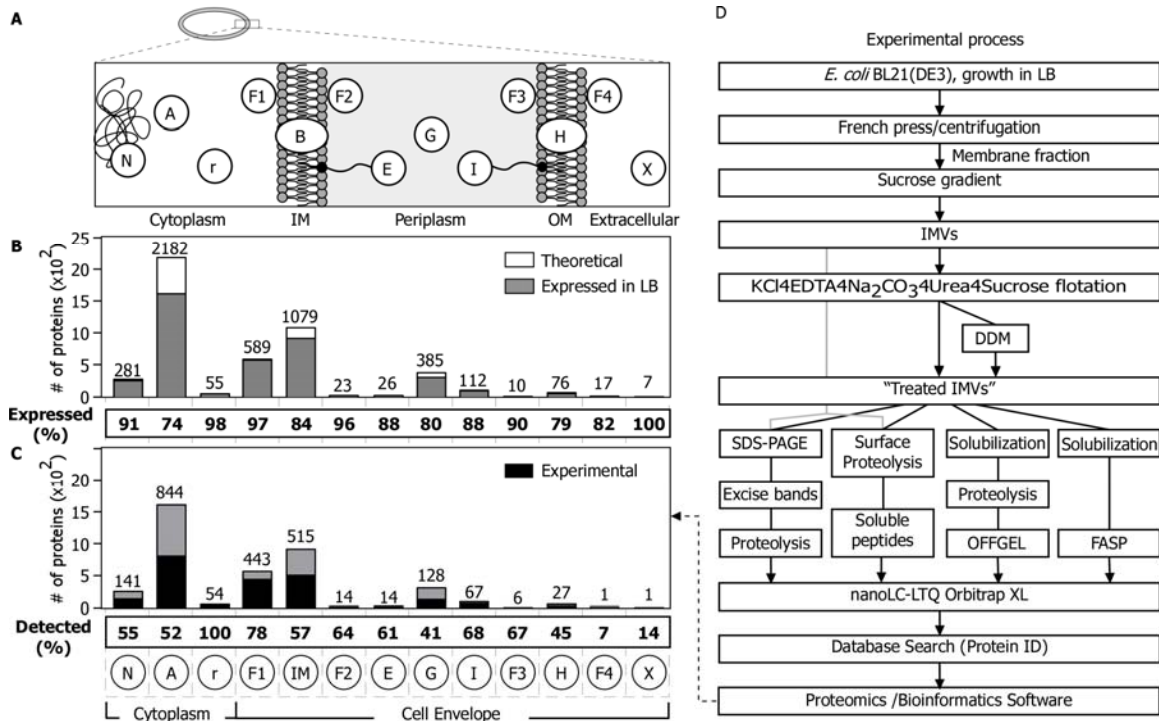
---

απουσία πηκτώματος (OFFGEL) και προετοιμασία δείγματος με διήθηση (FASP), και εξετάζουμε κατά πόσο η κάθε μέθοδος προσφέρει πλεονεκτήματα και προδιαθέτει την ταυτοποίηση πεπτιδίων.

Από την ανάλυση συμπεραίνουμε ότι ο συνολικός αριθμός των ταυτοποιημένων πρωτεϊνών είναι περίπου ίδιος ανάμεσα στις μεθόδους υποδεικνύοντας ότι οι περισσότερες από τις μεμβρανικές πρωτεΐνες είναι ανιχνεύσιμες από τα πολικά τους πεπτίδια. Ότι αφορά όμως την κάλυψη των μεμβρανικών πρωτεϊνών η επιλογή της μεθόδου είναι κριτικής σημασίας καθώς τα πεπτίδια που είναι ανιχνεύσιμα σε κάθε μέθοδο διαφέρουν σε μήκος και υδροφοβικότητα.

### 5.1 Υποκυτταρική ταξινόμηση του θεωρητικού πρωτεϊνώματος του *E.coli* BL21-DE3

Βασιστήκαμε στην εμπειριστατωμένη ταξινόμηση των πρωτεϊνών δυο στελεχών του *E.coli* (K-12 και BL21-DE3) (Orfanoudaki et al, 2014), η οποία ορίζει εννέα διακριτές κατηγορίες μη κυτταροπλασματικών πρωτεϊνών (**Εικόνα 2.1**) και μια ξεχωριστή κατηγορία για τις πρωτεΐνες που είναι διαλυτές αλλά αλληλεπιδρούν και με τον ΚΦ (περιφερικές πρωτεΐνες, F1). Η ταξινόμηση αναφέρει ότι το πρωτεϊνώμα του κυτταρικού φακέλου συνιστά σχεδόν το μισό του συνολικού πρωτεϊνώματος (~48%; **Εικόνα 2.1, υποκυτταρική θέση 'B'**) ξεπερνώντας κατά πολύ τις προηγούμενες προβλέψεις (Hu et al, 2009). Οι μεμβρανικές και οι περιφερικές πρωτεΐνες αποτελούν τα μεγαλύτερα υποσύνολα του ΚΦ (47% και 25% αντίστοιχα).



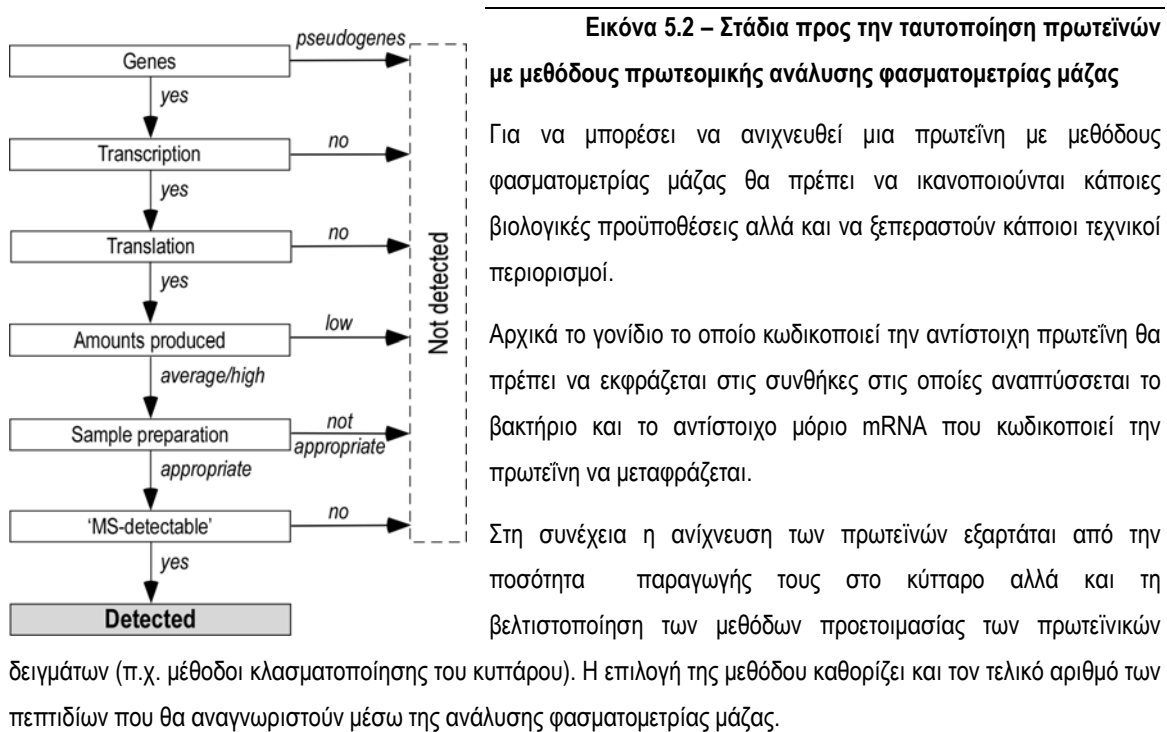
**Εικόνα 5.1 - Εμπειριστικώς ταξινομημένη ταξινόμηση του πρωτεϊνώματος του *E. coli* BL21-DE3 και πειραματική διαδικασία ανίχνευσης με μεθόδους πρωτεομικής ανάλυσης**

**A.** Αναπαράσταση της τομής ενός κυτάρου *E. coli* και των αντίστοιχων υποκυτταρικών θέσεων των πρωτεϊνών: κυταροπλασματικές (A), περιφερικές πρωτεΐνες της ΠΜ από την πλευρά του κυταροπλάσματος (F1), διαμεμβρανικές πρωτεΐνες της ΠΜ (B), περιφερικές πρωτεΐνες της ΠΜ από την πλευρά του περιπλάσματος (F2), λιποπρωτεΐνες της ΠΜ (E), περιπλασμικές πρωτεΐνες (G), λιποπρωτεΐνες της EM (I), περιφερικές πρωτεΐνες της EM από την πλευρά του περιπλάσματος (F3), διαμεμβρανικές πρωτεΐνες της EM (H), περιφερικές πρωτεΐνες της EM από την πλευρά του εξωκυτταρίου χώρου (F4), εξωκυτταρίες πρωτεΐνες (X). **B.** Κατανομή των θεωρητικών πρωτεϊνών του *E. coli* BL21-DE3 στα διάφορα υποκυτταρικά διαμερίσματα ('Theoretical') και ταυτόχρονη απεικόνιση του ανιχνεύσιμου πρωτεϊνώματος σε πλούσιο θρεπτικό μέσο ('Expressed in LB'). **C.** Κατανομή των πρωτεϊνών που ανιχνεύθηκαν πειραματικά ('Experimental'; (Papanastasiou et al, in preparation)). **D.** Διάγραμμα ροής των πειραματικών μεθόδων που ακολουθήθηκαν προς την πλήρη ταυτοποίηση του πρωτεϊνώματος του ΚΦ του *E. coli* BL21-DE3 (Papanastasiou et al, in preparation).

SDS-PAGE: αποδιατακτική ηλεκτροφόρηση; Surface proteolysis: Πρωτεόλυση στην επιφάνεια των AMK, OFFGEL: ηλεκτροφόρηση απουσία πηκτώματος; FASP: προετοιμασία δείγματος με διήθηση; IMVs: Inverted membrane vesicles; DDM: n-Dodecyl-b-D-maltoside (non-ionic detergent); LB: Luria-Bertani broth

## 5.2 Εκτίμηση του υποσύνολου του πρωτεϊνώματος που εκφράζεται όταν το βακτήριο αναπτύσσεται σε πλούσιο θρεπτικό μέσο.

Τα βακτήρια εκτίθενται σε μεγάλη ποικιλία περιβαλλοντικών αλλαγών, άλλες βραχυπρόθεσμες και άλλες που επιδρούν στην πορεία των γενεών. Καταφέρνουν και επιβιώνουν σε αυτές τις αλλαγές έχοντας ενσωματώσει στο γονιδίωμα τους γονίδια που τους προσδίδουν συγκεκριμένες ιδιότητες αλλά και αντίστοιχους μηχανισμούς που ενεργοποιούν ή κατασιγάζουν την έκφραση τους (Lopez-Mauy et al, 2008). Συνεπώς ανάλογα με τις περιβαλλοντολογικές συνθήκες εκφράζεται διαφορετικό υποσύνολο του γονιδιώματος.



Η ταυτοποίηση πρωτεϊνών χρησιμοποιώντας πρωτομικές μεθόδους που βασίζονται στην φασματομετρία μάζας (ΦΜ), ακολουθεί μια σειρά από γεγονότα που ξεκινάνε από το στάδιο της σύνθεσης μιας πρωτεΐνης στο κύτταρο σε ικανές ποσότητες για να ανιχνευτούν τα πεπτίδια της. Ως εκ τούτου ένα γονίδιο θα πρέπει να μεταγραφεί στις εκάστοτε συνθήκες που μεγαλώνει το βακτήριο και στην συνέχεια το αντίστοιχο mRNA να μεταφραστεί σε επίπεδο πρωτεΐνης. Στο *E. coli*, οι συγκεντρώσεις μορίων mRNA είναι της τάξης  $10^{-2}$  μέχρι  $10^2$  (Allen et al, 2003) ενώ οι αντίστοιχες ποσότητες σε επίπεδο πρωτεΐνης κυμαίνονται από 10 το  $10^5$  (και στις δύο περιπτώσεις εκφράζουν μόρια ανά κύτταρο) (de Sousa Abreu et al, 2009a; Vogel et al, 2012).

Ταυτόχρονη μελέτη σε επίπεδο mRNA και πρωτεΐνης έχουν δείξει ότι υπάρχει μικρή συσχέτιση ανάμεσα στις ποσότητες αυτών των δύο βιολογικών μορίων (Corbin et al, 2003; Covert et al, 2004; de Sousa Abreu et al, 2009b; Lu et al, 2007; Masuda et al, 2009; Taniguchi et al, 2010).

Αποφασίσαμε να εκτιμήσουμε το κλάσμα του πρωτεϊνώματος του ΚΦ στο *E.coli* BL21-DE3, που αναμένεται να εκφράζεται στις συνθήκες ανάπτυξης που χρησιμοποιήσαμε και το οποίο ονομάζουμε «ανιχνεύσιμο πρωτεϊνωμα». Συνδυάσαμε δεδομένα διαθέσιμα στην βιβλιογραφία, από συγκριτικές μελέτες για τις διαφορές στην πρωτεϊνική έκφραση που υπάρχουν ανάμεσα σε διαφορετικά στελέχη *E.coli* καθώς και τις διαφορές πρωτεϊνικής έκφρασης που προκύπτουν από τις συνθήκες ανάπτυξης (Πίνακας 5.1, Πίνακας 6.10).

Συγκεκριμένα τα δεδομένα προσδιορίσαμε ότι 3943 από τις 4842 πρωτεΐνες του BL21-DE3 αναμένεται να συνθέτονται όταν το βακτήριο αναπτύσσεται σε πλούσιο θρεπτικό μέσο (LB στους 37 °C ) συμπεριλαμβανομένου και 84% του μεμβρανικού πρωτεϊνώματος (ΜΠ) και 89% του πρωτεϊνώματος του ΚΦ (ΠΚΦ) χωρίς το ΜΠ (ΠΚΦ-ΜΠ) (Πίνακας 5.1).

**Πίνακας 5.1 – Ανιχνεύσιμο πρωτεϊνωμα του *E. coli* BL21-DE3**

Σύνοψη του θεωρητικού, βασικού (δηλ. χωρίς ψευδογονίδια και κινητά στοιχεία; δες ενότητα 2.2) και ανιχνεύσιμου πρωτεϊνώματος (με βάση ένδειξης σε επίπεδο mRNA και πρωτεΐνης) (Εικόνα 5.1; ενότητα 6.3.1) για το συνολικό πρωτεϊνωμα του *E. coli* BL21-DE3 και για τρία υποσύνολα αυτού (ΠΚΦ: πρωτεϊνωμα κυτταρικού φακέλου, ΜΠ: μεμβρανικό πρωτεϊνωμα, ΠΚΦ-ΜΠ: η διαφορά ανάμεσα στα δύο προηγούμενα). Το ανιχνεύσιμο πρωτεϊνωμα αναλύεται σε επίπεδο mRNA και πρωτεΐνης, αυτό που ανιχνεύτηκε στην παρούσα μελέτη και σε παλαιότερες πρωτεομικές αναλύσεις (άλλες μελέτες Πίνακας 6.10).

Υποσύνολο πρωτεϊνώματος	Συνολικό θεωρητικό πρωτεϊνωμα	Συνολικό βασικό θεωρητικό πρωτεϊνωμα	Ανιχνεύσιμο	mRNA	Ανιχνεύτηκε σε επίπεδο πρωτεΐνης (παρούσα μελέτη)	Βασικό Πρωτεϊνωμα που ανιχνεύτηκε σε επίπεδο πρωτεΐνης (παρούσα μελέτη)	Βασικό Πρωτεϊνωμα που ανιχνεύτηκε σε επίπεδο πρωτεΐνης (άλλες μελέτες)
<b>BL21-DE3</b>	<b>4842</b>	<b>4559</b>	<b>3943</b>	3550	2255	2195	3101
	100%	94.2%	81.4%	73.3%	46.6%	55.7%	64.0%
<b>ΠΚΦ</b>	<b>2324</b>	<b>2265</b>	<b>2023</b>	1817	1216	1196	1536
	100%	97.5%	87.0%	78.2%	52.3%	59.1%	66.1%
<b>ΜΠ</b>	<b>1079</b>	<b>1047</b>	<b>910</b>	833	515	503	602
	100%	97.0%	84.3%	77.2%	47.7%	55.3%	55.8%
<b>ΠΚΦ-ΜΠ</b>	<b>1245</b>	<b>1218</b>	<b>1113</b>	984	701	693	934
	100%	97.8%	89.4%	79.0%	56.3%	62.3%	75.0%

---

Επίσης για το 37% του πρωτεϊνώματος του ΚΦ έχει μετρηθεί η συγκέντρωση μέσα στο κύτταρο. Οι ποσότητες κυμαίνονται από πολύ μικρές μέχρι μεγάλες [0.15-3350 μόρια ανά κύτταρο; 0.32-1538 για το ΜΠ, (Taniguchi et al, 2010)]. Η χαμηλή ποσότητα έκφρασης μιας πρωτεΐνης ίσως υπονομεύει τη πειραματική της ανίχνευση από μεθόδους πρωτεομικής ανάλυσης (Bernsel et al, 2009; Speers et al, 2007; Weiner et al, 2008).

Η ανίχνευση των μεμβρανικών πρωτεϊνών επηρεάζεται σε μεγάλο βαθμό από την προετοιμασία του δείγματος αλλά και από την μετέπειτα ανίχνευση μέσω φασματομετρία μάζας. Κατά συνέπεια ανάλογα με το υποσύνολο του πρωτεϊνώματος που επιθυμούμε να ανιχνεύσουμε ενδέχεται να απαιτείται βελτιστοποίηση της πειραματικής διαδικασίας.

Για την απλή ταυτοποίηση μιας πρωτεΐνης η ανίχνευση ενός με δύο πεπτιδίων είναι αρκετή χρησιμοποιώντας σύγχρονους φασματογράφους. Συνεπώς μια μεμβρανική πρωτεΐνη μπορεί να ταυτοποιηθεί μέσω της ανίχνευσης 1-2 πολικών πεπτιδίων με άμεσες και γρήγορες προσεγγίσεις (shotgun approaches) (Yates, 1998). Εάν ο στόχος μιας ανάλυσης είναι η μεγαλύτερη κάλυψη της πρωτεϊνικής αλληλουχίας (protein coverage), δηλαδή η ανίχνευση όσον το δυνατόν περισσότερων διαφορετικών πεπτιδίων, τότε επιβάλλεται η εφαρμογή βημάτων βιοχημικής κλασματοποίησης σε επίπεδο, πρωτεΐνης, πεπτιδίων ή μεμβρανών. Τέλος για το στάδιο της ποσοτικοποίησης των πρωτεϊνών χρησιμοποιώντας μεθόδους χωρίς χημική επισήμανση των πρωτεϊνών (label-free methods), απαιτούνται περίτεχνες πειραματικές μεθοδολογίες που παρακάμπτουν τα τεχνικά εμπόδια. Για παράδειγμα βέλτιστη διαλυτοποίηση και αποτελεσματική πρωτεόλυση για να εμπλουτιστούν τα κλάσματα των μεμβρανικών πρωτεϊνών και να διασφαλιστεί η ανίχνευση των υδρόφοβων πεπτιδίων τους

Προς την κατεύθυνση αυτή, με τις τεχνολογικές καινοτομίες (Makarogon, 2000) ορισμένοι τεχνικοί περιορισμοί έχουν πλέον ξεπεραστεί επιτρέποντας ακόμα και την ποσοτικοποίηση μεμβρανικών κλασμάτων (Masuda et al, 2009). Ωστόσο, εάν συνυπολογίσουμε τις πρωτεομικές αναλύσεις μέχρι σήμερα, έχει ταυτοποιηθεί μόνο το ~64% του ανιχνεύσιμου πρωτεϊνώματος σε συνθήκες πλούσιου θρεπτικού μέσου (Πίνακας 5.1, Πίνακας 6.10) ενώ μόνο για το 35% έχει υπολογιστεί η ποσότητα (Ishihama et al, 2008; Masuda et al, 2009; Taniguchi et al, 2010).

### 5.3 Μελέτη των φυσικοχημικών χαρακτηριστικών των πρωτεϊνών του ΚΦ και των πεπτιδίων που παράγονται από αυτές στο στάδιο της πρωτεόλυσης

Μία από τις βασικές προκλήσεις της πρωτεομικής ανάλυσης του ΚΦ είναι η σαφής ανίχνευση των πρωτεϊνών που αλληλεπιδρούν με τις μεμβράνες (περιφερικές πρωτεΐνες) αλλά κυρίως των μεμβρανικών πρωτεϊνών. Αποφασίσαμε να διερευνήσουμε τις ιδιότητες του θεωρητικού ΠΚΦ και να προσδιορίσουμε σε τι ποσοστό είναι ανιχνεύσιμες από μεθόδους φασματομετρίας μάζας.

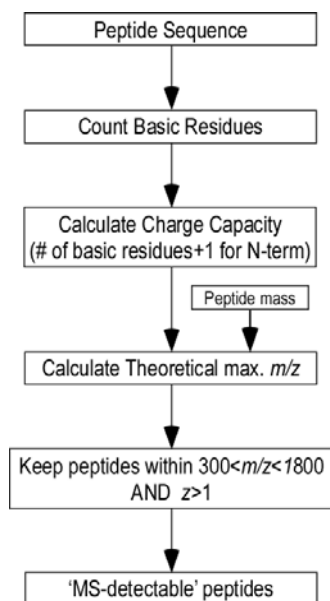
Ως εκ τούτου, αναλύσαμε υποσύνολα του ΠΚΦ σε σχέση με το ισοηλεκτρικό σημείο (pI), το μήκος και την υδροφοβικότητα τους (Εικόνα 5.4). Εξαιτίας των διαφορετικών φυσικοχημικών ιδιοτήτων που παρουσιάζει το ΜΠ αναλύθηκαν ξεχωριστά από το ΠΚΦ παρόλο που αποτελεί κομμάτι του. Οι κατανομές του ισοηλεκτρικού σημείου των κυτταροπλασματικών πρωτεϊνών και του ΠΚΦ-ΜΠ προέκυψαν παρόμοιες αν και ασύμμετρες (Εικόνα 5.4A). Οι κατανομές αυτές προέκυψαν διτροπικές (δύο τοπικά μέγιστα) παρουσιάζουν ένα μεγαλύτερο μέγιστο στην τιμή ~5.8 και ένα μικρότερο στην τιμή ~9.2. Οι μεμβρανικές πρωτεΐνες παρουσιάζουν ελάχιστα ενισχυμένη πιθανότητα εμφάνισης στο σημείο ~9.4 ενώ μειωμένη στο σημείο ~6 (Εικόνα 5.4A, κόκκινη καμπύλη) σε σχέση με τις δύο άλλες κατηγορίες.

Η ανάλυση του μήκους των πρωτεϊνών συναρτήσκει της υδροφοβικότητας (Kyte et al, 1982) έδειξε ότι οι κυτταροπλασματικές πρωτεΐνες και το ΠΚΦ-ΜΠ έχουν παρόμοιες κατανομές και ελάχιστα υδρόφιλο χαρακτήρα. Αντίθετα το ΜΠ φαίνεται να ακολουθεί διτροπική κατανομή με μια μετατόπιση προς τις πιο υδρόφοβες τιμές (Εικόνα 5.4B).

Στη συνέχεια προβλέψαμε τον αριθμό των διαμεμβρανικών περιοχών (TMs) χρησιμοποιώντας τα BE: TMHMM (Krogh et al, 2001) και Phobius (Käll et al, 2004) (Εικόνα 5.4C; δισδιάστατη κατανομή). Οι πρωτεΐνες με ένα έως δύο ΔΠ (κίτρινο) κυμαίνονται σε μεγάλο εύρος μήκους και υδροφοβικότητας ξεκινώντας από υδρόφιλες τιμές (συχνά λόγω της ύπαρξης μεγάλων και διαλυτών περιοχών) έως έντονα υδρόφοβες (συχνά μικρές πρωτεΐνες). Μεμβρανικές πρωτεΐνες που ανήκουν στην κατηγορία των 3-6 ΔΠ δεν εκκινούνται σε τόσο μεγάλα εύρη υδροφοβικότητας ενώ οι περισσότερες έχουν έντονα υδρόφοβο χαρακτήρα (πράσινο). Στην ακραία περίπτωση των πρωτεϊνών με περισσότερες από 7 ΔΠ παρατηρούμε ότι η κατανομή αφορά πρωτεΐνες

μεγαλύτερου μήκους (>200 αμινοξέα; καφέ κατανομή) και αντιστοιχεί στο δεύτερο μέγιστο της διτροπικής κατανομής των ΜΠ (κόκκινη γραμμή).

Έπειτα πραγματοποιήσαμε *in-silico* πρωτεόλυση με τρυψίνη (*in-silico* trypsin digestion) όπως γίνεται και πειραματικά στην από κάτω προς τα πάνω προσέγγιση της πρωτεομικής ανάλυσης ('bottom-up' approach). Αναλύσαμε τις φυσικοχημικές ιδιότητες που προκύπτουν από το σύνολο των θεωρητικών πεπτιδίων (Εικόνα 5.4D-E) αλλά και των ανιχνεύσιμων πεπτιδίων (Εικόνα 5.3). Στην παρούσα ανάλυση για να ανιχνευτεί πειραματικά ένα πεπτίδιο θα πρέπει ο λόγος μάζα προς φορτίο βρίσκεται στο εύρος  $300 < m/z < 1800$  σύμφωνα με τα όρια ανίχνευσης που εφαρμόσαμε στον φασματογράφο. Τα κριτήρια που πρέπει να πληρεί ένα πεπτίδιο για να ανιχνευτεί είναι: α) να επιδέχεται δύο και παραπάνω φορτία (καθώς μόνο πολλαπλά φορτισμένα πεπτίδια επιλέγονται για δεύτερου επιπέδου διαχωρισμό (MS/MS) ), β) να έχει μοριακό βάρος τουλάχιστον 600Da και γ) ο λόγος μάζας προς μέγιστο αριθμό φορτίων να μην υπερβαίνει τα 1800Da. Τα πεπτίδια που ικανοποιούν τα παραπάνω κριτήρια τα ονομάζουμε ανιχνεύσιμα ('MS-detectable'; Εικόνα 5.3). Μόνο το 1-2% των πεπτιδίων των κυτταροπλασματικών πρωτεϊνών ενώ το 8% των πεπτιδίων του ΜΠ δεν είναι ανιχνεύσιμα (not 'MS-detectable') (τα συγκεκριμένα δεδομένα δεν παρουσιάζονται).



**Εικόνα 5.3 Διάγραμμα ροής των κριτηρίων για την επιλογή των ανιχνεύσιμων πεπτιδίων με φασματομετρία μάζας ('MS-detectable' peptides)**

Πραγματοποιήσαμε *in-silico* πέψη του πρωτεϊνώματος για κάθε πεπτίδιο υπολογίσαμε το μέγιστο αριθμό πρόσληψης φορτίων με βάση τον αριθμό των βασικών αμινοξέων (Αργινίνη, Λυσίνη, Ιστιδίνη) και προσθέτοντας ένα επιπλέον φορτίο που μπορεί να λάβει το αμινοτελικό άκρο. Έπειτα υπολογίσαμε τον θεωρητικό λόγο  $m/z$  διαιρώντας την μάζα και το μέγιστο φορτίο. Ως ανιχνεύσιμα πεπτίδια ('MS-detectable') θεωρήσαμε αυτά με λόγο  $m/z$  μέσα στα όρια 300-1800 Da.

Τα ισοηλεκτρικά σημεία των θεωρητικών και ανιχνεύσιμων πεπτιδίων είναι ίδια και για τις τρεις κατηγορίες (καμπύλες με συνεχόμενες ή διακεκομμένες γραμμές Εικόνα 5.4D). Σημαντική διαφορά παρατηρούμε όμως, ανάμεσα στα πεπτίδια του ΜΠ και στα πεπτίδια των κυτταροπλασματικών πρωτεϊνών και του ΠΚΦ-ΜΠ σε επίπεδο μήκους έναντι

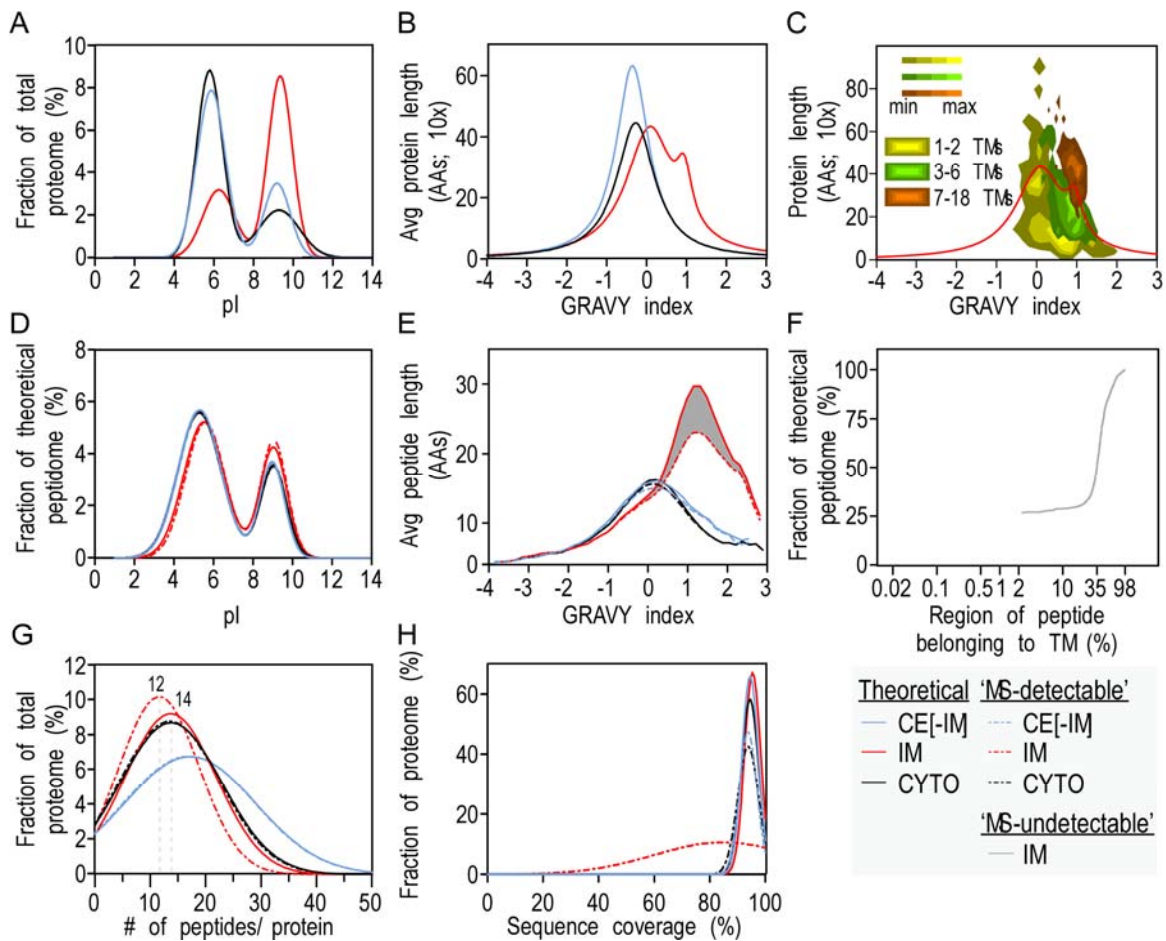
---

υδροφοβικότητας (Εικόνα 5.4Ε). Τα πεπτίδια του ΜΠ στην πλειοψηφία τους έχουν υψηλές τιμές υδροφοβικότητας (GRAVY index). Εάν επιλέξουμε τα ανιχνεύσιμα πεπτίδια (Εικόνα 5.4Ε; διακεκομμένη κόκκινη γραμμή) τότε αρκετά πεπτίδια του ΜΠ εξαιρούνται ως μη πιθανά αν ανιχνευθούν λόγω μικρής ικανότητας πρόσληψης φορτίων (Εικόνα 5.4Ε; γκρι περιοχή). Η σύγκριση των ιδιοτήτων των ανιχνεύσιμων πεπτιδίων του ΜΠ και των άλλων δύο κατηγοριών υποδεικνύει ότι στην πρώτη περίπτωση τα πεπτίδια είναι μεγαλύτερα και περισσότερο υδρόφοβα. Ανάλογα με την μέθοδο προετοιμασίας του δείγματος αυτά τα πεπτίδια ενδέχεται να χάνονται όπως για παράδειγμα συμβαίνει σε συγκεκριμένες μεθόδους όπως η εξαγωγή από πηκτώματα πολυακρυλαμίδης (gel-extraction).

Επιπλέον, εξετάσαμε εάν τα ανιχνεύσιμα πεπτίδια του ΜΠ σε σύγκριση τα μη ανιχνεύσιμα αντιστοιχούν σε ΔΠ των αντίστοιχων πρωτεϊνών. Για το λόγο αυτό αντιστοιχήσαμε τις αλληλουχίες των πεπτιδίων στις αντίστοιχες ΔΠ περιοχές που προβλέπει το Phobius (Kall et al, 2007) και υπολογίσαμε το ποσοστό των αμινοξέων των πεπτιδίων που βρίσκονται σε ΔΠ (Εικόνα 5.4D). Είναι ξεκάθαρο ότι τα ανιχνεύσιμα πεπτίδια περιέχουν μικρό αριθμό αμινοξέων που ανήκουν σε ΔΠ (< 1% κόκκινη διακεκομμένη γραμμή) σε αντίθεση με την πλειοψηφία των μη ανιχνεύσιμων πεπτιδίων που αποτελούνται κατά μεγάλο ποσοστό από διαμεμβρανικές περιοχές (συνεχόμενη γκρι γραμμή ; >2% και κατά μέσο όρο 35% περιεκτικότητα σε ΔΠ)

Η εξαίρεση των μη ανιχνεύσιμων πεπτιδίων αντιστοιχεί σε μόνο 10 πρωτεΐνες που αποκλείονται εντελώς από την ανάλυση ΦΜ καθώς από αυτές προκύπτουν μόνο μη ανιχνεύσιμα πεπτίδια (τα συγκεκριμένα δεδομένα δεν παρουσιάζονται). Οι 1062 μεμβρανικές πρωτεΐνες που απομένουν είναι ανιχνεύσιμες καθώς αναμένουμε να τις ταυτοποιήσουμε αξιόπιστα από κάποιο μοναδικό πολικό πεπτίδιο. Δεδομένου ότι το ΜΠ έχει λιγότερα ανιχνεύσιμα πεπτίδια ανά πρωτεΐνη (12 έναντι >15 πεπτιδίων για τις κυτταροπλασματικές και του ΠΚΦ-ΜΠ; Εικόνα 5.4G), οι πιθανότητες ταυτοποίησης τους είναι μικρότερη. Οι επιπλέον απώλειες των υδρόφοβων πεπτιδίων στο στάδιο της προετοιμασίας των δειγμάτων ενδέχεται σε ορισμένες περιπτώσεις να μειώσει δραματικά το ποσοστό κάλυψη μιας μεμβρανικής πρωτεΐνης (Εικόνα 5.4H). Σε αντίθεση με το ΜΠ η θεωρητική κάλυψη των πρωτεϊνών των άλλων δύο κατηγοριών επηρεάζεται ελάχιστα από την εξαίρεση των μη ανιχνεύσιμων πεπτιδίων (Εικόνα 5.4H).





**Εικόνα 5.4 – Φυσικοχημικές ιδιότητες των κυτταροπλασματικών (CYTO), μεμβρανικών (IM) πρωτεϊνών και των πεπτιδίων τους.**

**A-B.** Ισοηλεκτρικό σημείο, υδροφοβικότητα και μέσο μήκος των πρωτεϊνών ανά κατηγορία (δες ενότητα 6.3);

**C.** τρισδιάστατη κατανομή της υδροφοβικότητας σε σχέση με το μήκος για τις ΜΠ η οποία αναπαρίσται με ένα διάγραμμα χρωματικής πυκνότητας (contour plot). Οι ΜΠ χωρίζονται σε τρεις υποκατηγορίες με βάση τον αριθμό των ΜΠ που προβλέπονται από το TMHMM (Kall et al, 2004) και το Phobius (Kall et al, 2007).

**D-E** Κατανομές του ισοηλεκτρικού σημείου και της υδροφοβικότητας συναρτήσει του μέσου μήκους των θεωρητικών (συνεχόμενες γραμμές) και ανιχνεύσιμων (διακεκομμένες γραμμές) πεπτιδίων ανά υποκατηγορία. Η γκρι περιοχή αντιστοιχεί σε μεμβρανικά πεπτιδία τα οποία εξαιρέθηκαν αφού εφαρμόστηκαν τα κριτήρια που ορίζουν τα ανιχνεύσιμα πεπτιδία (**Εικόνα 5.3**; 'MS-detectable' peptides).

**F.** Κατανομή της αθροιστική πιθανότητα για την περιεκτικότητα των πεπτιδίων σε μεμβρανικά αμινοξέα (ποσοστό του πεπτιδίου που ανήκει σε μεμβρανική/ες περιοχή/ές). Δύο σύνολα πεπτιδίων συγκρίνονται, τα ανιχνεύσιμα ('MS-detectable') και τα μη ανιχνεύσιμα ('MS non-detectable').

---

**H, I** Κατανομές του αριθμού των ανιχνεύσιμων πεπτιδίων ανά πρωτεΐνη και της κάλυψης των πρωτεϊνών (ποσοστό αμινοξέων της αλληλουχίας που ανιχνεύθηκαν μέσω αντίστοιχων πεπτιδίων τρυψίνης).

Με συνεχόμενη γραμμή συμβολίζονται οι καμπύλες που προέκυψαν με την ανάλυση του θεωρητικού συνόλου πεπτιδίων ενώ με διακεκομμένη αντιστοιχούν στο υποσύνολο των ανιχνεύσιμων πεπτιδίων (MS-detectable)

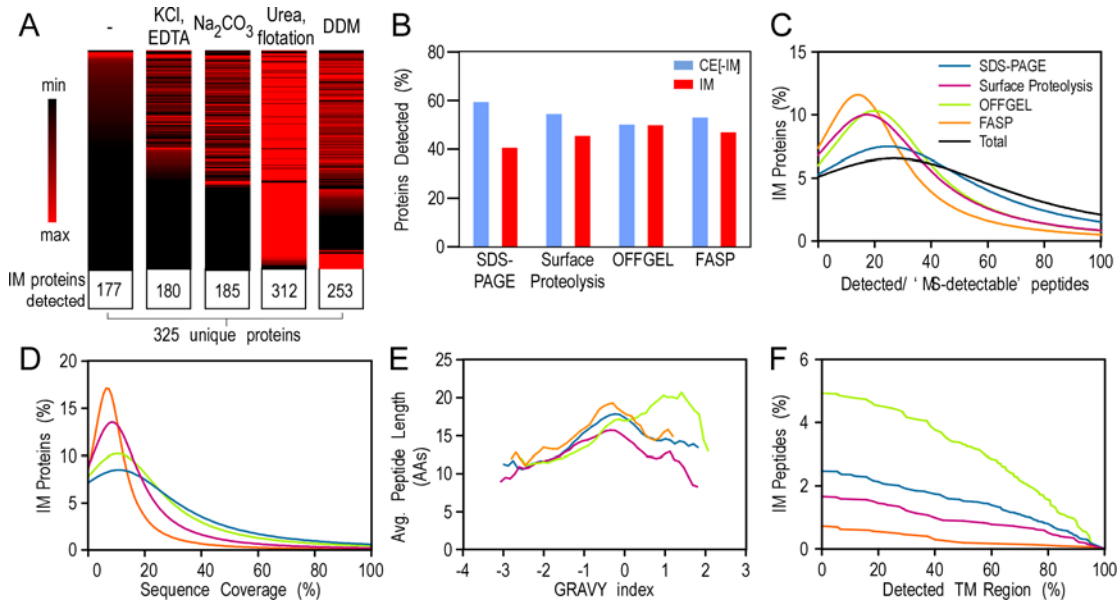
*CYTO*: Κυτταροπλασματικές πρωτεΐνες; *IM*: μεμβρανικές πρωτεΐνες (ΜΠ); *CE-[IM]*: πρωτεΐνωμα του ΚΦ χωρίς τις μεμβρανικές πρωτεΐνες (ΠΚΦ-ΜΠ); *Theoretical*: θεωρητικό πρωτεΐνωμα; *MS-detectable*: ανιχνεύσιμο με φασματομετρία μάζας; *GRAVY*: μέση υδροφοβικότητα πρωτεΐνης

---

#### 5.4 Πειραματική ανίχνευση των πρωτεϊνών του κυτταρικού φακέλου

Η πειραματική ανίχνευση του ΠΚΦ έγινε μέσω ανεστραμμένων κυστιδίων της εσωτερικής μεμβράνης (AMK; IMVs: inverted membrane vesicles ), που παράγονται από το στέλεχος BL21-DE3. Τα AMK επεξεργάστηκαν με διάφορους παράγοντες ώστε να γίνουν πιο πλούσια σε πρωτεϊνικό περιεχόμενο (Papanastasiou et al, 2013 ).

Τέσσερις μέθοδοι προετοιμασίας του δείγματος εφαρμόστηκαν: Α) αποδιατακτικά πηκτώματα πολυακρυλαμίδης (SDS-PAGE) και αποκοπή των αντίστοιχων πρωτεϊνικών ζωνών σε συνδυασμό με υγρή χρωματογραφία- ΦΜ (**nano-LC-MS-MS**) για πιο ολοκληρωμένη ανάλυση του ΠΚΦ, Β) πρωτεόλυση στην επιφάνεια των AMK, Γ) πρωτεόλυση σε διάλυμα από διαλυτοποιημένα πεπτίδια από AMK με ακόλουθη κλασματοποίηση απουσία πηκτώματος (OFFGEL; (Horth et al, 2006)) και τέλος προετοιμασία δείγματος με διήθηση (FASP; (Wisniewski et al, 2009b)) (Εικόνα 5.5; (Papanastasiou et al, in preparation)). Στη συνέχεια παρουσιάζουμε και συγκρίνουμε τα αποτελέσματα κάθε μεθόδου.



**Εικόνα 5.5 – Ανίχνευση μεμβρανικών πρωτεϊνών με διαφορετικές μεθόδους προετοιμασίας δείγματος, σύγκριση φυσικοχημικών ιδιοτήτων των αντίστοιχων πεπτιδίων**

**A.** Ενίσχυση των μεμβρανικών πεπτιδίων στα διάφορα στάδια επεξεργασίας του δείγματος κατά την πρωτεόλυση στην επιφάνεια των ΜΚ (**Papanastasiou et al, in preparation**). **B.** Σχετικές ποσότητες των μεμβρανικών πρωτεϊνών (IM) και των υπόλοιπων πρωτεϊνών του ΚΦ (CE-IM) που ταυτοποιήθηκαν σε κάθε μέθοδο προετοιμασίας δείγματος. **D.** Σύγκριση των μεθόδων σε επίπεδο κάλυψης των μεμβρανικών πρωτεϊνών. Μεγαλύτερο ποσοστό κάλυψης εξασφαλίζεται για μεγαλύτερο μέρος μεμβρανικών πρωτεϊνών στην περίπτωση προετοιμασίας δείγματος απουσία πηκτώματος πολυακρυλαμίδης (OFFGEL). **E-F.** Σύγκριση των μεθόδων σε επίπεδο πεπτιδίων που ανιχνεύθηκαν ανα μέθοδο: (E) κατανομή μήκους συναρτήσεως υδροφοβικότητας των ταυτοποιημένων πεπτιδίων, (F) περιεκτικότητα σε διαμεμβρανικές περιοχές (ποσοστό αμινοξέων που ανήκουν σε ΔΜ).

*Αποδιατακτικά πηκτώματα πολυακρυλαμίδης (SDS-PAGE) σε συνδυασμό με υγρή χρωματογραφία – ΦΜ*

Τα κλάσματα των ΑΜΚ αναλύθηκαν είτε αυτούσια, όπως λαμβάνονται απευθείας από την προπαρασκευαστική καθίζηση σε βαθμίδες σακχαρόζης (σε μη επεξεργασμένα ΑΜΚ), είτε έπειτα από επεξεργασία με χημικούς παράγοντες, όπως ουσίες που διαταράσσουν την δομή των μορίων του νερού (chaotropes) και επίπλευση σε βαθμίδες σακχαρόζης (σε επεξεργασμένα ΑΜΚ) (Εικόνα 5.2D) και στη συνέχεια διαχωρίστηκαν σε αποδιατακτικά πηκτώματα πολυακρυλαμίδης (SDS-PAGE gels).

---

Η ανάλυση των μη επεξεργασμένων AMK οδήγησε στην ταυτοποίηση 763 πεπτιδίων: 76% προερχόμενα από πρωτεΐνες του ΚΦ (48% μεμβρανικές) και το 24% αποτέλεσαν κυτταροπλασματική «επιμόλυνση». Για την απομάκρυνση των πρωτεϊνών που παρέμειναν δεσμευμένες στις μεμβράνες και αυτών που εγκλωβίστηκαν στην διαδικασία παρασκευής των AMK, έγινε μετέπειτα επεξεργασία με υψηλής συγκέντρωσης «αλάτι» και ανθρακικού νατρίου, υψηλό pH και ουρία. Για καλύτερη απομόνωση των μεμβρανικών πρωτεϊνών από τα επεξεργασμένα κυστίδια, εφαρμόστηκε επίπλευση σε βαθμίδες σακχαρόζης (Gibeaut et al, 1990). Στο τέλος των παραπάνω βημάτων επεξεργασίας ο αριθμός των ταυτοποιημένων πρωτεϊνών του ΚΦ αυξήθηκε κατά 86%. Συνδυαστικά οι πρωτεΐνες που ταυτοποιήθηκαν από όλα τα στάδια της επεξεργασίας αντιστοιχούν σε 1092 μοναδικές πρωτεΐνες του ΚΦ (Εικόνα 5.6)

*Πρωτεόλυση στην επιφάνεια ανεστραμμένων μεμβρανικών κυστιδίων*

Η ταυτοποίηση μεμβρανικών πρωτεϊνών με αυτήν την μέθοδο βασίζεται κυρίως στα τμήματα των μεμβρανικών πρωτεϊνών που βρίσκονται εκτεθειμένα στο διάλυμα τα οποία α) περιέχουν κατάλοιπα Λυσίνης/Αργινίνης και β) παράγουν πεπτίδια κατάλληλου μοριακού βάρους (Griffin et al, 2011).

Τα διαλυτά τμήματα των μεμβρανικών πρωτεϊνών ποικίλουν από πολύ μικρά (π.χ. 6 κατάλοιπα όπως η τοξική πρωτεΐνη (Fozo et al, 2008; Hemm et al, 2008a)) μέχρι πολύ μεγαλύτερα (π.χ. 1158 κατάλοιπα στην καρβοξυτελική περιοχή της πρωτεΐνης FtsK (Dorazi et al, 2000))

Για την βελτίωση της μεθόδου και ενίσχυση του αριθμού των μεμβρανικών πρωτεϊνών που ταυτοποιούνται, πραγματοποιήθηκε πρωτεόλυση στην επιφάνεια των AMK. Στην συγκεκριμένη προσέγγιση τα AMK πρωτεολύθηκαν με τρυψίνη, τα πεπτίδια που προέκυψαν συλλέχθηκαν έπειτα από αφαίρεση των μεμβρανών και αναλύθηκαν με ΦΜ. Η τρυψίνη πρωτεολύει μόνο τα επιφανειακά τμήματα των πρωτεϊνών και συνεισφέρει επίσης στην ταυτοποίηση περιφερικών πρωτεϊνών (PIM proteins, F1 στην ορολογία του STEPdb).

Η επιφανειακή πρωτεόλυση των AMK οδήγησε στην ταυτοποίηση 510 πρωτεϊνών του ΚΦ από τις οποίες 177 είναι μεμβρανικές πρωτεΐνες. Η επεξεργασία των AMK με «αλάτι», ανθρακικό νάτριο και pH 11 πριν την επιφανειακή πρωτεόλυση βελτίωσε ελάχιστα την ταυτοποίηση των πρωτεϊνών του ΚΦ. Η επεξεργασία με ουρία και επίπλευση σε κλάσματα σακχαρόζης βελτίωσε σημαντικά την μέθοδο (συνολικά 312 μεμβρανικές πρωτεΐνες από τις 515 πρωτεΐνες του ΚΦ).

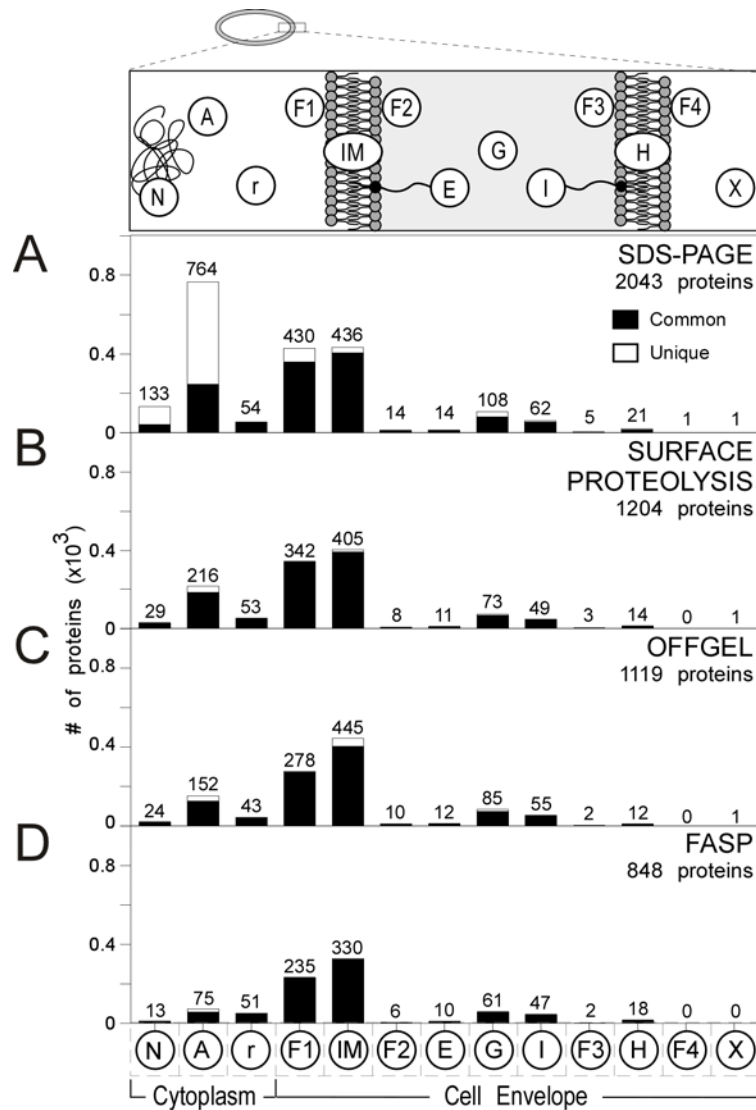
---

Εικάζουμε ότι οι συγκεκριμένες μεμβρανικές πρωτεΐνες «επισκιάζονταν» από περιφερικές πρωτεΐνες ή από κυτταροπλασματικές πρωτεΐνες λόγω τυχαίων αλληλεπιδράσεων. Επιπρόσθετα έγινε επεξεργασία των AMK με τον παράγοντα DDM (non-ionic detergent dodecyl maltoside) σε ικανές συγκεντρώσεις ώστε να διαταραχθούν οι υδρόφοβες αλληλεπιδράσεις (Papanastasiou et al, 2013 ) αλλά αρκετά χαμηλές ώστε να μην διαλυτοποιηθούν τα λιπίδια των AMK. Επτά επιπλέον μεμβρανικές πρωτεΐνες ταυτοποιήθηκαν με αυτή τη μέθοδο. Συνολικά ταυτοποιήθηκαν 906 πρωτεΐνες του ΚΦ από τις οποίες το 44% αντιστοιχεί σε μεμβρανικές πρωτεΐνες.

#### *Κλασματοποίηση απουσία πηκτώματος (OFFGEL)*

Τα επεξεργασμένα AMK πρωτεολύθηκαν σε διάλυμα αφού διαλυτοποιήθηκαν πλήρως μέσω του παράγοντα DDM (34mM, 200 x CMC) (Εικόνα 5.1D). Τα πεπτίδια διαχωρίστηκαν σε κλάσματα με βάση το ισοηλεκτρικό τους φορτίο σε υγρή φάση χρησιμοποιώντας ηλεκτροφόρηση απουσία πηκτώματος (OFFGEL electrophoresis; (Michel et al, 2003)). Επίσης διερευνήθηκε η ικανότητα της μεθόδου να ανιχνεύσει διαμεμβρανικά πεπτίδια σε διαφορετικές συνολικές ποσότητες πρωτεϊνών (50, 200 and 400 µg μεμβρανικών πρωτεϊνών).

Συνολικά ταυτοποιήθηκαν 900 πρωτεΐνες του ΚΦ, το 57% του ανιχνεύσιμου ΜΠ (Εικόνα 5.1). Τα πεπτίδια ανιχνεύτηκαν στην περιοχή pH 3.5-10 ενώ ελάχιστα βρέθηκαν στη περιοχή pH 7-8 κατά αντιστοιχία με την θεωρητική ανάλυση των πεπτιδίων (Cargile et al, 2003). Η διτροπική κατανομή του ισοηλεκτρικού σημείου φαίνεται να αποτελεί καθολικό χαρακτηριστικό στους προκαρυώτες ενώ φαίνεται να εξαρτάται από τον υποκυτταρικό εντοπισμό των πρωτεϊνών (Schwartz et al, 2001).



**Εικόνα 5.6 – Υποκυτταρικός εντοπισμός των πρωτεϊνών που ανιχνεύθηκαν με διαφορετικές μεθόδους προετοιμασίας δείγματος σε συνδυασμό με φασματομετρία μάζας**

Η κατανομή σε υποκυτταρικά διαμερίσματα των πρωτεϊνών που ανιχνεύθηκαν με Α) αποδιατακτικά πηκτώματα πολυακρυλαμίδης (SDS-PAGE), Β) πρωτεόλυση στην επιφάνεια των ανεστραμμένων μεμβρανικών κυστιδίων (AMK), Γ) πρωτεόλυση σε διάλυμα και ακόλουθη κλασματοποίηση απουσία πηκτώματος (OFFGEL) και Δ) προετοιμασία δείγματος με βοήθεια φίλτρων (FASP).

Η ταξινόμηση ακολουθεί τις κατηγορίες υποκυτταρικού εντοπισμού όπως ορίζονται στη βάση δεδομένων STEPdb (Εικόνα 5.1). Οι κοινές πρωτεΐνες (common) αναφέρονται στις πρωτεΐνες που ανιχνεύθηκαν σε όλες τις μεθόδους (μαύρο) ενώ οι μοναδικές (unique; άσπρο) σε αυτές που ανιχνεύθηκαν μόνο από την αντίστοιχη μέθοδο. Ο συνολικός αριθμός των πρωτεϊνών που ανιχνεύθηκαν ανά μέθοδο συνοψίζεται πάνω δεξιά. Η κατανομή των πρωτεϊνών που ανιχνεύθηκαν συνολικά από όλες τις μεθόδους απεικονίζεται στην Εικόνα 5.1.

---

### *Προετοιμασία δείγματος με διήθηση (FASP)*

Προς την ταυτοποίηση περισσότερου αριθμού μεμβρανικών πρωτεϊνών δοκιμάσαμε προετοιμασία δείγματος με διήθηση (FASP; (Wisniewski et al, 2009b)). Συνοπτικά, η μέθοδος FASP χρησιμοποιεί μια συσκευή υπερ-διήθησης μέσα στην οποία τα διαλυτοποιημένα πεπτιδία «συλλαμβάνονται» και πρωτεολύονται ενώ ταυτόχρονα γίνεται απομάκρυνση των χημικών παραγόντων. Τα πεπτιδία που προκύπτουν εγκλωβίζονται σε μια επιφάνεια διήθησης και απελευθερώνονται με την προσθήκη μικρο- ή μεγαλο- μοριακών ουσιών. Η τεχνική αυτή συνδυάζει τα πλεονεκτήματα των δύο μεθόδων: της πρωτεόλυσης σε πήκτωμα (οι χημικοί παράγοντες απομακρύνονται) αλλά και της πρωτεόλυσης σε διάλυμα (η πρωτεόλυση γίνεται σε συνθήκες αποδιατακτικές) και είναι ιδιαίτερα κατάλληλη για την μελέτη μεμβρανικών πρωτεϊνών (Wisniewski et al, 2009a). Τα επεξεργασμένα AMK διαλυτοποιήθηκαν με παράγοντα SDS (2%) και στη συνέχεια εισήχθησαν στην συσκευή υπερ-διήθησης (Wisniewski et al, 2009b).

Συνολικά ταυτοποιήθηκαν 709 πρωτεΐνες του ΚΦ, 47% αποτελούν μεμβρανικές πρωτεΐνες (Εικόνα 5.6D). Παρόλο που η διηθητική μεμβράνη που χρησιμοποιήθηκε έχει κατώτατο όριο στα 30 kDa, το ένα τρίτο των πρωτεϊνών που ανιχνεύθηκαν ήταν μικρότερου μοριακού βάρους γεγονός το οποίο καταδεικνύει ότι δεν απομακρύνθηκαν κατά τα στάδια «ξεπλύματος». Η ερμηνεία που αποδώσαμε είναι ότι τα όρια αυτά ίσως αντιστοιχούν σε αναδιπλωμένες πρωτεΐνες.

### **5.5 Σύγκριση μεταξύ των πειραματικών μεθόδων**

Αποφασίσαμε να διερευνήσουμε εάν υπάρχει κάποια προδιάθεση στην ανίχνευση πεπτιδίων με συγκεκριμένες ιδιότητες από κάθε μέθοδο. Συγκεντρώσαμε τα αποτελέσματα από όλες βιολογικές και τις τεχνικές επαναλήψεις ανά μέθοδο προετοιμασίας δείγματος και συγκρίναμε τον αριθμό των πρωτεϊνών που ανιχνεύθηκαν (Εικόνα 5.5B). Με τις μεθόδους SDS-PAGE και πρωτεόλυσης στην επιφάνεια AMK ταυτοποιήθηκαν περισσότερες πρωτεΐνες του ΠΚΦ-ΜΠ και αυτό αποδίδεται στο γεγονός ότι τα AMK αναλύθηκαν αμέσως μετά την κλασματοποίηση τους (Εικόνα 5.1D). Τα αντίστοιχα δείγματα δεν αναλύθηκαν με τις μεθόδους OFFGEL και FASP. Οι επιπλέον πρωτεΐνες που ταυτοποιήθηκαν αφορούν κυρίως περιφερικές πρωτεΐνες που απομακρύνθηκαν κατά την επεξεργασία των AMK με χημικούς παράγοντες που ακολούθησε (Papanastasiou et al, 2013 ).

---

Εάν συγκρίνουμε τα δείγματα που προέκυψαν έπειτα από επίπλευση του κλάσματος των ΑΜΚ σε βαθμίδες σακχαρόζης, τότε τα σχετικά ποσοστά ανάμεσα στο ΜΠ και του ΠΚΒ-ΜΠ που ανιχνεύθηκαν είναι σχεδόν τα ίδια σε όλες τις μεθόδους εκτός της επιφανειακής πρωτεόλυσης όπου έφτασε μέχρι και το 25%. Από το ΠΚΦ-ΜΠ ταυτοποιήθηκαν περιπλασμικές πρωτεΐνες αλλά και πρωτεΐνες της εξωτερικής μεμβράνης (Εικόνα 5.6) περισσότερες απ αυτές ταυτοποιήθηκαν με την μέθοδο SDS-PAGE πιθανότατα λόγω ότι εγκλωβίστηκαν στα ΑΜΠ και συνέχεια απελευθερώθηκαν με επεξεργασία του αποδιατακτικού παράγοντα SDS.

Στη συνέχεια συγκρίναμε τον αριθμό των ταυτοποιημένων πεπτιδίων ανά πρωτεΐνη (Εικόνα 5.5C) και την κάλυψη των αλληλουχιών (Εικόνα 5.5D), μεταξύ των τεσσάρων μεθόδων. Για τα πειραματικά ανιχνευμένα πεπτίδια που είχαν ένα έως δύο σημεία αποτυχίας αποκοπής από της Τρυψίνη, εφαρμόσαμε *in-silico* πρωτεόλυση και συμπεριλήφθησαν στην ανάλυση μόνο όσα προέκυψαν μεγαλύτερα των 4 αμινοξέων (αντίστοιχα με την θεωρητική ανάλυση; Εικόνα 5.4). Δεν παρατηρήθηκαν ιδιαίτερες διαφορές ανάμεσα στις μεθόδους όσον αφορά τα συγκεκριμένα χαρακτηριστικά.

Έπειτα διερευνήσαμε εάν κάποια μέθοδος συνεισφέρει περισσότερα πεπτίδια με συγκεκριμένες φυσικοχημικές ιδιότητες (Εικόνα 5.5E-F) . Με την μέθοδο OFFGEL ανιχνεύονται πολλά περισσότερα μεγάλα και υδρόφοβα πεπτίδια (Εικόνα 5.5E) τα οποία μάλιστα φαίνεται να αποτελούνται σε μεγάλο ποσοστό από μεμβρανικά τμήματα (Εικόνα 5.5F) και τα οποία εν τέλει συνεισφέρουν στο συνολικό αριθμό των ταυτοποιημένων μεμβρανικών πρωτεϊνών (42 επιπλέον μεμβρανικές πρωτεΐνες από το OFFGEL) αλλά και στην μεγαλύτερη κάλυψη των αλληλουχιών.

Συνδυαστικά από όλες τις μεθόδους προετοιμασίας δείγματος ταυτοποιήθηκαν 1216 πρωτεΐνες του ΚΦ (Εικόνα 5.1), 515 μεμβρανικές (~60% του ΠΚΦ και 57% του ανιχνεύσιμου ΜΠ σε πλούσιο θρεπτικό μέσο).

## **5.6 Μελέτη των φυσικοχημικών χαρακτηριστικών των μεμβρανικών πρωτεϊνών που δεν ανιχνεύθηκαν με καμία μέθοδο**

Στη συνέχεια διερευνήσαμε τις πιθανές αιτίες που οδήγησαν στην απώλεια ανίχνευσης των υπολειπομένων 403 (533 θεωρητικές από τις οποίες 403 ανήκουν στο βασικό πρωτεϊνώμα) μεμβρανικών πρωτεϊνών (Εικόνα 5.7). Από αυτές οι 371 έχουν ανιχνευθεί σε επίπεδο μεταγραφόμενου mRNA και 99 σε επίπεδο πρωτεΐνης (Patten et al, 2004a; Vijayendran et al,



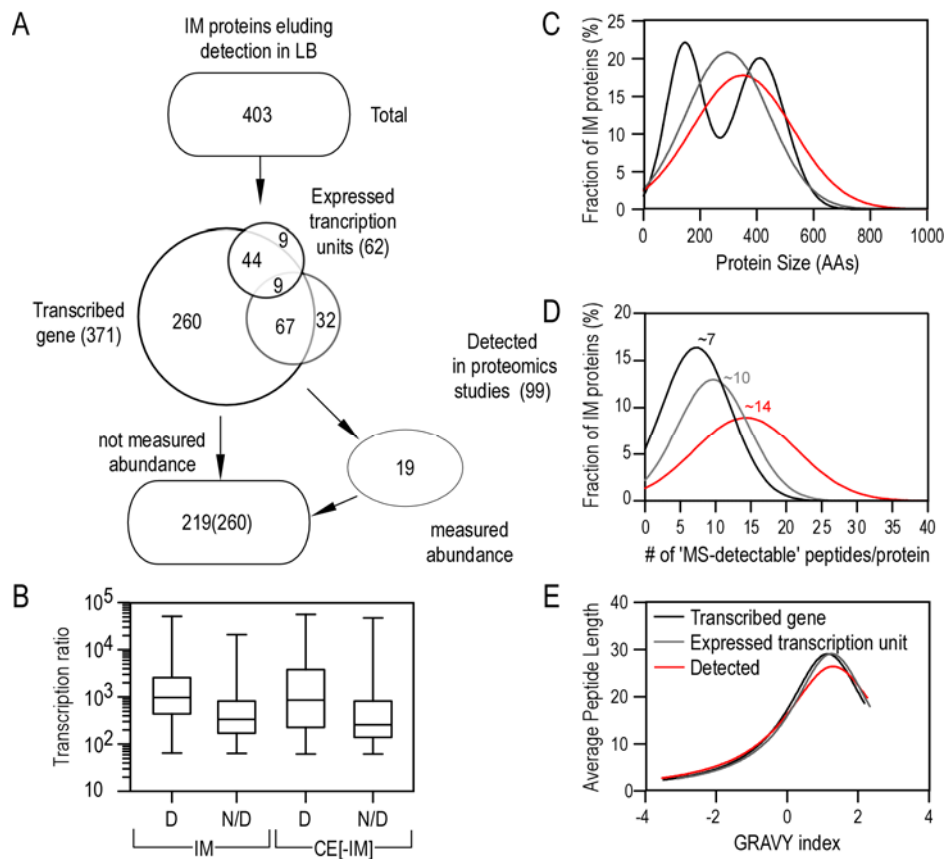
---

2007) 62 κωδικοποιούνται σε μεταγραφικές μονάδες (TUs) μαζί με άλλες πρωτεΐνες που τις ανιχνεύσαμε (δες ενότητα 6.3.4). Υποθέτουμε ότι μια μεταγραφική μονάδα που κωδικοποιεί παραπάνω από μία πρωτεΐνες (polycistronic mRNA), εκφράζεται όταν τουλάχιστον ένα από τα πεπτίδια που κωδικοποιεί έχουν ανιχνευθεί πειραματικά (Πίνακας 6.11; δες ενότητα 6.3.4). Συνολικά 403 από τις 533 πρωτεΐνες αναμένεται να εκφράζονται στις πειραματικές συνθήκες που μελετάμε (Εικόνα 5.7A).

Η μη ταυτοποίηση αυτών των πρωτεϊνών μπορεί να οφείλεται σε χαμηλά επίπεδα σύνθεσης (σε σχέση με όσες ανιχνεύθηκαν στην παρούσα ανάλυση) (Εικόνα 5.7B). Όμως δεν θεωρούμε ότι αυτός είναι ο μοναδικός παράγοντας καθώς πολλές από τις πρωτεΐνες που έχουμε ταυτοποιήσει παράγονται σε μικρές ποσότητες (π.χ. FadE με 10 και CysC με 16 μόρια ανά κύτταρο).

Εναλλακτικοί λόγοι μπορούν να σχετίζονται με το μικρό μέγεθος μιας πρωτεΐνης και την παραγωγή μικρού αριθμού πεπτιδίων. Οι 403 πρωτεΐνες αναλύθηκαν σε σχέση με το μήκος και τον αριθμό των πεπτιδίων που παράγουν. Χωρίστηκαν σε τρεις κατηγορίες 260 (από τις 371; δες λεζάντα εικόνας Εικόνα 5.7) πρωτεΐνες που έχουν βρεθεί μόνο σε επίπεδο mRNA, αυτές που ανήκουν σε μεταγραφικές μονάδες που εκφράζονται και αυτές που έχουν ανιχνευθεί σε άλλες πρωτεομικές αναλύσεις (Εικόνα 5.7C-E).

Οι πρωτεΐνες που έχουν ανιχνευθεί σε επίπεδο mRNA ακολουθούν διτροπική κατανομή (τοπικά μέγιστα στα σημεία ~160 και ~420) ένδειξη ότι στο σύνολο εμπεριέχονται πολλές «κοντές» πρωτεΐνες (Εικόνα 5.7C). Η *in silico* πρωτεόλυση σε συνδυασμό με τα κριτήρια που εφαρμόσαμε αποδεικνύει ότι αυτές οι πρωτεΐνες παράγουν κατά μέσο όρο ~7 ανιχνεύσιμα πεπτίδια σε σύγκριση με τις πρωτεΐνες των μεταγραφικών μονάδων (~10 και ~14 για το ταυτοποιημένο ΜΠ, Εικόνα 5.7D). Όλες οι πρωτεΐνες έχουν παρόμοια χαρακτηριστικά με αυτές που ταυτοποιήθηκαν (αντίστοιχα μήκη και υδροφοβικότητα πεπτιδίων Εικόνα 5.7E). Συνεπώς θεωρούμε ότι ο μικρός αριθμός πεπτιδίων ανά πρωτεΐνη αποτελεί τον πιο πιθανό παράγοντα για την απώλεια αυτών των πρωτεϊνών.



**Εικόνα 5.7 – Ανάλυση των μεμβρανικών πρωτεϊνών που δεν ταυτοποιήθηκαν**

**A.** Διάγραμμα σύνοψης των πρωτεϊνών που δεν ανιχνεύθηκαν: από τις 533 μεμβρανικές πρωτεΐνες που δεν ταυτοποιήθηκαν από καμία μέθοδο που εφαρμόσαμε, οι 403 αναμένουμε να είναι ανιχνεύσιμες σε συνθήκες πλούσιου θρεπτικού (LB; δες ενότητα 6.3.1). Από αυτές οι 371 έχει βρεθεί ότι μεταγράφονται σε επίπεδο mRNA (transcribed gene), 99 έχουν ανιχνευθεί σε επίπεδο πρωτεΐνης και 62 ανήκουν σε μεταγραφικές μονάδες (transcription units) που εκφράζονται (δες ενότητα 6.3.4), ενώ για μόνο 19 έχουν μετρηθεί οι ποσότητες μέσα στο κύτταρο (Iwasaki et al, 2010; Masuda et al, 2009; Taniguchi et al, 2010)

**B.** Οι πρωτεΐνες του ΜΠ και του ΠΚΦ-ΜΠ αναλύθηκαν σε σχέση με τα επίπεδα μεταγραφής, όπως έχουν μετρηθεί σε πειράματα μικροσυστοιχιών (Yoon et al, 2012). Για την ανάλυση χρησιμοποιήσαμε τις τιμές έντασης που μετρήθηκαν σε δύο βιολογικές επαναλήψεις. Πιο αναλυτικά στα συγκεκριμένα πειράματα συγκρίνεται η γονιδιακή έκφραση κατά την εκθετική (επισημασμένη με Cy5) και στατική φάση (επισημασμένη με Cy3) του *E.coli*. Εδώ απεικονίζουμε την μέση τιμή έντασης των δύο πειραμάτων για το κανάλι Cy5. Οι κατηγορίες του ΜΠ και του ΠΚΦ-ΜΠ συγκρίνονται για τα υποσύνολα των ταυτοποιημένων (D: detected) και μη ταυτοποιημένων (N/D/Q not detected) πρωτεϊνών. Δείχνουμε ότι οι πρωτεΐνες που δεν καταφέραμε να ταυτοποιήσουμε μεταγράφονται σε χαμηλότερα επίπεδα.

---

**C.** Κατανομή μήκους των πρωτεϊνών: πρωτεΐνες που έχουν ανιχνευθεί σε παλαιότερες πρωτεομικές αναλύσεις (κόκκινο) είναι μεγαλύτερες (~349 αμινοξέα) σε σύγκριση με αυτές που ανήκουν σε μεταγραφικές μονάδες που εκφράζονται (γκρί ~279 αμινοξέα). Πρωτεΐνες που έχουν ανιχνευθεί σε μεταγραφικό επίπεδο ακολουθούν διτροπική καμπύλη (θέσεις ~410 και ~145).

**D.** Η κατανομή των ανιχνεύσιμων πεπτιδίων ανά πρωτεΐνη: οι πρωτεΐνες που ανήκουν στις παραπάνω κατηγορίες συγκρίνονται σχετικά με τον αριθμό των ανιχνεύσιμων πεπτιδίων που παράγουν. Τα πεπτίδια που παρουσιάζονται εδώ είναι μήκους από 5 έως 65 αμινοξέα και οι πρωτεΐνες με 1-2 ανιχνεύσιμα πεπτίδια εξαιρούνται από την ανάλυση. Οι πρωτεΐνες που έχουν ανιχνευθεί σε προηγούμενες πρωτεομικές αναλύσεις έχουν περισσότερα ανιχνεύσιμα πεπτίδια (~14, κόκκινο). Πρωτεΐνες που έχουν βρεθεί μόνο σε επίπεδο μεταγραφικό έχουν λιγότερα πεπτίδια (~7, μαύρο) και ενδέχεται αυτός να είναι ο λόγος που δεν καταφέραμε να τις ανιχνεύσουμε.

**E.** Μέσο μήκος πεπτιδίου σε σχέση με την υδροφοβικότητα του: οι πρωτεΐνες από τις τρεις κατηγορίες δεν παρουσιάζουν σημαντικές διαφορές ως προς τις ιδιότητες των πεπτιδίων τους.

---

## 5.7 Σχετική ποσοτικοποίηση πρωτεϊνών χωρίς χημική σήμανση με ισότοπα (Label-free)

Για να υπολογίσουμε τις ποσότητες των μεμβρανικών και περιφερικών πρωτεϊνών (PIM proteins) χρησιμοποιήσαμε δύο διαφορετικές μεθόδους υπολογισμού της σχετικής ποσότητας των πρωτεϊνών (χωρίς χημική επισήμανση πρωτεϊνών) emPAI και NSAF. Η πρώτη υπολογίζει το λόγο ανάμεσα στον αριθμό των πειραματικά ανιχνευμένων πεπτιδίων (detected) και στον αριθμό των θεωρητικά ανιχνεύσιμων πεπτιδίων (observable) (Ishihama et al, 2005). Η δεύτερη μέθοδος λαμβάνει υπόψη τον αριθμό των φασμάτων κανονικοποιημένο ως προς το μήκος της αντίστοιχης πρωτεΐνης και στη συνέχεια κανονικοποιεί όλες τις τιμές με το συνολικό τους άθροισμα (Zybailon et al, 2006)

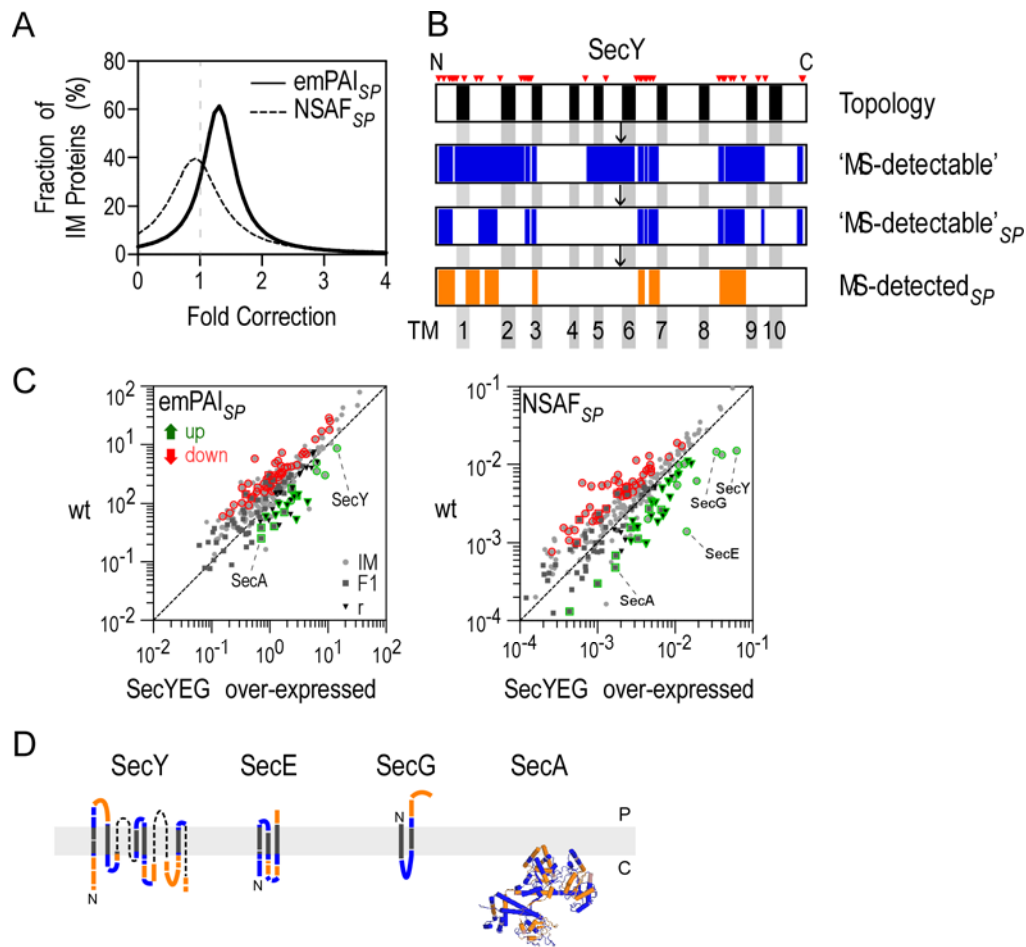
Διεξήχθη πρωτεόλυση στις επιφάνειες των αγρίου τύπου AMK (Papanastasiou et al, in preparation) ενώ πραγματοποιήθηκαν τέσσερις βιολογικές επαναλήψεις και τρεις ή τέσσερις τεχνικές επαναλήψεις σε κάθε περίπτωση για να εξασφαλιστεί η στατιστική εγκυρότητα. Υπολογίσαμε το βαθμό συσχέτισης του αριθμού φασμάτων κάθε πρωτεΐνης ανάμεσα στις βιολογικές επαναλήψεις (Spearman's correlation coefficients >0.873). Από τις πρωτεΐνες που ταυτοποιήθηκαν η πιο πολυπληθής κατηγορία ήταν οι μεμβρανικές πρωτεΐνες (282 συνολικά), μετά ακολουθούν οι περιφερικές (98 πρωτεΐνες). Επιλέξαμε ένα υποσύνολο από αυτές (221 μεμβρανικές, 50 ριβοσωμικές και 25 περιφερικές) με κριτήριο την συστηματική ανίχνευση τους

στις πειραματικές επαναλήψεις (δες ενότητα 6.3.5) και υπολογίσαμε τις ποσότητες τους κατά emPAI και NSAF.

Δεδομένου ότι η Τρυψίνη δεν εισχωρεί στα AMK αλλά πρωτεολύει μόνο τα τμήματα των πρωτεϊνών που βρίσκονται εκτεθειμένα στο διάλυμα, παράγονται μόνο τα αντίστοιχα πεπτίδια. Η εκτίμηση της σχετική ποσότητας των μεμβρανικών πρωτεϊνών με βάση τις δύο μεθόδους (emPAI, NSAF) αποδείχθηκε προβληματική λόγω μικρότερου αριθμού διαθέσιμων πεπτιδίων σε σχέση με τα θεωρητικά. Κατά συνέπεια τροποποιήσαμε τις εξισώσεις των δύο μεθόδων ως εξής: α) για το emPAI (στο εξής επονομαζόμενο ως emPAI<sub>SP</sub>) ο αριθμός των θεωρητικών πεπτιδίων αντικαταστάθηκε με τον αριθμό των ανιχνεύσιμων πεπτιδίων ('MS-detectable'<sub>SP</sub>; δες ενότητα 6.3.6) και β) για το NSAF (στο εξής επονομαζόμενο ως NSAF<sub>SP</sub>), το συνολικό μήκος κάθε πρωτεΐνης αντικαταστάθηκε με το συνολικό μήκος κάλυψης που προκύπτει από τα ανιχνεύσιμα πεπτίδια κάθε πρωτεΐνης (L<sub>SP</sub>; δες ενότητα 6.3.6).

Για να αποδείξουμε το βαθμό διόρθωσης ανάμεσα στις αρχικές και διορθωμένες τιμές των δύο μεθόδων ποσοτικοποίησης, υπολογίσαμε κατά πόσες φορές είναι μεγαλύτερες/μικρότερες οι διορθωμένες τιμές διαιρώντας τις καινούργιες τιμές με τις αρχικές (Εικόνα 5.8A). Οι κατανομές του ποσού διόρθωσης προσεγγίστηκαν με συναρτήσεις Lorentzian (coefficients of determination >0.96 για το emPAI και >0.98 για το NSAF). Στην περίπτωση του emPAI, για την πλειοψηφία των μεμβρανικών πρωτεϊνών (70%) οι τιμές διορθώθηκαν προς τα πάνω (1.23 φορές) ενώ στην περίπτωση των τιμών NSAF οι μισές περίπου πρωτεΐνες διορθώθηκαν προς τα κάτω (<1 φορές).

Για να δείξουμε πως λειτουργεί η διόρθωση στη πράξη επιλέξαμε σαν παράδειγμα τις πρωτεΐνες του εκκριτικού συστήματος Sec το οποίο αποτελείται από το μεμβρανικό κανάλι secYEG και την πρωτεΐνη κινητήρα του συστήματος, την SecA (Εικόνα 5.8D). Για το σύμπλοκο SecYEG που περιέχει ΔΠς η διόρθωση των ανιχνεύσιμων πεπτιδίων κυμαίνεται από 12-40% (μπλε Εικόνα 5.8D). Μεγαλύτερη λεπτομέρεια απεικονίζεται στην περίπτωση της SecY η οποία περιέχει 10 ΔΠς και τα πεπτίδια που αντιστοιχούν στις ΔΠς 3-4 και 7-8 δεν έχουν σημεία πρωτεόλυσης και κατά συνέπεια αποκλείονται από την ανίχνευση με ΦΜ. Εάν εφαρμόσουμε τα κριτήρια και επιλέξουμε τα ανιχνεύσιμα πεπτίδια της μεθόδου ('MS-detectable'<sub>SP</sub>) τότε το πραγματικά ανιχνεύσιμο μήκος της πρωτεΐνης αντιστοιχεί στο 27% του αρχικού (Εικόνα 5.8B). Οι αντίστοιχες διορθωμένες τιμές emPAI<sub>SP</sub> και NSAF<sub>SP</sub> είναι 1.8 και 1.9 φορές μεγαλύτερες από τις αρχικές.



**Εικόνα 5.8 – Ποσοτική ανάλυση των πρωτεϊνών του ΚΦ.**

**A.** Κατανομή του ποσοστού διόρθωσης των τιμών emPAI και NSAF με βάση τον ορισμό των ανιχνεύσιμων πεπτιδίων με βάση όρια ανίχνευσης της μεθόδου της πρωτεόλυσης επιφάνειας (δες ενότητα 6.3.6).

**B.** Τοπολογικός χάρτης των ΔΠ της πρωτεΐνης SecY (μαύρο). Με μπλε είναι τα θεωρητικά πεπτίδια που είναι ανιχνεύσιμα με βάση τα όρια του φασματογράφου (MS-detectable) και τα διορθωμένα με βάση όρια ανίχνευσης της μεθόδου της πρωτεόλυσης επιφάνειας (MS-detectable<sub>SP</sub>) (δες ενότητα 6.3.6) Στην δεύτερη περίπτωση η κάλυψη της πρωτεΐνης αντιστοιχεί στο 27%. Με πορτοκαλί απεικονίζονται τα πεπτίδια που ανιχνεύθηκαν στην πραγματικότητα με την μέθοδο της πρωτεόλυσης της επιφάνειας των AMK. Ο αύξων αριθμός των ΔΠ της SecY αναφέρεται στην τελευταία σειρά.

**C.** Σύγκριση της ποσότητας των πρωτεϊνών ανάμεσα στα αγρίου τύπου και secYEG κύτταρα. Οι ποσότητες έχουν υπολογιστεί με βάση το διορθωμένο emPAI (emPAI<sub>SP</sub>, αριστερά) και NSAF (NSAF<sub>SP</sub>, δεξιά) ενώ έχει πραγματοποιηθεί στατιστικό test για τον καθορισμό των πρωτεϊνών που η διαφορά στην ποσότητα ανάμεσα στις δύο συνθήκες (wt έναντι secYEG) είναι στατιστικά σημαντική (κόκκινο/πράσινο).

---

D. Τοπολογική απεικόνιση των πεπτιδίων της SecY, ενώ ακολουθείται αντίστοιχος χρωματικός κώδικά με το γράφημα B.

---

Τέλος απομονώσαμε AMK κυττάρων BL21-DE3 στα οποία υπερεκφράζονται (πλασμιδιακά) οι πρωτεΐνες του καναλιού SecYEG και τα αναλύσαμε με ΦΜ. Πραγματοποιήσαμε τρεις βιολογικές επαναλήψεις με 3 με 4 τεχνικές επαναλήψεις σε κάθε περίπτωση. Από τις 265 μεμβρανικές πρωτεΐνες που ανιχνεύθηκαν για τις 221 υπολογίσαμε τις αντίστοιχες ποσότητες ( $emPAI_{SP}$ ,  $NSAF_{SP}$ ).

Συγκρίναμε τις ποσότητες των πρωτεϊνών ανάμεσα στα κύτταρα αγρίου τύπου και SecYEG με στατιστικές αναλύσεις. Πραγματοποιώντας στατιστική ανάλυση με τους αριθμούς φασμάτων (SC: spectral count; δες ενότητα 6.3.10) βρήκαμε ότι 16 περιφερικές και 19 ριβοσωμικές πρωτεΐνες είναι στατιστικά σημαντικά διαφοροποιημένες ( $p$ -value<0.05, Εικόνα 5.8C). Εκ των οποίων όλες οι ριβοσωμικές και 9 περιφερικές βρέθηκαν να υπερεκφράζονται ενώ οι υπόλοιπες 7 περιφερικές βρέθηκαν σε μικρότερες ποσότητες στα SecYEG κύτταρα. Αντίστοιχη στατιστική ανάλυση με τις τιμές  $emPAI_{SP}$  (δες ενότητα 6.3.10) έδωσε τα ίδια αποτελέσματα όσον αφορά τις πρωτεΐνες του συστήματος Sec με εξαίρεση την SecE η οποία ανιχνεύεται με τα ίδια αξιόπιστα πεπτίδια και στις δύο συνθήκες (αγρίου τύπου και κύτταρα SecYEG) κατά συνέπεια η μέθοδος  $emPAI$  αποτυγχάνει να εντοπίσει κάποια μεταβολή στην ποσότητα. Τέλος η πρωτεΐνη SecA επίσης ανιχνεύεται σε μεγαλύτερες ποσότητες στα κύτταρα SecYEG πιθανά λόγω της υπερέκφρασης του υποδοχέα της, την SecY.

---

## 5.8 Συζήτηση

Στις ενότητες που προηγήθηκαν παρουσιάσαμε μια αναλυτική διαδικασία για τον προσδιορισμό των πρωτεϊνών του ΚΦ στο *Escherichia coli*. Η ανάλυση μας βασίζεται στην επισταμένη ταξινόμηση των πρωτεϊνών σε υποκυτταρικά διαμερίσματα (Orfanoudaki et al, 2014), στην ενδελεχή θεωρητική ανάλυση για τον καθορισμό ανιχνεύσιμων πεπτιδίων ('MS-detectable' peptides) και πως αυτά εφαρμόζονται στο ΠΚΦ, μια συνδυαστική πειραματική διαδικασία προετοιμασίας δείγματος και σε μια τροποποιημένη μέθοδο υπολογισμού της ποσότητας των πρωτεϊνών.

Ακολουθώντας ορθολογιστικά κριτήρια καταλήξαμε σε συμπεράσματα για τους περιορισμούς που υπεισέρχονται στην ανίχνευση των μεμβρανικών πρωτεϊνών του ΚΦ. Αυτά μπορούν να συνοψιστούν σε γενικούς κανόνες για την επίτευξη των τριών βασικών στόχων της πρωτεομικής (αναγνώριση, κάλυψη και ποσοτικοποίηση).

Επωφεληθήκαμε της πρόσφατης ταξινόμησης του πρωτεϊνώματος *E. coli* BL21-DE3 (Orfanoudaki et al, 2014) η οποία μας επέτρεψε να πραγματοποιήσουμε την πρώτη μέχρι σήμερα πλήρη σύγκριση των ιδιοτήτων των πρωτεϊνών ανά υποκυτταρικό διαμέρισμα καθώς επίσης και των αντίστοιχων πεπτιδίων που προκύπτουν από αυτές. Όσον αφορά το ισοηλεκτρικό σημείο οι μεμβρανικές πρωτεΐνες διαφοροποιούνται από τις υπόλοιπες κατηγορίες.

Σύμφωνα με την βιβλιογραφία οι κατανομές του ισοηλεκτρικού σημείου σχετίζονται με το μήκος των πρωτεϊνών και την ταξινόμηση των οργανισμών.(Kiraga et al, 2007) αλλά με τον υποκυτταρικό εντοπισμό (Kiraga et al, 2007; Schwartz et al, 2001) ενώ άλλοι το αμφισβητούν (Wu et al, 2006). Η μετατόπιση των μεμβρανικών πρωτεϊνών προς τις βασικές τιμές ισοηλεκτρικού φορτίου αποδίδεται στην παρουσία θετικά φορτισμένων αμινοξέων στις περιοχές κοντά στην επιφάνεια της μεμβράνης που α) σταθεροποιούν τις πρωτεΐνες στην μεμβράνη b) εξυπηρετούν την αλληλεπίδραση με διαλυτές πρωτεΐνες (Kiraga et al, 2007; Schwartz et al, 2001). Το χαρακτηριστικό αυτό μπορεί να αποδειχθεί χρήσιμο για την απομόνωση περισσότερων μεμβρανικών πρωτεϊνών από εκχυλίσματα ολόκληρων κυττάρων, π.χ. χρησιμοποιώντας την μέθοδο OFFGEL.

Σε μια πρόσφατη ανασκόπηση των διαφορετικών εφαρμογών της τεχνικής OFFGEL στην σύγχρονη πρωτεομική ανάλυση αναφέρουν ότι τα πλεονεκτήματα της μεθόδου για τον

---

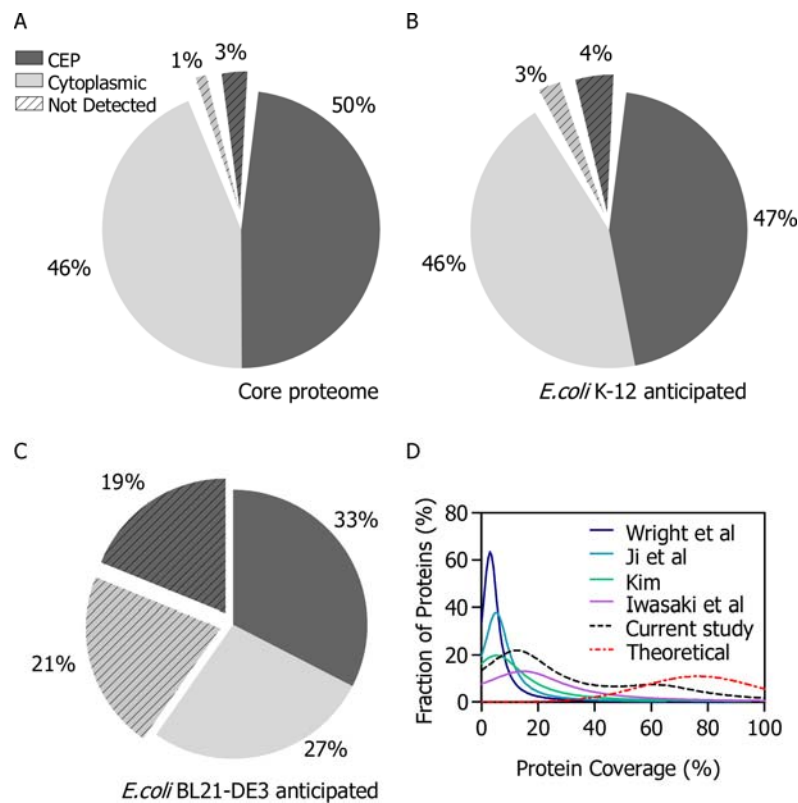
διαχωρισμό των μεμβρανικών πρωτεϊνών δεν έχει ακόμα εκτιμηθεί (Moreda-Piñeiro et al, 2014). Όταν πρωτεολυθούν οι πρωτεΐνες τα πεπτίδια που προκύπτουν δεν παρουσιάζουν αντίστοιχη μετατόπιση στις βασικές τιμές (Εικόνα 5.4D) πιθανώς επειδή τα κατάλοιπα που μπορούν να ιονιστούν είναι κατανεμημένα πιο ομοιόμορφα (Wu et al, 2006).

Ένα άλλο χαρακτηριστικό των μεμβρανικών πρωτεϊνών είναι ότι παράγουν λιγότερα πεπτίδια ανά πρωτεΐνη σε σχέση με τις άλλες κατηγορίες πρωτεϊνών (Εικόνα 5.4G), το οποίο αποδίδεται στην ύπαρξη των ΔΠ (μήκους ~20 αμινοξέων κατά μέσο όρο (Eichacker et al, 2004)). Οι περιοχές αυτές δεν περιέχουν πιθανά σημεία πρωτεόλυσης και κατά συνέπεια παράγουν μεγαλύτερα πεπτίδια με χαμηλή ικανότητα ιονισμού. Το γεγονός ότι το μήκος και η υδροφοβικότητα των πεπτιδίων αποτελεί καθοριστικό παράγοντα για την ταυτοποίηση των μεμβρανικών πρωτεϊνών έχει αναφερθεί σε παλαιότερη ανάλυση του φυτού *A. thaliana* (Eichacker et al, 2004) αλλά και του παθογόνου οργανισμού *S. aureus* (Wolff et al, 2008).

Ενδιαφέρον παρουσιάζει το γεγονός ότι συγκρίνοντας τις διαφορετικές μεθόδους δεν υπάρχει κάποια που να δίνει πλεονέκτημα στην ανίχνευση πρωτεϊνών που παράγονται σε μικρές ποσότητες από το κύτταρο. Ωστόσο τα πλεονεκτήματα κάθε μεθόδου αναδεικνύονται περισσότερο σημαντικά στις μικρές και λιγότερο άφθονες πρωτεΐνες. Η ταυτοποίηση μικρών πρωτεϊνών που παράγουν λιγότερα από 7 ανιχνεύσιμα πεπτίδια σε μικρή διαθεσιμότητα (από την άποψη μορίων ανά κύτταρο) φαίνεται να επιτυγχάνεται όταν εφαρμόζονται τεχνικές κλασματοποίησης. Αντίθετα η ο μικρός αριθμός ανιχνεύσιμων πεπτιδίων δεν φαίνεται να αποτελεί περιοριστικό παράγοντα για περισσότερο αφθονες πρωτεΐνες. Για παράδειγμα η μικρή υπομονάδα της ηλεκτρικής αφυδρογονάσης (DsbC) έχει μόνο τρία ανιχνεύσιμα πεπτίδια όμως υπολογίζεται στα περίπου 12.500 μόρια ανά κύτταρο και έχει ταυτοποιηθεί από όλες τις μεθόδους που εφαρμόσαμε. Συνεπώς οι τεχνικές κλασματοποίησης φαίνεται να είναι υποχρεωτικές για την εξερεύνηση των μικρών πρωτεϊνών του κυττάρου (54, 66, 67).

Είναι γενικά αποδεκτό ότι τα ποσοστά κάλυψης που επιτυγχάνονται στην περίπτωση των μεμβρανικών περιοχών είναι πολύ μικρά. Αυτό οφείλεται είτε στις μεθόδους προετοιμασίας των δειγμάτων, κάποια πεπτίδια απομονώνονται σε μικρές ποσότητες, είτε στα απορρυπαντικά και ιόντα που βρίσκονται σε μεγάλες ποσότητες κατά το στάδιο του ιονισμού τα οποία καταστέλλουν τα συγκεκριμένα πεπτίδια.





**Εικόνα 5.9 – Σύνοψη των πρωτεϊνών του ΚΦ που έχουν ταυτοποιηθεί μέχρι σήμερα από πρωτεομικές αναλύσεις**

Συνδυάζοντας τις πρωτεομικές αναλύσεις που έχουν πραγματοποιηθεί στα στελέχη *E. coli* K-12 και BL21-DE3 (Πίνακας 6.10) υπολογίσαμε πόσες πρωτεΐνες του κυτταροπλάσματος και του ΚΦ έχουν μέχρι σήμερα ταυτοποιηθεί.

- A.** Κοινό πρωτεϊνώμα ανάμεσα σε 43 στελέχη *E. coli* (Orfanoudaki et al, 2014).
- B.** Ανιχνεύσιμες πρωτεΐνες του *E. coli* K-12, που έχουν ταυτοποιηθεί μέχρι σήμερα από πρωτεομικές αναλύσεις εξειδικευμένες στο αντίστοιχο στέλεχος.
- C.** Ανιχνεύσιμες πρωτεΐνες του *E. coli* BL21-DE3, που έχουν μέχρι σήμερα ταυτοποιηθεί σε πρωτεομικές αναλύσεις στοχευόμενες στο αντίστοιχο στέλεχος.
- D.** Σύγκριση των πρωτεομικών αναλύσεων όσον αφορά το ποσοστό κάλυψης των πρωτεϊνών συμπεριλαμβανομένου και της παρούσας ανάλυσης.

Η *in-silico* ανάλυση του ΠΚΦ έδειξε ότι οι πρωτεΐνες διαφέρουν σε σχέση με την υδροφοβικότητα και το μέγεθος των αντίστοιχων πεπτιδίων τους (Εικόνα 5.4E) αλλά και σε σχέση με τον αριθμό των ανιχνεύσιμων πεπτιδίων τους (Εικόνα 5.4G). Όλα αυτά τα χαρακτηριστικά συνεισφέρουν στην μειωμένη κάλυψη των αλληλουχιών (Εικόνα 5.4H). Επιπλέον το εγγενώς

---

μικρό περιεχόμενο τους σε βασικά κατάλοιπα υπονομεύει περισσότερο την ανίχνευση τους. Στην παρούσα ανάλυση δεν βρέθηκε συσχέτιση ανάμεσα στη μέθοδο και τη τελική κάλυψη των μεμβρανικών πρωτεϊνών (Εικόνα 5.5D), αν και η ανίχνευση των πεπτιδίων με ΔΠ βελτιώθηκε σημαντικά με την μέθοδο OFFGEL (Εικόνα 5.5E και F).

Επιπλέον το μικρό ποσοστό κάλυψης των μεμβρανικών πρωτεϊνών οδήγησε σε εσφαλμένη ποσοτικοποίηση των πρωτεϊνών του ΚΦ από τις μέχρι σήμερα πρωτεομικές αναλύσεις γεγονός το οποίο καταδεικνύεται από το μικρό αριθμό δημοσιεύσεων που αναφέρουν τιμές για μεμβρανικές πρωτεΐνες. Η μέθοδος emPAI έχει χρησιμοποιηθεί σε πρωτεομικές αναλύσεις μεγάλης κλίμακας (Neilson et al, 2011) όπου έχει αναλυθεί το συνολικό πρωτεϊνωμα του K-12 (Krug et al, 2013) ή το διαλυτό πρωτεϊνωμα (Ishihama et al, 2008). Σε όλες τις περιπτώσεις οι μεμβρανικές πρωτεΐνες αντιμετωπίστηκαν με τον ίδιο τρόπο με τις διαλυτές ενώ η χαμηλή ανιχνευσιμότητα έχει αποδοθεί αποκλειστικά στο λόγο μάζας προς φορτίο και όχι στις μεθόδους προετοιμασίας του δείγματος ή στην περιεκτικότητα των πεπτιδίων σε ΔΠ.

Στη παρούσα ανάλυση χρησιμοποιήσαμε τις μεθόδους emPAI και NSAF για να υπολογίσουμε τις σχετικές ποσότητες των πρωτεϊνών και να εντοπίσουμε τις πρωτεΐνες που επηρεάζονται όταν υπερεκφράζεται το μεμβρανικό σύμπλοκο SecYEG, σε επεξεργασμένα AMK (Treated IMVs; Εικόνα 5.1). Μόνο ένα μικρό υποσύνολο πρωτεϊνών φαίνεται να αυξήθηκε υποδεικνύοντας ότι στην περίπτωση όπου το μεμβρανικό σύμπλοκο υπερεκφράζεται η αναγκαιότητα των υπόλοιπων βοηθητικών μονάδων του εκκριτικού συστήματος Sec υποβιβάζεται (π.χ. YidC, SecF, SecD κλπ.). Αντίστοιχα πολλές μεμβρανικές πρωτεΐνες βρέθηκαν σε σημαντικά μικρότερες ποσότητες σε σχέση με τα αγρίου τύπου κύτταρα, πιθανώς επειδή επισκιάζονται από τα πεπτίδια των SecYEG πρωτεϊνών.

Η πλειονότητα των ριβοσωμικών πρωτεϊνών ανιχνεύθηκαν σε μεγαλύτερες ποσότητες επιβεβαιώνοντας το σενάριο ότι βρίσκονται προσδεμένες στο σύμπλοκο SecYEG κατά την διαδικασία της έκκριση (Paranikou et al, 2007). Το ίδιο ισχύει και για την πρωτεΐνη SecA η οποία βρίσκεται επίσης προσδεμένη στο μεμβρανικό σύμπλοκο και αποτελεί το κινητήρα της εκκριτικής διαδικασίας (καταναλώνει ATP και παράγει μηχανικό έργο). Παρά την επεξεργασία των AMK με χαοτροπικούς και αποδιατακτικούς παράγοντες (όπου θα αναμέναμε ότι αφαιρούν κάποια μόρια) φαίνεται ότι η ποσότητα της SecA επηρεάζεται ελάχιστα (σύγκριση ανάμεσα στα επεξεργασμένα και μη επεξεργασμένα AMK).

---

Συνολικά ταυτοποιήσαμε το 57% του ανιχνεύσιμου πρωτεϊνώματος σε πλούσιο θρεπτικό μέσο (Πίνακας 5.1) ενώ αναφέρουμε 43 μεμβρανικές πρωτεΐνες που ανιχνεύθηκαν για πρώτη φορά με πρωτεομική ανάλυση. Για 503 μεμβρανικές πρωτεΐνες που ταυτοποιήσαμε, το UniProt δεν έχει καταχωρημένη την πληροφορία ότι η ύπαρξη τους έχει επιβεβαιωθεί σε επίπεδο πρωτεΐνης. Στο στέλεχος BL21-DE3 το 73% των μεμβρανικών πρωτεϊνών βρίσκεται στην κατηγορία 'inferred from homology' και 26% στην κατηγορία 'predicted'. Αυτό οφείλεται στο γεγονός ότι η αλληλούχιση του συγκεκριμένου στελέχους ολοκληρώθηκε πρόσφατα, και η κατηγοριοποίηση των πρωτεϊνών έχει προκύψει βάση ομολογίας με το στέλεχος *E.coli* K-12.

Συλλογικά από τις προηγούμενες πρωτεομικές αναλύσεις έχει ταυτοποιηθεί το 50% του ΚΦ που ανήκει στο πρωτεϊνώμα πυρήνα (Orfanoudaki et al, 2014) ενώ υπολείπεται το 3% (Εικόνα 5.9A). Αντίστοιχα, αν αναλύσουμε ξεχωριστά τις πρωτεϊκές αναλύσεις που βασίστηκαν στο πρωτεϊνώμα του *E.coli* K-12 και αυτές χρησιμοποίησαν το πρόσφατα διαθέσιμο πρωτεϊνώμα του BL21-DE3 παρατηρούμε ότι το έχει ταυτοποιηθεί το 47% του ΚΦ σε σχέση με το 33% στην περίπτωση του BL21-DE3 (Εικόνα 5.9B και C). Τέλος υπολογίσαμε το ποσοστό κάλυψης των μεμβρανικών πρωτεϊνών όπως προκύπτει δεδομένα τεσσάρων πρωτεομικών αναλύσεων (Πίνακας 6.10) και τα συγκρίναμε με τα αποτελέσματα της παρούσας ανάλυσης (Εικόνα 5.9D). Παρατηρούμε ότι θεωρητική κατανομή της αναμενόμενης κάλυψης των μεμβρανικών πρωτεϊνών κυμαίνεται από 60% έως 100%. ενώ οι πειραματικές κατανομές κυμαίνονται σε πολύ χαμηλότερα ποσοστά, μέσο ποσοστό κάλυψης που δεν υπερβαίνει το 20% (Masuda et al, 2009). Αντίθετα παρατηρούμε ότι ο συνδυασμός μεθόδων κλασματοποίησης και επεξεργασίας των AMK που ακολουθήσαμε στην παρούσα ανάλυση έχει οδηγήσει στην αύξηση της κάλυψης σημαντικού ποσοστού μεμβρανικών πρωτεϊνών (δεξιάς ώμος της μαύρης διακεκομμένης καμπύλης).

Συνοψίζοντας παρουσιάσαμε την πιο πλήρη μέχρι σήμερα θεωρητική ανάλυση των ιδιοτήτων των μεμβρανικών πρωτεϊνών με στόχο να μελετήσουμε και να συγκρίνουμε το εύρος ανίχνευσης διαφορετικών μεθόδων επεξεργασίας των δειγμάτων. Η ανάλυση αυτή οδήγησε σε σημαντικά συμπεράσματα για τις τεχνικές που αναμένεται να βελτιώσουν το ποσοστό κάλυψης των μεμβρανικών πρωτεϊνών.



---

## **ΚΕΦΑΛΑΙΟ 6 Υλικά και μέθοδοι**

### **6.1 Υποκυτταρική ταξινόμηση πρωτεϊνώματος *E.coli***

#### **6.1.1. Το πρωτεΐνωμα αναφοράς του *E.coli* K-12**

Αρχικές πηγές για την μελέτη της υποκυτταρικής ταξινόμησης του *E.coli* υπήρξαν οι βάσεις δεδομένων Uniprot (Dimmer et al, 2012) και EchoLOCATION (Horler et al, 2009) καθώς και το θεωρητικό πρωτεΐνωμα της ΠΜ (Bernsel et al, 2009). Ως πρωτεΐνωμα αναφοράς θεωρήσαμε το πρωτεΐνωμα του *E.coli* κ-12 MG1655 όπως ορίζεται στο Uniprot το οποίο και περιέχει 4303 πρωτεΐνες.

Η βάση δεδομένων EchoLOCATION περιέχει ένα πίνακα αντιστοίχισης γονιδίων σε υποκυτταρικές κατηγορίες. Η αντιστοίχιση των ονομάτων των γονιδίων στους αντίστοιχους αριθμούς πρόσβασης (accession numbers) δεν ήταν απλή καθώς τα ονόματα των γονιδίων δεν είναι μοναδικά. Σε πάνω από 100 περιπτώσεις πρωτεϊνών τα κύρια ονόματα γονιδίων ήταν κοινά με τα συνώνυμα άλλων.

Για την αντιστοίχιση των καταχωρημένων γονιδίων στο EchoLOCATION με αυτά στο Uniprot χρησιμοποιήσαμε το εργαλείο «αντιστοίχισης αριθμών πρόσβασης» στο Uniprot, όπου οι αριθμοί πρόσβασης EchoBASE αντιστοιχήθηκαν σε αριθμούς πρόσβασης Uniprot (Uniprot accessions). Σε περιπτώσεις όπου κάποιες καταχωρήσεις στο EchoLOCATION δεν συνοδεύονταν από αντίστοιχο αριθμό πρόσβασης EchoBASE αντιστοιχίσαμε τα ονόματα των γονιδίων με την βοήθεια MySQL ερωτημάτων. Για τις πρωτεΐνες με κοινά ονόματα γονιδίων, επιπλέον πληροφορίες διασταυρώθηκαν για την ταυτοποίηση τους.

#### **6.1.2. Βιοπληροφορικά εργαλεία και καθορισμός παραμέτρων.**

Η υπάρχουσα ταξινόμηση σε υποκυτταρικά διαμερίσματα στις τρεις πηγές δεδομένων συγκρίθηκαν μεταξύ τους και επανεξετάστηκαν σε σύγκριση με τις προβλέψεις διάφορων βιοπληροφορικών εργαλείων. Τα βασικά εργαλεία που χρησιμοποιήθηκαν ήταν: 1) SignalP4.0 (Petersen et al, 2011) και LipoP (Juncker et al, 2003) για την πρόβλεψη σηματοδοτικών πεπτιδίων που αναγνωρίζονται από τις πεπτιδάσες τύπου I και II αντίστοιχα, 2) TatP (Bendtsen et al, 2005b) το οποίο αναγνωρίζει μοτίβα έκκρισης του συστήματος Tat (twin arginine motifs), 3)

TMHMM v2.0 (Kall et al, 2004) και Phobius (Kall et al, 2007) τα οποία αναγνωρίζουν διαμεμβρανικές έλικες δηλαδή πιθανές μεμβρανικές πρωτεΐνες. Για όλες τις προβλέψεις που έγιναν με το Phobius το TMHMM, LipoP και το TatP χρησιμοποιήθηκαν οι τιμές των παραμέτρων ως είχαν. Στο SignalP επιλέξαμε ως ενδεχόμενο οι αλληλουχίες να περιέχουν διαμεμβρανικές έλικες. Οι λιποπρωτεΐνες της ΠΜ διαχωρίστηκαν από αυτές της ΕΜ βασιζόμενοι στο κριτήριο το αμινοξύ της θέσης +2 του ΩΤ (Ασπαραγίνη και Γλουταμίνη σημαίνει ότι παραμένει στην ΠΜ). Στο TatP χρησιμοποιήσαμε την προ-καθορισμένη κανονική έκφραση (Bendtsen et al, 2005b). Λόγω απόκλισης των προβλέψεων σε γνωστές πρωτεΐνες τύπου Tat για τον προσδιορισμό τους βασιστήκαμε αποκλειστικά σε βιοχημικά πειράματα (Tullman-Ercek et al, 2007)

Πιθανά δομικά στοιχεία καθορίστηκαν με χρήση εξειδικευμένων βιοπληροφορικών εργαλείων όπως το Amphipaseek για την πρόβλεψη αμφίφιλων α-ελικών. Σε άλλες περιπτώσεις βασιστήκαμε σε αμινοξική ομοιότητα με καλά χαρακτηρισμένες πρωτεΐνες.

Χρησιμοποιήσαμε βοηθητικά βιοπληροφορικά εργαλεία όπως το PSORT-B (Gardy et al, 2003), το *sosuiGramN* (Imai et al, 2008), το *LocTree3* (Goldberg et al, 2014) που βασίζεται σε μηχανές διανυσμάτων υποστήριξης και το *ClubSub-P*, το οποίο εκτελεί αναζητήσεις ομολογίας (Paramasivam et al, 2011). Στο *ClubSub-P* εφαρμόσαμε ως κατώτερο όριο ομολογίας της αλληλουχίας της άγνωστης πρωτεΐνης (query sequence) αλλά και της γνωστής (hit sequence) ίση με 70% ενώ ορίσαμε ως κατώτερο όριο του ποσοστού ταυτοποίησης (sequence identity threshold) ίσο με 40%.

### 6.1.3. Εκτίμηση αξιοπιστίας πειραματικών δεδομένων

Το επίπεδο αξιοπιστίας για κάθε πειραματική ένδειξη καθορίστηκε από την μεθοδολογία του αντίστοιχου πειράματος. Για παράδειγμα οι μέθοδοι κλασματοποίησης του κυττάρου σε συνδυασμό με ανίχνευση μέσω πρωτεομικής ανάλυσης είναι γνωστό ότι πάσχουν από θόρυβο από γειτονικά υποκυτταρικά διαμερίσματα (7). Εγείρονται επίσης αμφιβολίες όσον αφορά την αξιοπιστία των στατιστικών αναλύσεων που εφαρμόστηκαν σε κάθε μελέτη. Πιο σύγχρονες μελέτες τείνουν να χρησιμοποιούν πιο αυστηρά κριτήρια σε σχέση με την αξιόπιστη ταυτοποίηση πεπτιδίων/πρωτεϊνών ενώ χρησιμοποιούν και πιο σύγχρονους φασματογράφους.

Στην διαδικασία ταξινόμησης του *E.coli* σε υποκυτταρικά διαμερίσματα ακολουθήσαμε κάποιους κανόνες σε σχέση με τα πειραματικά αποδεικτικά στοιχεία που συγκεντρώθηκαν από

την βιβλιογραφική αναζήτηση. Η σειρά αξιοπιστίας των δεδομένων που εφαρμόσαμε ήταν: πειράματα μικροσκοπίας, βιοχημικά πειράματα και τέλος πρωτεομικές αναλύσεις (Πίνακας 6.1). Ως εκ τούτου όταν μια πρωτεΐνη έχει αναφερθεί ότι ταυτοποιήθηκε σε πρωτεομικές αναλύσεις έπειτα από κλασματοποίηση του κυττάρου αλλά έχει επίσης βρεθεί και με βιοχημικές μεθόδους, τότε στην δεύτερη περίπτωση τα αποδεικτικά στοιχεία θεωρήθηκαν μεγαλύτερης αξιοπιστίας.

**Πίνακας 6.1 - Ευρείας κλίμακας πρωτεομικές, γονιδιωματικές και βιοχημικές αναλύσεις στις οποίες βασίστηκε η υποκυτταρική ταξινόμηση του *E.coli*.**

Author	PMID	Referring Protein(s)
<b>Proteomic Studies</b>		
<b>Molloy MP et al</b>	[10806384]	MIPA SLP PA1 OMPX OMPW OMPC LAMB OMPA OMPF OMP PP TOLC OMPT BTUB YBHC PAL SLYB BAMC OMPW OMPX OMP T TSX OMPA FIU BAMA FHUA OMPC CIRA SLP AG43 PAL FEPA LAMB YBHC BAMC TOLC SLYB FADL FHUE OMPF CIRA FHUE BTUB FHUA FADL
<b>Gevaert K et al</b>	[12488465]	BAMC OMPC MPAA YGDI LOLB FIMD YRAP YFEY MPAA OMPX OMPF OMPA YEDD RLPA BAMD ODP1 YJEI
<b>Fountoulakis M et al</b>	[12624733]	BAMC PAL MLTA YNCD PA1 OMPA MIPA TSX OMPW OMP T OMP T BAMA BTUB AG43 FEPA CIRA NFRA FIMD FADL OMPF FECA TOLC LOLB FHUA LAMB FHUA OMPC
<b>Lopez-Campistrous A et al</b>	[15911532]	ARTI HYBA DACB DPPA OMPW ARTI CPDB DCRB TSX OMPX CPDB RBSB PAL RLPA USPA YIFL OSME YAJG PTNAB PTFAH BTUB TRPG LPOA RPOE FADL SLP LPTD YEDD LAMB OMPF TOLC RNE CIRA FECA ODP2 YBHC FHIA FABZ GLPD AK1H AG43 ECOT BAMC SLYB MIPA YNFB BAMB LOLB FHUA OMPC FIMD BAMD LPP PTNC YRAP YBJP OMP T
<b>Spelbrink RE et al</b>	[15919657]	ODO2 ODP1 FRDA PHSM YIBN ATPA ODO1 YHCB IMDH PBP2 CYDA CYOA ATPB ATPD RL17 PTM3C SECD PTTBC DLD PQIB ODP2 LEPA OPGH PTGCB PTNAB PTW3C YDIJ PLSB RL10 RL9 MGLA FTSH NUOG NUOCD SECG MSCS SECA GLPD TOLA SPEA SRMB RL15 PBPA LLDD NUOF CYDD RODZ YIDC PBPB
<b>Stenberg F et al</b>	[16079137]	MALK PTNC
<b>Weiner JH et al</b>	[17904518]	CPXP
<b>Metzger LE 4th et al</b>	[19883124]	LPXB
<b>Thein M et al</b>	[20932056]	GFCE YBHC NLPD YAJG OSME ECNB LPTE YJEI MLTC RLPA FADL FECA CIRA YGGG BAMD YEAY BLC BAMC SLYB MIPA OMPA PAL OMPW OMPX PA1 MLTA BAME BAMA LPTD SLP YIAD FEPA LAMB PHOE OMPF TOLC YFAZ MLAA YFEY CUSC YAIW OMPN BAMB FLGH YGER BSMA NLPE MLTB LPOA LPOB LOLB YDCL YRAP RCSF FIU YBJP YNCD
<b>Ohniwa RL et al</b>	[21541338]	RL3 RS6 RL30 RL14 HNS RL24 DPS YGAU RL21 ATPB ATPA ATPG ATPD FDOH OMPX OMPA SLYB BAMC BLC OMP T YBJP DHS A NUOI NUOCD DBHA DBHB STPA FIS HFQ IHFA IHFB TNAA EFTS CH60 YDGA DHSB LPP RL4 RL13 RPOA RS4 RS12 RL9 RL34 RL33 RL31 RL27 RL1 RL19 RL11 OSME ATPF YGDI LOLB OMPC SLP FLIC OMPF DAMX CYOA NUOA YAJC ECNB HFLK

<b>Papanastasiou M et al</b>	[23230279]	PHOL CFA ACEA OMPR CYSB SLYD DEAD ARCA ACCD GATD PTHP DAPD G3P1 RHLB SYS SYN SYT RPOC RPOB YJGA RIMJ FABB ALKH ALF1 THIO FABF ASPA FUMA DPS DEOD CISY ACCA PUR8 KBL ALF RHLE GLRX3 GLRX4 GNSA STPA HNS DBHB DBHA HINT NIFU GLMU ERPA HEMG DNAB CYSH YHBJ NADE MUKB MUKE CARB PPSA PPIB HEMH GSA ARGE ADD GABT ASNB SYA SYM FADB SYGA SYGB RNR SRMB KPYK2 RFAC ACCC LACI TREC ALR2 GUAA RLUF GCSP TKT2 ACON2 WECG TKT1 OPDA RADA FRMA GABD RFAP RFAQ LRP PLSX GCST FUMC CRP G
<b>Genomic studies</b>		
<b>Nenninger AA et al</b>	[21645131]	CSGD
<b>Ishihama A</b>	[23138451]	YKGD PAAX YCIT DHAR YCGE YCFQ RUTE YCAN YBJK MATA HYFR TYRR RCSB RCSA MCBR YDEO YNEL YAFY YCJZ YAGI YGAV MURR YPHH YEEY KDGR DMLR YEAM CREB CLCB CHPS LSRR YEFM BAER YGJM DICA DICC FRLR TTDR YGJI BIRA BGLJ YJIR YJIM YJIE SGCR YJHI IDNR YJGJ DSDC HEXR YDCQ AAER YHAJ NEMR SDIA GADE GADW MALT YGBI YFJR RSTA CSGD QSEB YGFI ADA PHOB SOXS ARAC FUR SLYA NANR FADR METJ FLHD ICIA HDFR NRDR TRPR LEXA NIKR FNR NSR MNTR RPOE ARCD OMPR CSP A ULAR YIAG PUUR ARCA MODE CUER NHAR_
<b>Biochemical studies</b>		
<b>Gonnet P et al</b>	[15174130]	MLTB YFGH YGDR YQHH YNFC RCSF LPP YBFN YBFP YBJP AMID YHFL YHDV YCAL YBHC HSLJ LOLB YDDW YNBE YDCL YFIB YOAF YRAP GFCB PGAB FLGH BLC BAMC SLYB PAL GFCE MLTA YEAY YCEB YCEK TCDA YGER YCJN YECR WZA MLAA YFEY CUSC BORD YAFT YBAY BAMB LPOB NLPE NLPI ACRA YEDD CYOA CSGG APBE METQ MLTD ACRE NLPC YJEI YFIL RLPA EMTA MLTC SPR OSMB YIIG BAMD ECNB ECNA SLP MDTE OSME YIAD YEHR SPRT LPTE NLPA YJBF FTSL NLPD
<b>Tullman-Ercek D et al</b>	[17218314]	AMIC TORZ OPGD CUEO AMIA NAPA TORC TORA FDHD FDOG EFEB FTSP FDNG DMSA YFHG MBHS MBHT WCAM YAHJ YDHX YNFE YAGS YAGT NRFC YEDY FHUD NAPG YAGM YCBK YNFF
<b>Microscopy study</b>		
<b>Watt RM et al</b>	[17272300]	RSEP PYRH METK NUSA PPNK MSRB YGEH YFJP KAD ENGB MNME YQIK

#### 6.1.4. Σύγκριση ανάμεσα στα στελέχη K-12 και BL21-DE3 του *E.coli*

Οι ομόλογες πρωτεΐνες ανάμεσα στα δύο στελέχη *E.coli* (κοινό πρωτεΐνωμα) προσδιορίστηκαν με το εργαλείο BLAST (Camacho et al, 2009) για την αναζήτηση ομοιότητας και πολλαπλής στοίχισης ακολουθιών. Κάθε πρωτεΐνωμα αντιστοιχήθηκε με το άλλο, κάθε πρωτεΐνη του *E.coli* K-12 αναζητήθηκε στο BL21 και το αντίστροφο. Οι πρωτεΐνες που βρέθηκαν με ποσοστό ταυτοποίησης (identity) μικρότερο του 40% (δηλ. το ποσοστό των κατάλοιπων που ταχτοποιήθηκε ανάμεσα στις δύο ακολουθίες διαιρεμένο με το συνολικό μήκος αυτών των πρωτεϊνών) ή με  $e\text{-value} \geq 10^{-3}$  απορρίφθηκαν.

Στην συνέχεια ορίσαμε το πρωτεΐνωμα πυρήνα ανάμεσα σε όλα τα στελέχη *E.coli* συγκρίνοντας 43 στελέχη *E.coli* (Πίνακας 6.2). Για κάθε πρωτεΐνη στο *E.coli* K-12 αναζητήθηκε η



ποιο ομόλογη σε καθένα από τα 43 στελέχη *E.coli*. Στη συνέχεια επιλέχθηκαν μόνο πρωτεΐνες με ποσοστό ταυτοποίησης (identity) μεγαλύτερο του 40% και e-value μικρότερο από  $10^{-3}$ . Ορίσαμε ως πρωτεΐνωμα πυρήνα μόνο όσες πρωτεΐνες με τα παραπάνω κριτήρια υπήρχαν κοινές σε όλα τα στελέχη.

Και στις δύο αναλύσεις χρησιμοποιήσαμε τον αλγόριθμο “best hit” που εμπεριέχεται στην ρουτίνα blastp με κατώφλι ίσο με 0.2. Το πρωτεΐνωμα πυρήνα αποτελείται από 2583 proteins (Πίνακας 2.1).

**Πίνακας 6.2 - Κατάλογος με τα 43 στελέχη *E.coli* με βάση τα οποία προσδιορίστηκε το πρωτεΐνωμα πυρήνα**

Organism (Uniprot taxon id)	Taxon ID	Size of the complete Proteome	Strain Name
<a href="#">83333</a>	83333	4305	Escherichia coli (strain K12)
<a href="#">83334</a>	83334	6485	Escherichia coli O157:H7
<a href="#">199310</a>	199310	5335	Escherichia coli O6:H1 (strain CFT073 / ATCC 700928 / UPEC)
<a href="#">216592</a>	216592	4884	Escherichia coli O44:H18 (strain 042 / EAEC)
<a href="#">316385</a>	316385	3990	Escherichia coli (strain K12 / DH10B)
<a href="#">316401</a>	316401	4778	Escherichia coli O78:H11 (strain H10407 / ETEC)
<a href="#">331111</a>	331111	4915	Escherichia coli O139:H28 (strain E24377A / ETEC)
<a href="#">331112</a>	331112	4308	Escherichia coli O9:H4 (strain HS)
<a href="#">362663</a>	362663	4605	Escherichia coli O6:K15:H31 (strain 536 / UPEC)
<a href="#">364106</a>	364106	5192	Escherichia coli (strain UTI89 / UPEC)
<a href="#">405955</a>	405955	4870	Escherichia coli O1:K1 / APEC
<a href="#">409438</a>	409438	4979	Escherichia coli (strain SE11)
<a href="#">413997</a>	413997	4142	Escherichia coli (strain B / REL606)
<a href="#">431946</a>	431946	4461	Escherichia coli O150:H5 (strain SE15)
<a href="#">439855</a>	439855	4876	Escherichia coli (strain SMS-3-5 / SECEC)
<a href="#">444450</a>	444450	5279	Escherichia coli O157:H7 (strain EC4115 / EHEC)
<a href="#">469008</a>	469008	4842	Escherichia coli (strain B / BL21-DE3)
<a href="#">481805</a>	481805	4173	Escherichia coli (strain ATCC 8739 / DSM 1576 / Crooks)
<a href="#">536056</a>	536056	4583	Escherichia coli (strain ATCC 33849 / DSM 4235 / NCIB 12045 / K12 / DH1)
<a href="#">544404</a>	544404	5255	Escherichia coli O157:H7 (strain TW14359 / EHEC)
<a href="#">566546</a>	566546	5063	Escherichia coli (strain ATCC 9637 / CCM 2024 / DSM 1116 / NCIMB 8666 / NRRL B-766 / W)
<a href="#">573235</a>	573235	5326	Escherichia coli O26:H11 (strain 11368 / EHEC)
<a href="#">574521</a>	574521	4594	Escherichia coli O127:H6 (strain E2348/69 / EPEC)
<a href="#">585034</a>	585034	4330	Escherichia coli O8 (strain IAI1)
<a href="#">585035</a>	585035	4805	Escherichia coli O45:K1 (strain S88 / ExPEC)
<a href="#">585055</a>	585055	4797	Escherichia coli (strain 55989 / EAEC)
<a href="#">585056</a>	585056	4975	Escherichia coli O17:K52:H18 (strain UMN026 / ExPEC)

<a href="#">585057</a>	585057	4554	Escherichia coli O7:K1 (strain IAI39 / ExPEC)
<a href="#">585395</a>	585395	4964	Escherichia coli O103:H2 (strain 12009 / EHEC)
<a href="#">585396</a>	585396	5238	Escherichia coli O111:H- (strain 11128 / EHEC)
<a href="#">585397</a>	585397	4940	Escherichia coli O81 (strain ED1a)
<a href="#">595495</a>	595495	4637	Escherichia coli (strain ATCC 55124 / KO11)
<a href="#">595496</a>	595496	4042	Escherichia coli (strain K12 / MC4100 / BW2952)
<a href="#">655817</a>	655817	4786	Escherichia coli OR:K5:H- (strain ABU 83972)
<a href="#">685038</a>	685038	4575	Escherichia coli O83:H1 (strain NRG 857C / AIEC)
<a href="#">701177</a>	701177	5085	Escherichia coli O55:H7 (strain CB9615 / EPEC)
<a href="#">714962</a>	714962	4748	Escherichia coli O18:K1:H7 (strain IHE3034 / ExPEC)
<a href="#">869729</a>	869729	4779	Escherichia coli (strain UM146)
<a href="#">885275</a>	885275	4908	Escherichia coli (strain 'clone D i14')
<a href="#">885276</a>	885276	4908	Escherichia coli (strain 'clone D i2')
<a href="#">1133852</a>	1133852	5105	Escherichia coli O104:H4 (strain 2011C-3493)
<a href="#">1133853</a>	1133853	5044	Escherichia coli O104:H4 (strain 2009EL-2071)
<a href="#">1134782</a>	1134782	5123	Escherichia coli O104:H4 (strain 2009EL-2050)

### 6.1.5. Ανάπτυξη της βάσης δεδομένων STEPdb

Τα περιεχόμενα της βάσης δεδομένων STEPdb οργανώθηκαν και είναι διαθέσιμα μέσω ενός εξυπηρετητή MySQL. Η ιστοσελίδα του STEPdb υλοποιήθηκε με το πρόγραμμα PHPMaker ένα σύστημα διαχείρισης περιεχομένου τύπου PHP και Javascript και είναι διαθέσιμη μέσω εξυπηρετητή Apache. Δομικές αλλαγές του περιεχομένου της ιστοσελίδας υλοποιήθηκαν με την βιβλιοθήκη jQuery. Τα γραφήματα της πιθανότητας αταξίας (Disorder probability graphs) έγιναν με αντικειμενοστρεφή βιβλιοθήκη JpGraph (<http://jpgraph.net/>). Οι κατανομές των υποκυτταρικών κατηγοριών υλοποιήθηκαν χρησιμοποιώντας “Highcharts” (<http://www.highcharts.com/>). Τα σχέδια των πρωτεϊνικών συμπλόκων έγιναν με την βιβλιοθήκη γραφικών (graphics drawing library) της PHP.

## 6.2 Μοντέλα διαχωρισμού εκκρινόμενων από κυττροπλασματικές πρωτεΐνες

### 6.2.1. Πρόβλεψη σημείου εκκίνησης ώριμου τμήματος και καθορισμός του ελάχιστου δυνατού μήκους προς ανάλυση

Προσδιορίσαμε το σημείο έναρξης του ΩΤ χρησιμοποιώντας βιοπληροφορικά εργαλεία για την πρόβλεψη του σημείου αποκοπής των σηματοδοτικών πεπτιδίων. Για τους δύο τύπους εκκρινόμενων πρωτεϊνών επωφεληθήκαμε δύο εξειδικευμένων βιοπληροφορικών εργαλείων: α) το SignalP 4.1 (Petersen et al, 2011) για την πρόβλεψη τύπου I σηματοδοτικών πεπτιδίων

(υποκατηγορίες περιπλασμικών πρωτεϊνών και πρωτεϊνών της εξωτερικής μεμβράνης και β) το Lipop (Juncker et al, 2003) πρόβλεψη τύπου II σηματοδοτικών πεπτιδίων (λιποπρωτεΐνες).

Πίνακας 6.3 – Κατάλογος πρωτεϊνών με τα μικρότερα σε μήκος ώριμα τμήματα

Όνομα πρωτεΐνης (UniProt)	Κωδικός πρωτεΐνης (UniProt)	Υποκυτταρική Ταξινόμηση (STEPdb)	Μήκος ΣΠ	Μήκος ΩΤ	Συνολικό Μήκος
YJEI_ECOLI	P0AF70	I	18	99	117
YOBA_ECOLI	P0AA57	G	26	98	124
YBAV_ECOLI	P0AAR8	G	25	98	123
YCGJ_ECOLI	P76001	G	24	98	122
BAME_ECOLI	P0A937	I	19	94	113
YFIM_ECOLI	P46126	E	14	93	107
YEBY_ECOLI	P64506	G	20	93	113
OSME_ECOLI	P0ADB1	I	20	92	112
MLIC_ECOLI	P28224	I	17	92	109
YBFN_ECOLI	P75734	I	16	92	108
YECR_ECOLI	P76308	I	15	92	107
YACC_ECOLI	P0AA95	G	24	91	115
YOHN_ECOLI	P64534	G	21	91	112
YMGD_ECOLI	P0AB46	G	19	90	109
YIDQ_ECOLI	P0ADM4	E	20	90	110
HDEA_ECOLI	P0AES9	G	21	89	110
CSGC_ECOLI	P52107	G	21	89	110
YEGR_ECOLI	P76406	I	17	88	105
CUSF_ECOLI	P77214	G	22	88	110
YPEC_ECOLI	P64542	G	21	87	108
YDBL_ECOLI	P76076	G	21	87	108
YJDP_ECOLI	Q6BEX5	G	22	87	109
YKGJ_ECOLI	P0AAL9	G	24	85	109
PSIF_ECOLI	P0AFM4	G	21	85	106
PSPE_ECOLI	P23857	G	19	85	104
BSMA_ECOLI	P39297	I	24	85	109
YNFB_ECOLI	P76170	F3	28	85	113
PTFB1_ECOLI	P69808	G	24	84	108
YSAB_ECOLI	Q2M7M3	I	17	82	99
ASR_ECOLI	P36560	G	21	81	102
YMDA_ECOLI	P75917	G	22	81	103
BORD_ECOLI	P77330	I	16	81	97
YNFD_ECOLI	P76172	G	21	80	101
HDEB_ECOLI	P0AET2	G	29	79	108
YICS_ECOLI	Q2M7X4	G	21	76	97
YAAX_ECOLI	P75616	G	23	75	98
YDDL_ECOLI	P77519	G	21	75	96
YPDI_ECOLI	O32528	I	18	73	91
YBJH_ECOLI	P0AAY4	G	22	72	94
YDBJ_ECOLI	P0ACW2	I	16	72	88
YNJH_ECOLI	P76227	G	18	72	90

YJFY_ECOLI	P0AF86	G	20	71	91
YEHE_ECOLI	P33344	G	22	71	93
YJFN_ECOLI	P0AF82	G	21	70	91
YAH0_ECOLI	P75694	G	21	70	91
YOAF_ECOLI	P64493	I	16	68	84
YQHH_ECOLI	P65298	I	19	66	85
YHCN_ECOLI	P64614	G	22	65	87
YBIJ_ECOLI	P0AAX3	G	22	64	86
MCBA_ECOLI	P0AAX6	G	22	64	86
YJBT_ECOLI	A5A628	G	29	63	92
BHSA_ECOLI	P0AB40	F3	22	63	85
RZOQ_ECOLI	C1P601	E	22	62	84
YCEK_ECOLI	P0AB31	I	15	60	75
YJBE_ECOLI	P0AF45	G	21	59	80
LPP_ECOLI	P69776	I, F4	20	58	78
YHDV_ECOLI	P64622	I	16	57	73
YGDI_ECOLI	P65292	I	19	56	75
YKGI_ECOLI	P75687	G	22	56	78
YNCJ_ECOLI	P64459	G	22	54	76
YGDR_ECOLI	P65294	I	19	53	72
MARB_ECOLI	P31121	G	21	51	72
OSMB_ECOLI	P0ADA7	I	23	49	72
YIFL_ECOLI	P0ADN6	I	19	48	67
YNBE_ECOLI	P64448	I	16	45	61
RZOR_ECOLI	P58042	I	19	42	61
RZOD_ECOLI	P58041	I	19	41	60
YDCA_ECOLI	P0ACW4	G	20	37	57
YHFL_ECOLI	P64627	I	19	36	55
HOKC_ECOLI	P0ACG4	G	21	29	50
HOKD_ECOLI	P0ACG6	G	22	29	51
ECNB_ECOLI	P0ADB7	I	21	27	48
ECNA_ECOLI	P0ADB4	I	18	23	41

### 6.2.2. Αναπαράσταση των αμινοξικών αλληλουχιών

Στα περισσότερα μοντέλα που εκπαιδεύσαμε με το GEMS συμπεριλάβαμε ως παράμετρο την πληροφορία της πρωτοταγούς αμινοξικής αλληλουχίας. Για την διαδικασία της εκπαίδευσης οι αλληλουχίες μετατράπηκαν από πληροφορία κειμένου σε αριθμητική αναπαράσταση. Κάθε αμινοξύ κωδικοποιήθηκε με ένα δυαδικό αριθμό 20 ψηφίων (Πίνακας 6.4).

Σε ορισμένα μοντέλα δοκιμάσαμε να ομαδοποιήσουμε τα αμινοξέα με βάση τις φυσικοχημικές τους ιδιότητες. Προέκυψαν δύο εναλλακτικές ομαδοποιήσεις μια σε 11 και δεύτερη σε 9 ομάδες (Πίνακας 6.5 και Πίνακας 6.6)

**Πίνακας 6.4 – Απλή αναπαράσταση των αμινοξέων**

Αντιστοίχιση ενός δυαδικού αριθμού σε κάθε ένα από τα 20 αμινοξέα από τα οποία απαρτίζονται οι πρωτεΐνες. Τα αμινοξέα ταξινομήθηκαν με βάση τις φυσικοχημικές τους ιδιότητες και στη συνέχεια αντιστοιχήθηκαν σε δυαδικούς αριθμούς των 20 ψηφίων. Κάθε δυαδική αναπαράσταση έχει μοναδικό μοναδιαίο ψηφίο.

Name	One letter Code	Numeric Representation	Binary Coding
<b>Hydrophobic/Non Polar</b>			
Methionine	M	(1)	100000000000000000
Alanine	A	(2)	010000000000000000
Valine	V	(3)	001000000000000000
Leucine	L	(4)	000100000000000000
Isoleucine	I	(5)	000010000000000000
Proline	P	(6)	000001000000000000
Phenylalanine	F	(7)	000000100000000000
Tryptophan	W	(8)	000000010000000000
<b>Hydrophobic/Polar</b>			
Glycine	G	(9)	000000001000000000
Serine	S	(10)	000000000100000000
Cysteine	C	(11)	000000000010000000
Asparagine	N	(12)	000000000001000000
Glutamine	Q	(13)	000000000000100000
Tyrosine	Y	(14)	000000000000010000
Threonine	T	(15)	000000000000001000
<b>Polar/positive/Basic</b>			
Lysine	K	(16)	0000000000000001000
Arginine	R	(17)	0000000000000000100
Histidine	H	(18)	0000000000000000010
<b>Polar/negative/acidic</b>			

Aspartic Acid	D	(19)	00000000000000000010
Glutamic Acid	E	(20)	00000000000000000001

**Πίνακας 6.5 – Συμπαγής αναπαράσταση των αμινοξέων**

Ο πίνακας συνοψίζει εννέα διαφορετικές ομάδες αμινοξέων με βάση τις φυσικοχημικές τους ιδιότητες και την απεικόνισή τους σε δυαδική αριθμούς.

Property	One letter Code	Amino Acids	Numeric Representation	Binary Coding
Acidic	@	D,E	(1)	100000000
Basic	+	K,R	(2)	010000000
Small	s	A,G	(3)	001000000
Polar uncharged	o	T,S	(4)	000100000
Hydrophobic	h	I,L,V,M	(5)	000010000
Bulky	b	F,Y,W	(6)	000001000
Proline	p	P	(7)	000000100
Polar	q	N,Q,H	(8)	000000010
Cysteine	C	C	(9)	000000001

**Πίνακας 6.6 – Χαλαρή αναπαράσταση των αμινοξέων**

Ο πίνακας συνοψίζει έντεκα διαφορετικές ομάδες αμινοξέων με βάση τις φυσικοχημικές τους ιδιότητες και την απεικόνισή τους σε δυαδική αριθμούς.

Property	One letter Code	Amino Acids	Numeric Representation	Binary Coding
Hydrophobic	Ph	L,I,F	(1)	1000000000
Small	Sm	V,G,A,P	(2)	0100000000
R	R	R	(3)	0010000000
K	K	K	(4)	0001000000
D	D	D	(5)	0000100000
E	E	E	(6)	0000010000
Hydroxyl	X	Y,T,S	(7)	0000001000
Polar	pol	N,Q,C	(8)	0000000100
H	H	H	(9)	0000000010
M	M	M	(10)	0000000001
W	W	W	(11)	0000000000

### 6.2.3. Συναρτήσεις πυρήνα (kernels)

Δύο εναλλακτικές για συναρτήσεις πυρήνα είναι διαθέσιμες στο GEMS, πολυωνυμικός και Gaussian (radial basis function kernels). Δοκιμάσαμε να παράγουμε μοντέλα και με τις δύο μεθόδους. Τα μοντέλα της δεύτερης περίπτωσης είχαν ελάχιστα υψηλότερη απόδοση (1%-2% AUC) σε σχέση πρώτη. Όμως επειδή ο στόχος μας ήταν η καλύτερη κατανόηση των χαρακτηριστικών που επιλέγονται επιλέξαμε της πολυωνυμικές συναρτήσεις πυρήνα. Σε κάθε περίπτωση το GEMS ρυθμίστηκε ώστε να επιλέγει την τάξη του πολυωνύμου με κριτήριο το AUC. Να σημειωθεί ότι στα περισσότερα μοντέλα που εκπαιδεύτηκαν ήταν γραμμικά (πολυώνυμο πρώτης τάξης), υποδηλώνοντας γραμμική ανεξαρτησία ανάμεσα στις διαφορετικές θέσεις στην αλληλουχία.

---

#### 6.2.4. Εκτίμηση της απόδοσης των μοντέλων

Το GEMS ενσωματώνει επίσης τεχνικές για την εκτίμηση της απόδοσης από τα δεδομένα εκπαίδευσης (cross-validation). Δύο τεχνικές: LOOCV (Leave one out cross validation) και CV (N-fold cross validation) υλοποιούνται. Δοκιμάσαμε και τις δύο τεχνικές ενώ απορρίψαμε την τεχνική LOOCV επειδή σημαντικά πιο αργή. Κατά συνέπεια υιοθετήσαμε μια 10-fold cross validation σε όλα τα μοντέλα που εκπαιδεύσαμε. Ως μέτρο απόδοσης χρησιμοποιήσαμε το AUC (Area Under the Curve).

#### 6.2.5. Επιλογή Χαρακτηριστικών

Το GEMS ενσωματώνει στην διαδικασία εκπαίδευσης αλγόριθμους επιλογής χαρακτηριστικών: HITON-PC (Causal discovery method, outputs parents and children), HITON-MB (Causal discovery method, outputs Markov blanket) (Aliferis et al, 2003), S2N-OVR (Signal to Noise Ratio One versus rest), S2N-OVO (Signal to Noise Ratio One versus one). Όταν επιλέγονται πολλοί αλγόριθμοι ταυτόχρονα το GEMS συγκρίνει τα μοντέλα που προκύπτουν με τα διαφορετικά υποσύνολα χαρακτηριστικών και επιλέγει το βέλτιστο.

Στην παρούσα ανάλυση χρησιμοποιήσαμε και συγκρίναμε τους αλγόριθμους HITON-PC και S2N-OVR. Παρατηρήσαμε ότι τα υποσύνολα χαρακτηριστικών που επιλέγονται από τον HITON-PC είναι πολύ μικρότερα σε σχέση με αυτά του αλγόριθμου S2N-OVR, με αυτό να αντανακλάται στην πώση της απόδοσης των μοντέλων (της τάξης του 0.01%). Κατά συνέπεια αποφασίσαμε να υιοθετήσουμε την χρήση του αλγόριθμου HITON-PC για την εκπαίδευση όλων των μοντέλων (τα συγκεκριμένα δεδομένα δεν παρουσιάζονται)

#### 6.2.6. Σύνολα εκπαίδευσης και αξιολόγησης (test and train sets)

Σε κάθε κλάση χωρίσαμε τα δεδομένα με τυχαία επιλογή σε δύο σύνολα, το σύνολο εκπαίδευσης (train set) και το σύνολο αξιολόγησης (test set) με αναλογία 20%/80%. Την ίδια στρατηγική ακολουθήσαμε και για τα μοντέλα που εκπαιδεύσαμε για κάθε υποκατηγορία των εκκρινόμενων πρωτεϊνών (π.χ. περιπλασμικές).



Πίνακας 6.7 – Σύνοψη των συνόλου εκπαίδευσης και αξιολόγησης

Κλάση	Σύνολο εκπαίδευσης	Σύνολο αξιολόγησης	Σύνολο
Λιποπρωτεΐνες	94	24	118
Περιπλασμικές	229	57	286
Πρωτεΐνες της EM	51	13	64
Εκκρινόμενες	374	94	468
Κυτταροπλασματικές	1519	380	1899

### 6.2.7. Βάρη των επιλεγμένων χαρακτηριστικών

Στις περιπτώσεις όπου προέκυψαν γραμμικά μοντέλα διαχωρισμού στο GEMS, επαικταδεύσαμε τα μοντέλα στην MATLAB 2009b με βάση τα επιλεγμένα χαρακτηριστικά και υπολογίσαμε τα βάρη τους στην γραμμική συνάρτηση απόφασης. Για γραμμική συνάρτηση διαχωρισμού τα το διάνυσμα των συντελεστών βάρους υπολογίζεται ως:

$$\vec{w} = \sum_i^m \alpha_i \vec{x}_i$$

όπου  $x_1, x_2, \dots, x_m \in X$  είναι τα διανύσματα υποστήριξης και  $a_i > 0$  είναι οι αντίστοιχοι πολλαπλασιαστές Lagrange (δες Παράρτημα A5).

Χρησιμοποιώντας τα βάρη των επιλεγμένων χαρακτηριστικών αναπαραστήσαμε κάθε μοντέλο με ένα σχεδιάγραμμα τύπου logo (π.χ. **Εικόνα 3.4**). Στην αναπαράσταση αυτή κάθε αμινοξύ αναπαριστάται με το αντίστοιχο γράμμα στην κωδικοποίηση ενός γράμματος των αμινοξέων (one letter code). Τα χαρακτηριστικά που επιλέγονται στην ίδια θέση στην αλληλουχία αναπαριστώνται ως μια στοίβα από γράμματα.

Τα χαρακτηριστικά με θετικά βάρη απεικονίζονται πάνω από το άξονα  $x=0$  ενώ τα αρνητικά κάτω από αυτόν. Το συνολικό άθροισμα των θετικών και αρνητικών βαρών υπολογίστηκε ανά θέση:

$$s_i^- = \sum_{k=1}^n w_{ik} \quad w_{ik} < 0$$

$$s_i^+ = \sum_{k=1}^p w_{ik} \quad w_{ik} > 0$$

όπου  $w_{ik}$  το βάρος του χαρακτηριστικού  $k$  στη θέση  $i$  της αλληλουχίας,  $n$  και  $p$  ο αριθμός των αρνητικών και θετικών χαρακτηριστικών στη θέση  $i$  της αλληλουχίας αντίστοιχα.

Στη συνέχεια τα άθροισμα αυτά κανονικοποιήθηκαν από 0 έως 1 για τα θετικά και από 0 έως -1 για τα αρνητικά αθροίσματα.

$$S_i^- = \frac{\sum_{k=1}^n w_{ik}}{\min_{\text{for all } i} \sum_{k=1}^n w_{ik}} \quad w_{ik} < 0$$

$$S_i^+ = \frac{\sum_{k=1}^p w_{ik}}{\max_{\text{for all } i} \sum_{k=1}^p w_{ik}} \quad w_{ik} > 0$$

όπου  $S_i^-$  και  $S_i^+$  τα κανονικοποιημένα αθροίσματα των θετικών και αρνητικών χαρακτηριστικών στην θέση  $i$  της αλληλουχίας.

Το συνολικό ύψος κάθε στοίβας από γράμματα είναι ίσο με το κανονικοποιημένο άθροισμα στην συγκεκριμένη θέση ενώ το ύψος κάθε γράμματος μέσα στην ίδια στοίβα υπολογίστηκε αναλογικά με το βάρος του αντίστοιχου χαρακτηριστικού:

$$H_{ik} = \frac{w_{ik}}{\sum_{k=1}^n w_{ik}} \cdot |S_i|$$

όπου  $H_{ik}$  το συνολικό ύψος του γράμματος που αναπαριστά το χαρακτηριστικό  $k$  στη θέση  $i$  της αλληλουχίας

### 6.2.8. Ψεύδο-αμινοξική σύσταση (Pseudo amino-acid Composition)

Η έννοια της ψεύδο-αμινοξική σύστασης προτάθηκε ως τρόπος αναπαράστασης ο οποίος δεν χάνει εντελώς την πληροφορία της σειράς των αμινοξέων (Chou, 2005; Yu et al, 2010). Είναι στην ουσία μια επέκταση της αναπαράστασης με 20 τιμές που εκφράζουν την συχνότητα κάθε αμινοξέος (amino acid frequency). Διαφορετικές παραλλαγές της ψεύδο-αμινοξική σύστασης

έχουν προταθεί έκτοτε. Στην παρούσα ανάλυση πραγματοποιήσαμε την δική μας υλοποίηση η οποία βασίζεται σε δύο διαφορετικές κλίμακες υδροφοβικότητας (Kyte-Doolittle και Engelman).

Δεδομένου μιας πρωτεϊνικής αλληλουχίας P η οποία απαρτίζεται από L αμινοξέα  $P=[R1 R2 R3 \dots RL]$  η ψευδο-αμινοξική σύσταση ορίζεται ως:

$$P=[p_1, p_2, \dots, p_{20}, p_{20+1}, p_{20+2}, \dots, p_{20+\lambda}], \quad \lambda < L \text{ όπου}$$

$$P_u = \begin{cases} \frac{f_u}{\sum_{i=1}^{20} f_i + w \sum_{k=1}^{\lambda} \tau_k} & (1 \leq u \leq 20) \\ \frac{w \tau_{u-20}}{\sum_{i=1}^{20} f_i + w \sum_{k=1}^{\lambda} \tau_k} & (21 \leq u \leq 20 + \lambda) \end{cases}$$

$$\tau_k = \frac{1}{L - k} \sum_{i=1}^{L-k} \Phi_{i,i+k} \quad (k < L) \quad (2)$$

Όπου  $f_u$  είναι οι συχνότητες των 20 αμινοξέων στην πρωτεΐνη P,  $\tau_k$  ο  $k$ -στος παράγοντας συσχέτισης αλληλουχίας (sequence-correlation factor) όπως υπολογίζεται στην εξίσωση (2) και  $w$  ένας συντελεστής βάρους.

Στην δική μας υλοποίηση χρησιμοποιήσαμε τις κλίμακες υδροφοβικότητας Kyte-Doolittle και Engelman. Κατά συνέπεια με την ψευδο-αμινοξική σύσταση αναπαράστηκαμε έμμεσα την έννοια του γινομένου της υδροφοβικότητας δύο γειτονικών αμινοξέων. Η απόσταση των γειτονικών θέσεων καθορίζεται από τον αριθμό  $\lambda$ . Ορίσαμε την τιμή του βάρους  $w$  στην εξίσωση ίση με 0.5 ενώ το  $\lambda$  ίσο με 60, υπολογίσαμε δηλαδή 120 παράγοντες  $\tau_k$  οι τιμές των οποίων εισήχθησαν στην διαδικασία της εκπαίδευσης.

$$\tau_1 = \frac{1}{L-1} \sum_{i=1}^{L-1} \Phi_{i,i+1}^{Ky}$$

$$\tau_2 = \frac{1}{L-2} \sum_{i=1}^{L-2} \Phi_{i,i+2}^{Ky}$$

.

.

$$\tau_\lambda = \frac{1}{L-1} \sum_{i=1}^{L-1} \Phi_{i,i+1}^{En}$$

$$\tau_{\lambda+1} = \frac{1}{L-2} \sum_{i=1}^{L-2} \Phi_{i,i+2}^{En}$$

όπου

$\Phi_{i,i+j}^{Ky} = H_i H_{i+j}$  H: Kyte Doolittle Hydrophobicity

$\Phi_{i,i+j}^{En} = h_i h_{i+j}$  h: Engelman Hydrophobicity

### 6.2.9. Παράγωγα εκκρινόμενων πρωτεϊνών, εκτίμηση πειραματικών δεδομένων

Η αναπαράσταση του ποσού της εκκρινόμενης πρωτεΐνης υπήρξε διαφορετική ανάμεσα στις μελέτες. Στην περίπτωση που το πόσο εκκρινόμενο υλικό αναπαριστάται με ραβδογράμματα ή ως εικόνες από πηκτώματα πολυακρυλαμιδίου που ποσοτικοποιήθηκαν και κανονικοποιήθηκαν επί το ποσό της αγρίου τύπου πρωτεΐνης, τότε τα ποσοστά που αναφέρουν οι συγγραφείς χρησιμοποιήθηκαν ως αυτούσια. Σε περιπτώσεις όπου υπήρχαν διαθέσιμα πηκτώματα πολυακρυλαμιδίου χωρίς αντίστοιχη ποσοτικοποίηση τότε εμείς πραγματοποιήσαμε αντίστοιχη μέτρηση.

**Πίνακας 6.8 – Παράγωγα εκκρινόμενων πρωτεϊνών**

Πειραματικά δεδομένα που συλλέχθηκαν από την βιβλιογραφία για διαφορετικά παράγωγα εκκρινόμενων πρωτεϊνών. Αφορούν συνολικά 120 έκδοχα εκκρινόμενων πρωτεϊνών τα οποία έχουν προκύψει από μεταλλάξεις στην περιοχή του σηματοδοτικού πεπτιδίου ή/και στο ΩΤ των αγρίου τύπου πρωτεϊνών και συγκεκριμένα 37 μεταλλάξεις στο ΣΠ, και 73 με μεταλλάξεις στο ΩΤ και 10 με μεταλλάξεις και στις δύο περιοχές. Η απόδοση της έκκρισης έχει υπολογιστεί ποσοστιαία ως προς την ποσότητα της εκκρινόμενης πρωτεΐνης αγρίου τύπου. Τα ποσοστά κυμαίνονται από 0 έως και πάνω από το 100% (δηλαδή εκκρίνονται περισσότερο από την αγρίου τύπου πρωτεΐνη).

Παράγωγο	Πηγή	PMID	Περιοχή Μετάλλαξης	Περιγραφή	Πρωτεΐνη	Έκκριση (% επί της αγρίου τύπου)*
malE Delta12-18			SP		MalE	0
malE delta12-18 R1			SP		MalE	0
malE Delta12-18 R2			SP		MalE	0
malE Delta12-18 R3	<a href="#">Bankaitis V et al., Cell. [1984]</a>	6327054	SP	MalE with truncated hydrophobic core where efficiency is restored by mutations on the mature domain.	MalE	0
malE Delta12-18 R4			SP		MalE	0
malE Delta12-18 R5			SP		MalE	0
malE+			B		MalE	100
malE Delta12-18-R6			B		MalE	95
K(2,3)			M		PhoA	11
K(5,6)			M		PhoA	46
K(13,14)			M		PhoA	89
K(19,20)			M	Study of the negative charge bias of the early mature domain. Translocation of PhoA is affected by the insertion of Lysine and arginines in the early mature domain.	PhoA	97
K(29,30)	<a href="#">Kajava AV et al., J Bacteriol. [2000]</a>	PMC111264	M		PhoA	92
R(5,6)			M		PhoA	21
R(13,14)			M		PhoA	26
R(19,20)			M		PhoA	53
R(29,30)			M	PhoA	20	
L(-5)	<a href="#">Kajava AV et al., J Biol Chem. [2002]</a>	12393890	SP	Study of hydrophobic bulky residues in position -5 of the signal peptide of PhoA	PhoA	23
P(-6)L(-5)			SP		PhoA	70
L(-5)P(-4)			SP		PhoA	80
I(-5)			SP		PhoA	9
P(-6)I(-5)			SP		PhoA	103
I(-5)P(-4)			SP	PhoA	78	

Y(-5)			SP		PhoA	4
P(-6)Y(-5)			SP		PhoA	99
Y(-5)P(-4)			SP		PhoA	23
L(-5)S(-4)			SP		PhoA	48
L(-5)V(-3)			SP		PhoA	41
L(-5)S(-2)			SP		PhoA	40
P(-6/ -5/ -4)			SP		PhoA	68
1 Ala			M		PhoA	100
1 Phe			M		PhoA	100
1 Gly			M		PhoA	100
1 Cys			M		PhoA	100
1 Lys			M		PhoA	100
1 His	<a href="#">Karamyshev AL et al., J Mol Biol. [1998]</a>	9545377	M	Single amino acid substitutions at positions from -5 to +1 of the signal peptide of PhoA	PhoA	100
1 Glu			M		PhoA	100
1 Pro			M		PhoA	0
1 Tyr			M		PhoA	100
1 Leu			M		PhoA	100
1 Gln			M		PhoA	100
1 Ser			M		PhoA	100
WT-K5L5	<a href="#">Kim J et al., J Bacteriol. [2000]</a>	10869093	M	Mutations within the signal peptide or/and the early mature domain of alkaline phosphatase. Study of the interplay among these characteristics (efficiency and SecB requirement)	PhoA	17
WT-K5L5-mat			B		PhoA	17
WT-K3N2L5			M		PhoA	40
WT-K2N3L5			M		PhoA	48
WT-K1N4L5			M		PhoA	65
WT-N5L5			M		PhoA	77
3L7A-K5L5			B		PhoA	15
4L6A-K5L5			B		PhoA	15
5L5A-K5L5			B		PhoA	17
6L4A-K5L5			B		PhoA	30
7L3A-K5L5			B	PhoA	90	
9L1A-K5L5			B	PhoA	90	
5L5A			SP		PhoA	63
7L3A			SP		PhoA	105

9L1A			SP		PhoA	105
hMBP-AP			SP		PhoA	100
hOA-AP			SP		PhoA	100
hPC-AP			SP		PhoA	0
h'PC-AP	<a href="#">Laforet et al., J Biol Chem. [1989]</a>	2668291	SP	Replacement of PhoA signal peptide hydrophobic core with cores from MalE or M13 major coat protein	PhoA	0
hCPC-AP			SP		PhoA	30
hPC-AP K-F			SP		PhoA	0
hPC-AP K-Q			SP		PhoA	0
hcmPC-AP			B		PhoA	40
2AB			M		PhoA	5
X2			M		PhoA	99
X3	<a href="#">Li P et al., Proc Natl Acad Sci U S A. [1988]</a>	3051001	M	Mutations on the early mature domain of PhoA to alter the net charge.	PhoA	95
X4			M		PhoA	22
X6			M		PhoA	41
CO(0)			M		PhoA	100
C2(+2)			M		PhoA	70
C3(+3)			M		PhoA	20
C4(-3)			M	PhoA	100	
C5(+4)			M	PhoA	0	
C6(+6)	<a href="#">MacIntyre S et al., Mol Gen Genet. [1990]</a>	2199818	M	Mutations on the early mature domain of PhoA to alter the net charge.	PhoA	0
C7(-6)			M		PhoA	50
C8(-1)			M		PhoA	40
C9(-1)			M		PhoA	100
C10(+1)			M		PhoA	85
C11(-4)			M		PhoA	60
C12(-4)			M		PhoA	80
A(-20)	<a href="#">Nesmeyanova et al., FEBS Lett. [1997]</a>	9042967	SP	Positively charged residues in the N-terminus of PhoA signal peptide and their significance in the secretion efficiency	PhoA	68
F(-20)			SP		PhoA	69
G(-20)			SP		PhoA	69
H(-20)			SP		PhoA	98
P(-20)			SP		PhoA	65
C(-20)			SP		PhoA	63
Y(-20)			SP		PhoA	88

E(-20)			SP		PhoA	75
Δ23			M		LamB	0
Delta39			M		LamB	0
Delta55			M		LamB	0
Delta51			M		LamB	0
Delta54			M		LamB	0
Delta61	<a href="#">Rasmussen BA and Silhavy TJ [1987]</a>	20616663	M	Series of in-frame deletions that remove varying lengths of early lamB sequenceearly mature domain of maltoporin (LamB)	LamB	62
Delta36			M		LamB	66
Delta56			M		LamB	70
Delta49			M		LamB	85
Delta64			M		LamB	96
Delta2			M		LamB	88
BOT			M		AmpC	0
B1T			M		AmpC	0
B2T			M		AmpC	0
B3T			M		AmpC	3
B4T			M		AmpC	5
B5T			M		AmpC	10
B8T			M		AmpC	25
B11T			M		AmpC	50
B12T	<a href="#">Summers RG et al., J Biol Chem. [1989], 20074-81</a>	2511198	M	Fusion of AmpC (beta-lactamase) signal peptide with chicken muscle triosephosphate isomerase.	AmpC	100
D226N			M	Substitutions of positively charges in early mature achive secretion.	AmpC	0
Delta22-69A			M		AmpC	0
R3S			M		AmpC	100
R3P			M		AmpC	100
R3P,D226N			M		AmpC	100
K4Q			M		AmpC	0
K4Q,D226A			M		AmpC	0
R3P,K4R			M		AmpC	0
R3K			M		AmpC	0
R3K,D226N			M		AmpC	0



---

### 6.2.10. Σύγκριση με άλλα βιοπληροφορικά εργαλεία

Συγκρίναμε την απόδοση των μοντέλων που εκπαιδεύτηκαν με το GEMS με τρία παρόμοια βιοπληροφορικά εργαλεία: SignalP 4.1 (Petersen et al, 2011), Phobius (Kall et al, 2007) και LipoP (Juncker et al, 2003). Για την πρόβλεψη με SignalP 4.0 χρησιμοποιήσαμε το web-interface και από τα αποτελέσματα χρησιμοποιήσαμε τις τιμές πιθανότητας Smax για τον υπολογισμό του αντίστοιχου AUC. Η τιμή Smax για κάθε αλληλουχία αποτελεί την μέγιστη τιμή του S-score σε όλη την αλληλουχία ενώ το S-score είναι η πιθανότητα ένα αμινοξύ να ανήκει στο πεπτιδίο σήμα. Στην περίπτωση του Phobius εγκαταστήσαμε τοπικά το εργαλείο και χρησιμοποιήσαμε το υπάρχον αρχείο που ορίζει τις παραμέτρους για την πρόβλεψη (model file). Στη συνέχεια υπολογίσαμε την μέγιστη πιθανότητα για ύπαρξη σηματοδοτικού πεπτιδίου (cleavage site probability) για κάθε αλληλουχία και υπολογίσαμε το AUC με βάση αυτές τις τιμές. Τέλος για τις προβλέψεις με το LipoP χρησιμοποιήσαμε το web-interface και τις αντίστοιχες τιμές πιθανότητας (scores) για το υπολογισμό του AUC.

### 6.2.11. Εκτίμηση των υδρόφοβων περιοχών

ΣΠ προσδένονται σε διαφορετικές θέσεις από ότι τα ώριμα τμήματα (Gouridis et al, 2009)

Για τον προσδιορισμό των υδρόφοβων περιοχών των ώριμων τμημάτων, των σηματοδοτικών πεπτιδίων αλλά και των κυτταροπλασματικών πρωτεϊνών βασιστήκαμε στην κλίμακα υδροφοβικότητας των Kyte και Doolittle (Kyte et al, 1982). Στην κλίμακα αυτή τα αμινοξέα με υδρόφοβο χαρακτήρα έχουν θετικές τιμές ενώ υδρόφιλα αρνητικές.

Χρησιμοποιώντας μέγεθος παραθύρου ίσο με 9 και μοντέλο μεταβολής συντελεστών βάρους (Normal weight variation model) που αντιστοιχεί στην κανονική κατανομή, υπολογίσαμε το προφίλ υδροφοβικότητας κάθε πρωτεΐνης. Πιο συγκεκριμένα το προφίλ υδροφοβικότητας αντιστοιχεί στην καμπύλη που προκύπτει αν σε κάθε θέση της αλληλουχίας υπολογίσουμε το μέσο όρο της γειτονικής υδροφοβικότητας. Ο αριθμός των γειτονικών αμινοξέων που συνυπολογίζονται στην υδροφοβικότητα μιας θέσης καθορίζεται από το μέγεθος του παράθυρου.

Τα προφίλ υδροφοβικότητας κανονικοποιήθηκαν με βάση την μέγιστη συνολική υδροφοβικότητα που καθορίζεται από το μέγεθος παραθύρου που επιλέξαμε και την κλίμακα υδροφοβικότητας. Η τιμή ως προς την οποία κανονικοποιήσαμε το προφίλ προκύπτει από την

---

μέγιστη υδροφοβικότητα της κλίμακας που αντιστοιχεί στο κατάλοιπο της Ισολευκίνης (4.5), το μέγεθος παράθυρου και το μοντέλο μεταβολής συντελεστών βάρους. Στη συνέχεια προσδιορίσαμε τις θέσεις αλλά και το μήκος υδρόφοβων περιοχών (καμπύλη πάνω από το μηδέν) υπολογίζοντας την αρχή και το τέλος των θετικών τιμών στην καμπύλη.

Στην περίπτωση των μεμβρανικών πρωτεϊνών της ΠΜ (inner membrane proteins) υπολογίσαμε την υδροφοβικότητα το μήκος αλλά και την θέση της πρώτης διαμεμβρανικής περιοχής (ΔΠ). Για την πρόβλεψη των ΔΠ χρησιμοποιήσαμε το βιοπληροφορικό εργαλείο Phobius (Kall et al, 2007).

#### 6.2.12. Υπολογισμός προδιάθεσης σχηματισμού συσσωματώσεων

Η προδιάθεση σχηματισμού συσσωματώσεων (aggregation propensity) είναι τάση μιας αλληλουχίας να αλληλεπιδράσει τυχαία με άλλες αλληλουχίες σχηματίζοντας δεσμούς που οδηγούν σε μεγαλο-μοριακούς σχηματισμούς ή αλλιώς συσσωματώματα.

Χρησιμοποιήσαμε το βιοπληροφορικό εργαλείο AGGRESKAN (Conchillo-Sole et al, 2007) το οποίο υπολογίζει την πιθανότητα σχηματισμού συσσωματώσεων κατά μήκος μιας αλληλουχίας και στην συνέχεια εντοπίζει περιοχές με αυξημένη πιθανότητα. Τέλος υπολογίζει τιμές όπως ο αριθμό τέτοιων περιοχών ανά 100 αμινοξέα και την συνολική πιθανότητα κάθε αλληλουχίας. Για την ανάλυση μας χρησιμοποιήσαμε την ιστοσελίδα του AGGRESKAN. Η ανάλυση έγινε για τις κατηγορίες των προ-μορφών και ώριμων τμημάτων των εκκρινόμενων πρωτεϊνών αλλά και για τις κυτταροπλασματικές πρωτεΐνες.

#### 6.2.13. Υπολογισμός μέσου προφίλ αταξίας

Χρησιμοποιήσαμε το βιοπληροφορικό εργαλείο IUPred (Dosztanyi et al, 2005a) για να προβλέψουμε το προφίλ πιθανότητας αταξίας για κάθε πρωτεΐνη. Το εργαλείο εγκαταστάθηκε τοπικά. Τα προφίλ αταξίας υπολογίστηκαν με τον αλγόριθμο που συνυπολογίζει μια περιοχή 2-100 γειτονικών αμινοξέων (επιλογή *-long*). Τέλος δημιουργήσαμε το μέσο προφίλ αταξίακάθε υποκατηγορίας πρωτεϊνών (εκκρινόμενες, κυτταροπλασματικές και ώριμα τμήματα) υπολογίζοντας σε κάθε θέση τις αλληλουχίας, το μέσο όρο πιθανότητας αταξίας όλων των πρωτεϊνών.

---

#### 6.2.14. Εκπαίδευση μοντέλων με την πληροφορία της ενέργειας αλληλεπίδρασης των αμινοξέων

Αναπτύξαμε μοντέλα διαχωρισμού των ώριμων τμημάτων από κυτταροπλασματικές πρωτεΐνες χρησιμοποιώντας ως μεταβλητές εκπαίδευσης την συνολική ενέργεια αλληλεπίδρασης ανά ιδιοδιάνυσμα του πίνακα πιθανότητας ενέργειας  $P$  (Εικόνα 3.5). Συγκεκριμένα για κάθε πρωτεΐνη υπολογίσαμε 20 τιμές συνολικής ενέργειας αλληλεπίδρασης για κάθε ένα από τα αμινοξέα. Η συνολική ενέργεια αλληλεπίδρασης που οφείλεται στα αμινοξέα τύπου  $i$  που αλληλεπιδρούν με τα υπόλοιπα αμινοξέα.

#### 6.2.15. *In silico* συνδυασμοί σηματοδοτικών πεπτιδίων με αντίστοιχα ώριμα τμήματα

Τα ΣΠ 468 εκκρινόμενων πρωτεϊνών ανακατεύτηκαν με τα αντίστοιχα ώριμα τμήματα και για τις αλληλουχίες που προέκυψαν προβλέφθηκε η κατηγορία (κυτταροπλασματική ή εκκρινόμενη) χρησιμοποιώντας το βασικό μοντέλο πρόδρομης μορφής (Πίνακας 3.2). Αντίστοιχα οι σηματοδοτικές αλληλουχίες συνδυάστηκαν με 1899 κυτταροπλασματικές πρωτεΐνες (η Μεθειονίνη στη θέση +1 παραλείφθηκε) και έγινε πρόβλεψη έκκρισης ή όχι κάθε συνδυασμού. Επειδή οι συνδυασμοί που προκύπτουν είναι υπολογιστικά δαπανηρό να επεξεργαστούν όλοι επιλέξαμε τυχαία το 1.5 % των εκκρινόμενων συνδυασμών και το 3% των κυτταροπλασματικών

#### 6.2.16. Πρόβλεψη εκκρινόμενων πρωτεϊνών σε άλλα βακτήρια

Συλλέξαμε τα πρωτεϊνώματα για 10 Gram<sup>+</sup> και 25 Gram<sup>-</sup> βακτήρια (Πίνακας 6.9) από το Uniprot (Dimmer et al, 2012) και συνδυάσαμε τρία BE για την πρόβλεψη σηματοδοτικών πεπτιδίων: SignalP 4.1 (Petersen et al, 2011) το οποίο προβλέπει ΣΠ τύπου I, το LipoP (Juncker et al, 2003) για την πρόβλεψη σηματοδοτικών πεπτιδίων τύπου II και το PRED-TAT (Bagos et al, 2010) το οποίο μπορεί να προβλέψει ΣΠ τύπου Tat ή Sec (τύπου I και II). Για τον προσδιορισμό των πιθανών εκκρινόμενων πρωτεϊνών τύπου Sec εφαρμόσαμε τα παρακάτω κριτήρια:

Αρχικά εξαιρέσαμε όλες πρωτεΐνες προτείνονται από το PRED-TAT ότι περιέχουν ΣΠ τύπου Tat, ανεξάρτητα από τη πρόβλεψη των υπόλοιπων BE. Στη συνέχεια ορίσαμε τις πιθανές λιποπρωτεΐνες ως αυτές που προβλέπονται από το LipoP ως τύπου II εκκρινόμενες πρωτεΐνες («SPII»). Τέλος ορίσαμε ως τύπου I εκκρινόμενες πρωτεΐνες όλες προβλέπονται από το SignalP

ότι περιέχουν ΣΜ, το LipoP ως «SPI» και το PRED-TAT ως «Sec», με την προϋπόθεση ότι και τα τρία BE συμφωνούν για το σημείο αποκοπής του ΣΠ.

**Πίνακας 6.9 – Κατάλογος Gram<sup>+</sup> και Gram<sup>-</sup> βακτηρίων και σύνοψη του αριθμού των πιθανών εκκρινόμενων πρωτεϊνών τύπου Sec**

Για 10 Gram<sup>+</sup> και 25 Gram<sup>-</sup> βακτήρια έγινε εκτίμηση του αριθμού των εκκρινόμενων πρωτεϊνών του συστήματος Sec Συνυπολογίζοντας τις προβλέψεις από τρία BE (SignalP 4.1, LipoP, PRED-TAT).

\* Η ανάλυση αφορά τον προσδιορισμό Sec εκκρινόμενων πρωτεϊνών

A/A	Στέλεχος	Ταξινόμηση Gram	Συνολικός Αριθμός Πρωτεϊνών	Αριθμός Αναγνώρισης στελέχους (UniProt)	Αριθμός πιθανών Εκκρινόμενων Πρωτεϊνών*
<b>Gamma</b>					
1	Salmonella bongori N268-08	-	4751	1197719	376
2	Yersinia pestis bv. Antiqua (strain Antiqua)	-	4136	360102	331
3	Citrobacter freundii UCI 31	-	4932	1400136	472
4	Klebsiella pneumoniae (strain 342)	-	5739	507522	518
5	Pseudomonas fluorescens	-	7426	294	783
6	Acinetobacter baumannii (strain ACICU)	-	3746	405416	359
7	Coxiella burnetii (strain RSA 331 / Henzerling II)	-	1892	360115	90
8	Legionella pneumophila (strain Philadelphia 1 / ATCC 33152 / DSM 7513)	-	2950	272624	225
9	Haemophilus influenzae	-	3568	727	349
<b>Beta</b>					
10	Neisseria gonorrhoeae	-	6015	485	561
11	Bordetella pertussis (strain CS)	-	3275	1017264	286
12	Ralstonia pickettii (strain 12J)	-	4891	402626	490
<b>Alpha</b>					
13	Bartonella quintana JK 19	-	1308	1134507	27
14	Brucella melitensis biotype 2 (strain ATCC 23457)	-	3125	546272	216
15	Candidatus Liberibacter asiaticus str. Ishi-1	-	1068	931202	31
<b>Proteobacteria</b>					
16	Helicobacter pylori (strain HPAG1)	-	1542	357544	107
17	Campylobacter lari (strain RM2100 / D67 / ATCC BAA-1060)	-	1545	306263	96
18	Campylobacter coli 2548	-	1809	887315	95
<b>Chlamydiae</b>					
19	Chlamydia trachomatis (strain D/UW-3/Cx)	-	897	272561	46
20	Chlamydia pneumoniae	-	4081	83558	205
<b>Bacteroidetes</b>					

21	<i>Bacteroides fragilis</i> (strain YCH46)	-	4598	295405	856
22	<i>Capnocytophaga canimorsus</i> (strain 5)	-	2395	860228	324
	<b>Mollicutes</b>				
23	<i>Mycoplasma pneumoniae</i> 309	-	708	1112856	50
	<b>Bacilli</b>				
24	<i>Streptococcus equinus</i> ( <i>Streptococcus bovis</i> )	-	1996	1335	59
	<b>Spirochetes</b>				
25	<i>Borrelia hermsii</i> YOR	-	1591	1293576	168
1	<i>Enterococcus faecalis</i> (strain 62)	+	3011	936153	154
2	<i>Streptococcus pneumoniae</i> (strain 70585)	+	2179	488221	74
3	<i>Streptococcus uberis</i> (strain ATCC BAA-854 / 0140J)	+	1761	218495	59
4	<i>Staphylococcus aureus</i> (strain NCTC 8325)	+	2892	93061	112
5	<i>Listeria ivanovii</i> WSLC3009	+	2773	1457190	160
6	<i>Bacillus cereus</i> (strain ATCC 10987)	+	5835	222523	292
	<b>Clostridia</b>				
7	<i>Clostridium tetani</i> (strain Massachusetts / E88)	+	2416	212717	117
	<b>Actinobacteria</b>				
8	<i>Mycobacterium tuberculosis</i> (strain ATCC 25177 / H37Ra)	+	3993	419947	125
9	<i>Mycobacterium paratuberculosis</i> (strain ATCC BAA-968 / K-10)	+	4321	262316	135
10	<i>Mycobacterium bovis</i>	+	4353	1765	133

### 6.3 Μελέτη των περιορισμών ανίχνευσης των πρωτεϊνών του ΚΦ με μεθόδους πρωτεομικής

#### 6.3.1. Προσδιορισμός του ανιχνεύσιμου πρωτεϊνώματος σε συνθήκες πλούσιου θρεπτικού μέσου (LB)

Το πρωτεΐνωμα του *E.coli* BL21-DE3 που αναμένουμε να εκφράζεται όταν το βακτήριο αναπτύσσεται σε πλούσιο θρεπτικό μέσο (LB broth) προσδιορίστηκε από την συνδυαστική εκτίμηση γονιδιωματικών και πρωτεοματικών δεδομένων, συμπεριλαμβανομένου πειραμάτων μικροσυστοιχιών (Oberto et al, 2009; Patten et al, 2004a; Wang et al, 2005b; Yoon et al, 2012), μια μελέτη ανίχνευσης ξεχωριστών κυττάρων (single molecule) (Taniguchi et al, 2010) (Table S2) και πρωτεομικές αναλύσεις (Πίνακας 6.10, 23 studies) που πραγματοποιήθηκαν σε αντίστοιχες συνθήκες ανάπτυξης και σε κοντινά στελέχη με αυτό που μελετάμε. Τα δεδομένα μικροσυστοιχιών συλλέχθηκαν από τις βάσεις δεδομένων: ArrayExpress (Rustici et al, 2013) και GEO (Barrett et al,

---

2013; Edgar et al, 2002). Στην πλειονότητα των δεδομένων από μικροσυστοιχές τα αντίστοιχα γονία χαρακτηρίζονται 'Present', 'Marginal' ή 'Absent' ανάλογα με το επίπεδο έκφρασης (για παράδειγμα 'Present' αναφέρεται στην σίγορη έκφραση του γονιδίου etc). Από την ανάλυση των Obero *et al* (Obero et al, 2009) χρησιμοποιήσαμε τα δεδομένα από δύο από δύο πειραματικές επαναλήψεις (GSE11183 και GSM286826-7). Τα συγκεκριμένα δεδομένα αφορούν τα γονίδια που εκφράζονται σε ένα εργαστηριακό στέλεχος *E.coli* το J02057 στην εκθετική φάση. Τα δύο πειράματα συγχωνεύθηκαν καθώς τα αποτελέσματα τους ήταν ακριβώς τα ίδια. Από την ανάλυση των Yoon *et al* (Yoon et al, 2012) λάβαμε υπ όψιν δύο πειράματα (GSM325748 και GSM325750), όπου συγκρίνονται τα γονίδια που εκφράζονται στο στέλεχος B REL606 σε εκθετική (επισημασμένο με Cy5) και στατική φάση (επισημασμένο με Cy3). Χρησιμοποιήθηκαν οι κανονικοποιημένες τιμές έντασης Cy5.

Προσδιορίσαμε το γονιδίωμα που αναμένουμε να εκφράζεται σε συνθήκες LB (3770 πρωτεΐνες), λαμβάνοντας υπ όψιν μόνο γονίδια που ανιχνεύθηκαν με μεγάλη σιγουριά [τιμές ένταση μεγαλύτερες από μηδέν (Yoon et al, 2012) και επίπεδο ανίχνευσης 'Present' στα υπόλοιπα πειράματα μικροσυστοιχιών] σε τουλάχιστον δύο πειράματα. Οι δύο πειραματικές επαναλήψεις των Yoon *et al* (Yoon et al, 2012) και Patten *et al* (Patten et al, 2004a) (εννέα πειράματα στο σύνολο) συνεισέφεραν ξεχωριστά. Οι πρωτεΐνες που από τους Taniguchi *et al* (Taniguchi et al, 2010) (Table S2A) προστέθηκαν στο ανιχνεύσιμο πρωτεϊνωμασ.

**Πίνακας 6.10 – Ταυτοποίηση πρωτεϊνών του ΚΦ από πρωτεομικές αναλύσεις**

Κατάλογος των πρωτεομικών αναλύσεων φασματομετρίας μάζας με στόχο την ανίχνευση των πρωτεϊνών του ΚΦ που έχουν δημοσιευθεί μέχρι σήμερα. Αναφέρουμε το στέλεχος *E.coli* που μελετήθηκε καθώς και οι αντίστοιχες συνθήκες ανάπτυξης του, σε κάθε πρωτεομική ανάλυση.

Για να υπολογίσουμε τον συνολικό αριθμό των πρωτεϊνών που ανιχνεύθηκαν σε κάθε μελέτη αντιστοιχίσαμε την υποκυτταρική ταξινόμηση του STEPdb στα αντίστοιχα αναγνωριστικά πρωτεϊνών (protein IDs) που δημοσίευσαν συγγραφείς. Χωρίζουμε το συνολικό πρωτεϊνώμα σε δύο κατηγορίες αυτό του ΚΦ και το ΜΠ και υπολογίζουμε το συνολικό αριθμό των πρωτεϊνών που ανιχνεύθηκαν ανά κατηγορία και τα εκφράζουμε ως ποσοστά επί του θεωρητικού και του ανιχνεύσιμου πρωτεϊνώματος (ανά κατηγορία).

Author	Year	<i>E. coli</i> strain	Growth Medium	Total # proteins detected	# CEP proteins detected	# IM proteins detected	% CEP detected of theoretical CE proteome	% IM detected of theoretical IM proteome	% IM detected of anticipated IM proteome
Molloy <i>et al</i>	1999	K12 W3110	Minimal (MOPS)	13	8	0	0.3%	0.0%	0.0%
Molloy <i>et al</i>	2000	K12 W3110	Minimal (M9)	39	39	0	1.7%	0.0%	0.0%
Gevaert <i>et al</i>	2002	K12 HB2151	LB	886	458	82	19.7%	7.6%	9.3%
Yan <i>et al</i>	2002	ER1647	Minimal (MOPS)	161	102	3	4.4%	0.3%	0.3%
Corbin <i>et al</i>	2003	K12 MG1655	Minimal	1140	596	128	25.6%	11.9%	14.5%
Fountoulakis & Gasser	2003	K12 JM109	LB	358	253	6	10.9%	0.6%	0.7%
Taoka <i>et al</i>	2004	K12 JM109	LB	1469	688	157	29.6%	14.6%	17.8%
Butland <i>et al</i>	2005	DY330	-	1364	623	207	26.8%	19.2%	23.5%
Lopez-Campistrous <i>et al</i>	2005	K12 BW30270	Minimal	574	341	27	14.7%	2.5%	3.1%
Spelbrink <i>et al</i>	2005	BL21 (DE3)	LB	56	52	22	2.2%	2.0%	2.5%
Stenberg <i>et al</i>	2005	BL21 (DE3)	LB	54	54	23	2.3%	2.1%	2.6%
Arifuzzaman <i>et al</i>	2006	K12 W3110	LB	2049	889	249	38.3%	23.1%	28.2%
Baars <i>et al</i>	2006	EK413 (MC4100 derivative)	Minimal (M9)	295	214	51	9.2%	4.7%	5.8%
Huang <i>et al</i>	2006	K12	LB	104	79	2	3.4%	0.2%	0.2%
Ji <i>et al</i>	2006	K12 ATCC 47076	LB	631	483	210	20.8%	19.5%	23.8%

Lasserre <i>et al</i>	2006	DH5a	LB	299	197	30	8.5%	2.8%	3.4%
Marani <i>et al</i>	2006	K12 MG1655	Minimal (M9)	6	6	0	0.3%	0.0%	0.0%
Cirulli <i>et al</i>	2007	K12	LB	73	64	31	2.8%	2.9%	3.5%
Wagner <i>et al</i>	2007	BL21 (DE3)	LB	181	139	30	6.0%	2.8%	3.4%
Zhang <i>et al</i>	2007	K12 MG1655	Minimal (MOPS)	435	420	216	18.1%	20.0%	24.5%
Lu <i>et al</i>	2007	K12 N3433	Minimal (MOPS, glucose)	448	228	11	9.8%	1.0%	1.2%
Lee <i>et al</i>	2007	K12 (DH5a)	LB	134	106	4	4.6%	0.4%	0.5%
Jarchow <i>et al</i>	2008	MC4100	LB	42	41	0	1.8%	0.0%	0.0%
Qian <i>et al</i>	2008	BL21 (DE3)	LB	22	21	1	0.9%	0.1%	0.1%
Wagner <i>et al</i>	2008	BL21 (DE3)	LB	213	177	63	7.6%	5.8%	7.1%
Xia <i>et al</i>	2008	BL21 (DE3) & K12 W3110	Rich (R2)	81	73	1	3.1%	0.1%	0.1%
Iwasaki <i>et al</i>	2009	K12 BW25113	LB	1659	1024	430	44.1%	39.9%	48.8%
Masuda <i>et al</i>	2009	K12 BW25113	LB	1802	1018	365	43.8%	33.8%	41.4%
Vertommen <i>et al</i>	2009	MC4100 derivative	LB /	64	64	1	2.8%	0.1%	0.1%
Hemm <i>et al</i>	2010	K12 MG1655	Minimal (M63) (glucose, glycerol)	42	24	21	1.0%	1.9%	2.4%
Iwasaki <i>et al</i>	2010	K12 BW25113	LB	2598	1216	406	52.3%	37.6%	46.0%
Muller <i>et al</i>	2010	K12	LB	405	151	12	6.5%	1.1%	1.4%
Pan <i>et al</i>	2010	K12	LB	54	51	7	2.2%	0.6%	0.8%
Price <i>et al</i>	2010	FTL10	Minimal (M63)	72	72	43	3.1%	4.0%	4.9%
Thein <i>et al</i>	2010	BL21	LB	631	461	146	19.8%	13.5%	16.6%
Wickstrom <i>et al</i>	2010	WAM121	LB	132	83	10	3.6%	0.9%	1.1%
Lewis <i>et al</i>	2010	K-12 MG1655	Minimal (M9, glucose)	852	430	47	18.5%	4.4%	5.3%
Maddalo <i>et al</i>	2011	BL21 (DE3)	LB	102	102	49	4.4%	4.5%	5.6%
Han <i>et al</i>	2012	B REL606 & K12 MG1655	Rich (R2)	402	213	6	9.2%	0.6%	0.7%
Wright <i>et al</i>	2012	-	-	2413	1126	355	48.5%	32.9%	40.2%
Mancuso <i>et al</i>	2012	-	-	1047	451	45	19.4%	4.2%	5.1%
Kim (PRIDE, PRD000485)	2012	-	-	2279	1097	375	47.2%	34.8%	42.5%
Papanastasiou <i>et al</i>	2013	BL21-DE3	LB	747	363	1	15.6%	0.1%	0.1%



---

Tanca <i>et al</i>	2013	SSM5456	LB	554	278	63	12.0%	5.8%	7.1%
Krug <i>et al</i>	2013	K12 (BW25113)	Luria/Miller	2607	1254	494	54.0%	45.8%	56.0%
This study		BL21 (DE3)	LB	2243	1211	513	52.1%	47.9%	58.2%

---

### 6.3.2.Υπολογισμός της κατανομή GRAVY σε συνάρτηση με το μήκος και δισδιάστατη κατανομή των μεμβρανικών πρωτεϊνών

Οι ΔΠ των μεμβρανικών πρωτεϊνών προβλέφθηκαν με τα BE: TMHMM (Krogh et al, 2001) και Phobius (Kall et al, 2007) με τις προεπιλεγμένες παραμέτρους. Υπολογίσαμε τον μέσο όρο των ΔΠ που προβλέπουν τα δύο BE. Με κριτήριο τον αριθμό των ΔΠ οι μεμβρανικές πρωτεΐνες χωρίστηκαν σε τρεις κατηγορίες 1-2, 3-6 και 7-18 ΔΠς. Οι δισδιάστατες κατανομές του GRAVY συναρτήσεως του μήκους σχεδιάστηκαν από την MATLAB (Εικόνα 5.4C). Το εύρος των τιμών GRAVY (-1 to 2.6) χωρίστηκε σε 40 περιοχές ενώ ο μήκος (19-1342 AAs) σε 30. Σε κάθε περιοχή τιμών GRAVY υπολογίστηκε η μέση τιμή μήκους. Οι κατανομές αυτές προσεγγίστηκαν ως κατανομές τύπου Lorentzian.

### 6.3.3.Προβλεπόμενο διαμεμβρανικό κομμάτι των πεπτιδίων

Τα ανιχνεύσιμα ('MS-detectable') και μη ανιχνεύσιμα ('MS-undetectable') πεπτίδια αναλύθηκαν με βάση τις προβλεπόμενες ΔΠ των αντίστοιχων πρωτεϊνών. Οι ΔΠ προβλέφθηκαν από το Phobius (Kall et al, 2007). Στη συνέχεια για κάθε πεπτίδιο υπολογίστηκε ο αριθμός των αμινοξέων που ανήκουν σε ΔΠ και διαιρέθηκε με το συνολικό μήκος του πεπτιδίου εκφραζόμενο ως ποσοστό (predicted TM region (%);Εικόνα 5.4F).

### 6.3.4.Ανάλυση των μεταγραφικών μονάδων (TUs) που εκφράζονται

Οι μεταγραφικές μονάδες (TUs: transcription units) του *E.coli* BL21-DE3 κατεβάστηκαν από την βάση δεδομένων EcoCyc (Keseler et al, 2009). 407 μεμβρανικές πρωτεΐνες βρέθηκαν ότι ανήκουν σε μεταγραφικές μονάδες που κωδικοποιούν μια πρωτεΐνη ενώ 315 σε πολυκιστρονικές μεταγραφικές μονάδες. Για 62 μεταγραφικές μονάδες τουλάχιστον μία από τις πρωτεΐνες που κωδικοποιούν βρέθηκε σε επίπεδο πρωτεΐνης στην παρούσα ανάλυση. Αυτές οι μεταγραφικές μονάδες θεωρήθηκαν ότι εκφράζονται στην πληρότητα τους και ότι όλες οι πρωτεΐνες που κωδικοποιούν αναμένεται να είναι ανιχνεύσιμες (Πίνακας 6.11).

Πίνακας 6.11 – Μεταγραφικές μονάδες που αναμένεται να εκφράζονται σε επίπεδο πρωτεΐνης

BL21-DE3 Entry Name	Gene Name	Sub-cellular Location	Operon (EcoCyc)	Detected in this study (YES/NO)	Expected to be expressed based on corresponding TU* (YES/NO)	
C6EGJ5	acrE	E	acrE-acrF	C6EGJ5 (YES)	C6EGJ5 (NO)	
C6EGJ4	acrF	B		C6EGJ4 (NO)	C6EGJ4 (YES)	
C6EH90	agaA	A	agaZ-agaV-agaW-agaE-agaF-agaA	C6EH90 (NO)	C6EH90 (NO)	
C6EH92	agaE	B		C6EH92 (NO)	C6EH92 (YES)	
C6EH91	agaF	F1		C6EH91 (NO)	C6EH91 (NO)	
C6EH94	agaV	A		C6EH94 (YES)	C6EH94 (NO)	
C6EH93	agaW	B		C6EH93 (NO)	C6EH93 (YES)	
C6EH95	kbaZ	A		C6EH95 (NO)	C6EH95 (NO)	
C6EIH7	cmtA	B		cmtA-yggP-yggF-yggD-yggC	C6EIH7 (NO)	C6EIH7 (YES)
C6EII1	yggC	A	C6EII1 (YES)		C6EII1 (NO)	
C6EII0	yggD	N	C6EII0 (NO)		C6EII0 (NO)	
C6EIH9	yggF	A	C6EIH9 (NO)		C6EIH9 (NO)	
C6EIH8	yggP	A	C6EIH8 (NO)		C6EIH8 (NO)	
C6ELE0	cvpA	B	cvpA-purF		C6ELE0 (NO)	C6ELE0 (YES)
C6ELE1	purF	F1		C6ELE1 (YES)	C6ELE1 (NO)	
C6EI88	dmsA	F2	dmsA-dmsB-dmsC	C6EI88 (YES)	C6EI88 (NO)	
C6EI87	dmsB	F2		C6EI87 (NO)	C6EI87 (NO)	
C6EI86	dmsC	B		C6EI86 (NO)	C6EI86 (YES)	
C6EAD1	fhuB	B	fhuC-fhuD-fhuB	C6EAD1 (NO)	C6EAD1 (YES)	
C6EAD3	fhuC	F1		C6EAD3 (YES)	C6EAD3 (NO)	
C6EAD2	fhuD	G		C6EAD2 (YES)	C6EAD2 (NO)	
C6EJ58	gcvA	N	gcvA-ygdD-ygdE	C6EJ58 (YES)	C6EJ58 (NO)	
C6EJ59	ygdD	B		C6EJ59 (NO)	C6EJ59 (YES)	
C6EKG6	hcaB	A	hcaE-hcaF-hcaC-hcaB-hcaD-yphA	C6EKG6 (NO)	C6EKG6 (NO)	
C6EKG7	hcaC	A		C6EKG7 (NO)	C6EKG7 (NO)	
C6EKG5	hcaD	A		C6EKG5 (NO)	C6EKG5 (NO)	
C6EKG9	hcaE	A		C6EKG9 (YES)	C6EKG9 (NO)	
C6EKG8	hcaF	A		C6EKG8 (YES)	C6EKG8 (NO)	
C5W7P7	yphA	B		C5W7P7 (NO)	C5W7P7 (YES)	
C6EBL8	hpaD	A		hpaG-hpaE-hpaD-hpaF-hpaH-hpaI-hpaX-hpaA	C6EBL8 (NO)	C6EBL8 (NO)
C6EBL9	hpaF	A	C6EBL9 (NO)		C6EBL9 (NO)	
C6EBL6	hpaG	A	C6EBL6 (NO)		C6EBL6 (NO)	
C6EBM0	hpaH	A	C6EBM0 (YES)		C6EBM0 (NO)	
C6EBM2	hpaX	B	C6EBM2 (NO)		C6EBM2 (YES)	
C6EBM3	ybl222	A	C6EBM3 (NO)		C6EBM3 (NO)	
C6EBL7	ybl227	A	C6EBL7 (YES)		C6EBL7 (NO)	
C6EI10	hyaA	B	hyaA-hyaB-hyaC-hyaD-hyaE-hyaF		C6EI10 (NO)	C6EI10 (YES)
C6EHN5	hyaB	F1			C6EHN5 (YES)	C6EHN5 (NO)
C6EHN4	hyaC	B		C6EHN4 (NO)	C6EHN4 (YES)	
C6EHN3	hyaD	A		C6EHN3 (NO)	C6EHN3 (NO)	
C6EHN2	hyaE	A		C6EHN2 (NO)	C6EHN2 (NO)	
C6EHN1	hyaF	A		C6EHN1 (NO)	C6EHN1 (NO)	

C6EJE4	hycB	A		C6EJE4 (NO)	C6EJE4 (NO)
C6EJE5	hycC	B		C6EJE5 (NO)	C6EJE5 (YES)
C6EJE6	hycD	B		C6EJE6 (NO)	C6EJE6 (YES)
C6EJE7	hycE	F1	hycB-hycC-hycD-hycE-	C6EJE7 (NO)	C6EJE7 (NO)
C6EJE8	hycF	A	hycF-hycG-hycH-hycl	C6EJE8 (NO)	C6EJE8 (NO)
C6EJR2	hycG	A		C6EJR2 (YES)	C6EJR2 (NO)
C6EJR3	hycH	A		C6EJR3 (NO)	C6EJR3 (NO)
C6EJR4	hycl	A		C6EJR4 (NO)	C6EJR4 (NO)
C6EKL9	focB	B		C6EKL9 (NO)	C6EKL9 (YES)
C6EKM7	hyfD	B		C6EKM7 (NO)	C6EKM7 (YES)
C6EKM6	hyfE	B		C6EKM6 (NO)	C6EKM6 (YES)
C6EKM5	hyfF	B	hyfD-hyfE-hyfF-hyfG-	C6EKM5 (NO)	C6EKM5 (YES)
C6EKM4	hyfG	F1	hyfH-hyfI-hyfJ-hyfR-	C6EKM4 (NO)	C6EKM4 (NO)
C6EKM3	hyfH	A	focB	C6EKM3 (NO)	C6EKM3 (NO)
C6EKM2	hyfI	A		C6EKM2 (NO)	C6EKM2 (NO)
C6EKM1	hyfJ	A		C6EKM1 (NO)	C6EKM1 (NO)
C5W7J3	hyfR	N		C5W7J3 (YES)	C5W7J3 (NO)
C5WBS7	dinF	B	lexA-dinF	C5WBS7 (NO)	C5WBS7 (YES)
C6EDX7	lexA	N		C6EDX7 (YES)	C6EDX7 (NO)
C5WA29	livF	F1		C5WA29 (NO)	C5WA29 (NO)
C6EFB9	livG	F1	livH-livM-livG-livF	C6EFB9 (YES)	C6EFB9 (NO)
C6EFB7	livH	B		C6EFB7 (NO)	C6EFB7 (YES)
C5WA31	livM	B		C5WA31 (NO)	C5WA31 (YES)
C6EF98	nikA	G		C6EF98 (NO)	C6EF98 (NO)
C6EF97	nikB	B		C6EF97 (NO)	C6EF97 (YES)
C6EF96	nikC	B	nikA-nikB-nikC-nikD-	C6EF96 (NO)	C6EF96 (YES)
C6EF95	nikD	F1	nikE-nikR	C6EF95 (NO)	C6EF95 (NO)
C6EF94	nikE	F1		C6EF94 (YES)	C6EF94 (NO)
C6EF93	nikR	N		C6EF93 (YES)	C6EF93 (NO)
C6EFK8	cysG	A	nirC-cysG	C6EFK8 (YES)	C6EFK8 (NO)
C6EFK9	nirC	B		C6EFK9 (NO)	C6EFK9 (YES)
C6EJV0	nrdE	A	nrdE-nrdF	C6EJV0 (YES)	C6EJV0 (NO)
C6EJU9	nrdF	B		C6EJU9 (NO)	C6EJU9 (YES)
C5WBV5	nrfE	B	nrfE-nrfF-nrfG	C5WBV5 (NO)	C5WBV5 (YES)
C6ED50	nrfF	B		C6ED50 (NO)	C6ED50 (YES)
C6ED49	nrfG	I		C6ED49 (YES)	C6ED49 (NO)
C6EAC6	btuF	G	pfs-yadT-yadS	C6EAC6 (YES)	C6EAC6 (NO)
C6EAC7	yadS	B		C6EAC7 (NO)	C6EAC7 (YES)
C6VFL0	phnE	B		C6VFL0 (NO)	C6VFL0 (YES)
C6ED21	phnF	N		C6ED21 (YES)	C6ED21 (NO)
C6ED22	phnG	A	phnE-phnF-phnG-	C6ED22 (NO)	C6ED22 (NO)
C6ED23	phnH	A	phnH-phnI-phnJ-phnK	C6ED23 (NO)	C6ED23 (NO)
C6ED24	phnI	A		C6ED24 (NO)	C6ED24 (NO)
C6ED25	phnJ	A		C6ED25 (NO)	C6ED25 (NO)
C6ED26	phnK	F1		C6ED26 (NO)	C6ED26 (NO)

C6EJ43	ppdA	G		C6EJ43 (NO)	C6EJ43 (NO)
C6EJ44	ppdB	B	ppdA-ppdB-ygdB-	C6EJ44 (NO)	C6EJ44 (YES)
C6EJ46	ppdC	B	ppdC-recC	C6EJ46 (NO)	C6EJ46 (YES)
C6EJ47	recC	A		C6EJ47 (YES)	C6EJ47 (NO)
C6EJ45	ygdB	G		C6EJ45 (NO)	C6EJ45 (NO)
C6EGQ4	hemK	A		C6EGQ4 (YES)	C6EGQ4 (NO)
C6EGQ1	kdsA	A	prfA-hemK-ychQ-ychA-	C6EGQ1 (YES)	C6EGQ1 (NO)
C6EGQ5	prfA	A	kdsA	C6EGQ5 (YES)	C6EGQ5 (NO)
C6EGQ2	ychA	A		C6EGQ2 (YES)	C6EGQ2 (NO)
C6EGQ3	ychQ	B		C6EGQ3 (NO)	C6EGQ3 (YES)
C6EG53	rbsA	F1		C6EG53 (NO)	C6EG53 (NO)
C6EG51	rbsB	G	rbsA-rbsC-rbsB	C6EG51 (YES)	C6EG51 (NO)
C6EG52	rbsC	B		C6EG52 (NO)	C6EG52 (YES)
C6EC54	sgcB	F1		C6EC54 (NO)	C6EC54 (NO)
C6EC55	sgcC	B	sgcX-sgcB-sgcC-sgcQ	C6EC55 (NO)	C6EC55 (YES)
C6EC56	sgcQ	A		C6EC56 (YES)	C6EC56 (NO)
C6EBT0	sgcX	A		C6EBT0 (NO)	C6EBT0 (NO)
C6EJS9	srlA	B		C6EJS9 (NO)	C6EJS9 (YES)
C6EJS7	srlB	F1	srlA-srlE-srlB-srlD	C6EJS7 (NO)	C6EJS7 (NO)
C6EJS6	srlD	A		C6EJS6 (YES)	C6EJS6 (NO)
C6EJS8	srlE	B		C6EJS8 (NO)	C6EJS8 (YES)
C6EI46	ssuA	G		C6EI46 (NO)	C6EI46 (NO)
C6EI49	ssuB	F1	ssuE-ssuA-ssuD-ssuC-	C6EI49 (NO)	C6EI49 (NO)
C6EI48	ssuC	B	ssuB	C6EI48 (NO)	C6EI48 (YES)
C6EI47	ssuD	A		C6EI47 (YES)	C6EI47 (NO)
C6EI45	ssuE	A		C6EI45 (YES)	C6EI45 (NO)
C6ELL9	tauA	G		C6ELL9 (NO)	C6ELL9 (NO)
C6ELL8	tauB	F1	tauA-tauB-tauC-tauD	C6ELL8 (YES)	C6ELL8 (NO)
C6ELL7	tauC	B		C6ELL7 (NO)	C6ELL7 (YES)
C6ELL6	tauD	A		C6ELL6 (YES)	C6ELL6 (NO)
C6EAY8	thiP	B	thiP-yabJ	C6EAY8 (NO)	C6EAY8 (YES)
C6EAY9	yabJ	F1		C6EAY9 (YES)	C6EAY9 (NO)
C6EDN0	uidA	F1	uidA-uidB-uidC	C6EDN0 (NO)	C6EDN0 (NO)
C6EDN1	uidB	B		C6EDN1 (NO)	C6EDN1 (YES)
C6EDN2	uidC	H		C6EDN2 (YES)	C6EDN2 (NO)
C6ECF5	sgaB	F1	ulaA-sgaB-ptxA-sgaH-	C6ECF5 (NO)	C6ECF5 (NO)
C6ECF1	sgaE	A	sgaU-sgaE	C6ECF1 (NO)	C6ECF1 (NO)
C6ECF3	sgaH	A		C6ECF3 (YES)	C6ECF3 (NO)
C6ECF6	ulaA	B		C6ECF6 (NO)	C6ECF6 (YES)
C6ECF4	ulaC	F1		C6ECF4 (NO)	C6ECF4 (NO)
C6ECF2	ulaE	A		C6ECF2 (NO)	C6ECF2 (NO)
C6EIQ4	nfsA	A	ybjC-nfsA	C6EIQ4 (YES)	C6EIQ4 (NO)
C6EIQ5	ybjC	B		C6EIQ5 (NO)	C6EIQ5 (YES)
C6EIT9	ybjI	A	ybjJ-ybjI	C6EIT9 (YES)	C6EIT9 (NO)
C6EIT8	ybjJ	B		C6EIT8 (NO)	C6EIT8 (YES)
C6EHF7	yceI	G	yceJ-yceI	C6EHF7 (YES)	C6EHF7 (NO)
C6EHF6	yceJ	B		C6EHF6 (NO)	C6EHF6 (YES)

C6EFS3	ycjM	A		C6EFS3 (NO)	C6EFS3 (NO)
C6EFS2	ycjN	I		C6EFS2 (YES)	C6EFS2 (NO)
C6EFS1	ycjO	B		C6EFS1 (NO)	C6EFS1 (YES)
C6EFS0	ycjP	B		C6EFS0 (NO)	C6EFS0 (YES)
C6EFR9	ycjQ	A	ycjM-ycjN-ycjO-ycjP- ycjQ-ycjR-ycjS-ycjT- ycjU-ycjV	C6EFR9 (NO)	C6EFR9 (NO)
C6EFR8	ycjR	F1		C6EFR8 (NO)	C6EFR8 (NO)
C6EFR7	ycjS	A		C6EFR7 (NO)	C6EFR7 (NO)
C6EFR6	ycjT	A		C6EFR6 (NO)	C6EFR6 (NO)
C6EFR5	ycjU	A		C6EFR5 (NO)	C6EFR5 (NO)
C6EFR4	ycjV	F1		C6EFR4 (NO)	C6EFR4 (NO)
C6EEM1	yddA	B	yddA-yddB	C6EEM1 (NO)	C6EEM1 (YES)
C6EEM2	yddB	G		C6EEM2 (YES)	C6EEM2 (NO)
C5W545	yeaV	B	yeaV-yeaW	C5W545 (NO)	C5W545 (YES)
C6EC29	yeaW	A		C6EC29 (YES)	C6EC29 (NO)
C6EAI0	osmF	G		C6EAI0 (YES)	C6EAI0 (NO)
C6EAI3	yehW	B	yehZ-yehY-yehX-yehW	C6EAI3 (NO)	C6EAI3 (YES)
C6EAI2	yehX	F1		C6EAI2 (YES)	C6EAI2 (NO)
C6EAI1	yehY	B		C6EAI1 (NO)	C6EAI1 (YES)
C6E9V5	yfaU	A	yfaV-yfaU	C6E9V5 (YES)	C6E9V5 (NO)
C6E9V4	yfaV	B		C6E9V4 (NO)	C6E9V4 (YES)
C6EKJ0	pbpC	B	yfhM-pbpC	C6EKJ0 (NO)	C6EKJ0 (YES)
C6EKI9	yfhM	E		C6EKI9 (YES)	C6EKI9 (NO)
C6EE99	lyxK	A		C6EE99 (NO)	C6EE99 (NO)
C6EE96	sgbE	A		C6EE96 (NO)	C6EE96 (NO)
C6EE98	sgbH	A		C6EE98 (NO)	C6EE98 (NO)
C6EE97	sgbU	A	yiaM-yiaN-yiaO-lyxK- sgbH-sgbU-sgbE	C6EE97 (YES)	C6EE97 (NO)
C6EEA2	yiaM	B		C6EEA2 (NO)	C6EEA2 (YES)
C6EEA1	yiaN	B		C6EEA1 (NO)	C6EEA1 (YES)
C6EEA0	yiaO	G		C6EEA0 (NO)	C6EEA0 (NO)
C6EF14	dtd	A		C6EF14 (YES)	C6EF14 (NO)
C6EF16	yihX	A	yihX-rbn-dtd-yiiD	C6EF16 (NO)	C6EF16 (NO)
C6EF15	yihY	B		C6EF15 (YES)	C6EF15 (NO)
C5WBC9	yiiD	B		C5WBC9 (NO)	C5WBC9 (YES)
C6EE46	fabR	N	yijC-yijD	C6EE46 (YES)	C6EE46 (NO)
C6EE45	yijD	B		C6EE45 (NO)	C6EE45 (YES)
C6ED43	mdtO	B		C6ED43 (NO)	C6ED43 (YES)
C6ED42	sdsR	B	yjcR-yjcQ-yjcP	C6ED42 (NO)	C6ED42 (YES)
C6ED44	yjcP	I		C6ED44 (YES)	C6ED44 (NO)
C6EBQ1	iadA	A		C6EBQ1 (YES)	C6EBQ1 (NO)
C6EBQ0	yjiG	B	yjiH-yjiG-iadA	C6EBQ0 (NO)	C6EBQ0 (YES)
C6EBP9	yjiH	B		C6EBP9 (NO)	C6EBP9 (YES)
C6EKG1	yphD	B		C6EKG1 (NO)	C6EKG1 (YES)
C6EKG0	yphE	F1	yphF-yphE-yphD	C6EKG0 (YES)	C6EKG0 (NO)
C6EKF9	yphF	G		C6EKF9 (NO)	C6EKF9 (NO)
C6EBW8	znuB	B	znuC-znuB	C6EBW8 (NO)	C6EBW8 (YES)
C6EBW9	znuC	F1		C6EBW9 (YES)	C6EBW9 (NO)

### 6.3.5. Κριτήρια για το καθορισμό των πρωτεϊνών που ανιχνεύονται συστηματικά

Πριν υπολογίσουμε τις σχετικές ποσότητες των πρωτεϊνών, επιλέξαμε το σύνολο των πρωτεϊνών που ανιχνεύονται με συστηματικό τρόπο (επαναλαμβανόμενα) θέτοντας ορισμένα κριτήρια με βάση τις τεχνικές και βιολογικές επαναλήψεις. Στα πειράματα όπου δεν έγινε επεξεργασία των ΑΜΠ θέσαμε ως κριτήριο να έχουν ανιχνευθεί σε τουλάχιστον δύο τεχνικές επαναλήψεις εντός της βιολογικής επανάληψης και σε τουλάχιστον δύο βιολογικές επαναλήψεις. Στα επεξεργασμένα ΑΜΠ (untreated) θέσαμε ως κριτήριο μια πρωτεΐνη να έχει βρεθεί σε τουλάχιστον δύο τεχνικές επαναλήψεις (2/3) εντός της βιολογικής επανάληψης και σε τουλάχιστον μία βιολογική επανάληψη (1/2) στα αγρίου τύπου κύτταρα αλλά και στα κύτταρα με υπερέκφραση των SecYEG.

### 6.3.6. Μοντελοποίηση και προσδιορισμός πεπτιδίων που ανιχνεύονται με την πειραματική διαδικασία της πρωτεόλυσης επιφάνειας ΑΜΚ

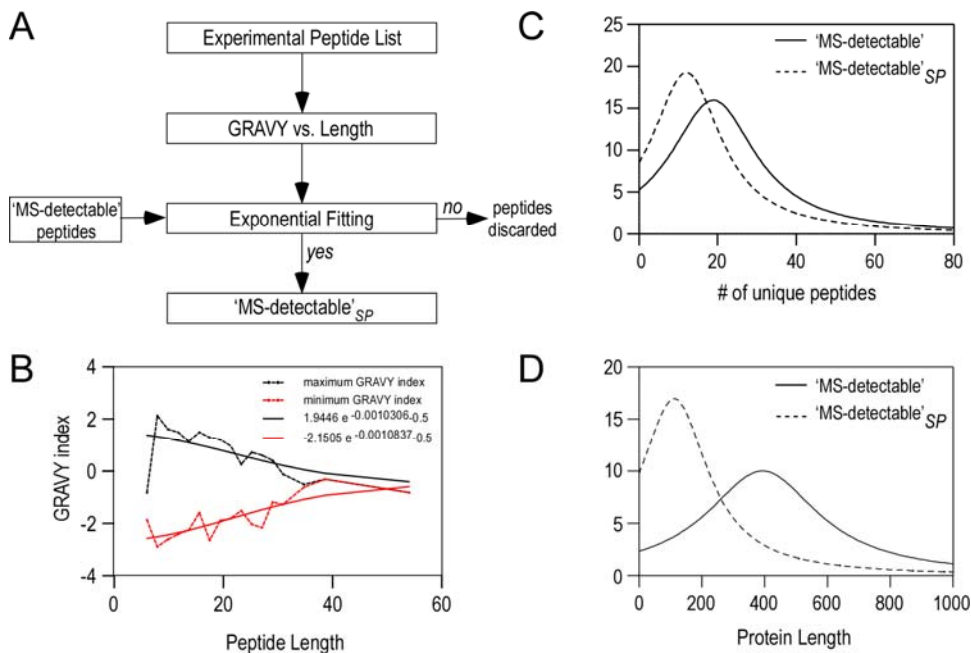
Η πρωτότυπη εξίσωση υπολογισμού των τιμών emPAI βασίζεται στον υπολογισμό του λόγου ανάμεσα στον αριθμό των μοναδικών πεπτιδίων που ανιχνεύθηκαν πειραματικά (detected) και στον θεωρητικό αριθμό πεπτιδίων (observable).

$$emPAI = 10^{\frac{\#detected}{\#observable}}$$

Αποδείξαμε με την θεωρητική ανάλυση των πεπτιδίων των μεμβρανικών πρωτεϊνών ότι στην περίπτωση των μεμβρανικών πρωτεϊνών πολύ λιγότερα πεπτίδια μπορούν να ανιχνευθούν (Εικόνα 5.4E) επειδή: α) ο λόγος μάζας προς φορτίο (m/z) βγαίνει εκτός ορίων του φασματογράφου, β) υπάρχει μικρή πιθανότητα πρόσληψης θετικών φορτίων (<3) γ) υπάρχουν περιορισμοί που προκύπτουν από την μέθοδο προετοιμασίας του δείγματος (π.χ. υδρόφοβα και μεγάλα πεπτίδια δεν απομονώνονται εύκολα από πηκτώματα).

Αποφασίσαμε για πιο σωστό υπολογισμό της ποσότητας των μεμβρανικών πρωτεϊνών να τροποποιήσουμε τον αρχικό τύπο για τον υπολογισμό του emPAI χρησιμοποιώντας αντί για τον θεωρητικό αριθμό πεπτιδίων τον αριθμό των ανιχνεύσιμων πεπτιδίων από την μέθοδο της πρωτεόλυσης στην επιφάνεια ΑΜΚ ('MS-detectable'<sub>SP</sub> Εικόνα 6.1 C και D).

$$emPAI = 10^{\frac{\#detected}{\#MS\ detectable_{SP}}}$$



**Εικόνα 6.1-** Ανάλυση για τον ορισμό των πεπτιδίων που ανιχνεύονται με την μέθοδο της πρωτεόλυσης της επιφάνειας AMK ('MS-detectable' SP).

A. Διάγραμμα ροής που συνοψίζει τα βήματα που ακλουθήσαμε για να ορίσουμε τα πεπτίδια που ανιχνεύονται με την μέθοδο της πρωτεόλυσης της επιφάνειας AMK ('MS-detectable' SP)

B. Κατανομές μέγιστου (μαύρο) και ελάχιστου (κόκκινο) GRAVY συναρτήσε μήκους πεπτιδίου και η προσέγγιση τους με εκθετικές εξισώσεις.

C. Σύγκριση αριθμού ανιχνεύσιμων πεπτιδίων ('MS-detectable') και αριθμού ανιχνεύσιμων πεπτιδίων με την μέθοδο πρωτεόλυσης στην επιφάνεια AMK ('MS-detectable' SP)

D. Σύγκριση πραγματικού μήκους πρωτεϊνών και διορθωμένου με βάση τον αντίστοιχο ορισμό των πεπτιδία που ανιχνεύονται με την μέθοδο της πρωτεόλυσης της επιφάνειας AMK

Αρχικά υπολογίσαμε τις καμπύλες μέγιστης και ελάχιστης υδροφοβικότητας συναρτήσε του μήκους των πεπτιδίων ανιχνεύθηκαν πειραματικά από την μέθοδο της επιφανειακής πρωτεόλυσης, δηλαδή για κάθε τιμή μήκους πεπτιδίων (4-60) υπολογίσαμε το μέγιστο και ελάχιστο GRAVY index. Στις συνέχεια οι καμπύλες αυτές μοντελοποιήθηκαν με εκθετικές συναρτήσεις του τύπου  $\alpha \cdot e^{\pm \beta \cdot x^2} - \gamma$  (Εικόνα 6.1B). Οι εκθετικές καμπύλες που προέκυψαν χρησιμοποιήθηκαν για τον καθορισμό των ανιχνεύσιμων πεπτιδίων με την μέθοδο της πρωτεόλυσης επιφάνειας ('MS-detectable' SP) ως εξής: το GRAVY index ενός πεπτιδίου δεν θα πρέπει να είναι μεγαλύτερο αλλά ούτε και μικρότερο από το θεωρητικό μέγιστο και ελάχιστο με



βάση τις θεωρητικές καμπύλες που μοντελοποιούν τα πεπτίδια που ανιχνεύονται με την συγκεκριμένη μέθοδο. Με άλλα λόγια το GRAVY index ενός πεπτιδίου θα πρέπει να βρίσκεται στην ενδιάμεση περιοχή που ορίζουν αυτές οι δύο καμπύλες.

### 6.3.7. Πρωτεόλυση επιφάνειας – υπολογισμός ανιχνεύσιμου μήκους πρωτεΐνης ( $L_{SP}$ )

Επιλέξαμε τα θεωρητικά ανιχνεύσιμα πεπτίδια (“MS-detectable” με βάση την μοντελοποίηση των χαρακτηριστικών των πεπτιδίων που ανιχνεύτηκαν με την πρωτεόλυση επιφάνειας (δες προηγούμενη ενότητα). Με βάση το υποσύνολο των ανιχνεύσιμων πεπτιδίων με πρωτεόλυση επιφάνειας (SP “MS-detectable” peptides) υπολογίσαμε το αναμενόμενο ανιχνεύσιμο μήκος κάθε πρωτεΐνης (Εικόνα 6.1D)

### 6.3.8. Διόρθωση τιμών σχετικής ποσότητας NSAF ( $NSAF_{SP}$ )

Η πρωτότυπη εξίσωση υπολογισμού των τιμών NSAF υπολογίστηκαν ως:

$$NSAF_i = \frac{\frac{SC_i}{L_i}}{\sum_{i=1}^n \frac{SC_i}{L_i}}$$

και διορθώθηκαν ως:

$$NSAF^{SP}_i = \frac{\frac{SC_i}{L_i^{SP}}}{\sum_{i=1}^n \frac{SC_i}{L_i^{SP}}}$$

όπου  $n$  είναι ο αριθμός των πρωτεϊνών που ανιχνεύθηκαν σε κάθε δείγμα (δες ενότητα 6.3.5),  $SC_i$  ο αριθμός φασμάτων για την πρωτεΐνη  $i$ ,  $L_i$  το πραγματικό μήκος της πρωτεΐνης  $i$  και  $L_i^{SP}$  το ανιχνεύσιμο μήκος από πρωτεόλυση επιφάνειας (δες προηγούμενη ενότητα). Ο βαθμός διόρθωσης μεταξύ τιμών  $NSAF_{SP}$  και πρωτότυπων τιμών NSAF υπολογίστηκε ως ο λόγος  $NSAF_{SP}/NSAF$ .

### 6.3.9. Διόρθωση τιμών σχετικής ποσότητας emPAI (emPAI<sub>SP</sub>)

Οι τιμές emPAI διορθώθηκαν με βάση τον αριθμό των ανιχνεύσιμων πεπτιδίων στην πρωτεόλυση επιφάνειας. Δεν είναι όλα τα θεωρητικά πεπτίδια ανιχνεύσιμα λόγω α) τεχνικά όρια  $m/z$  του φασματογράφου μάζας β) λόγω άλλων περιορισμών που προκύπτουν από την επεξεργασία των δειγμάτων. Δοκιμάσαμε δύο εναλλακτικούς τρόπους διόρθωσης του υπολογισμού των τιμών emPAI: έναν χρησιμοποιώντας τον αριθμό των ανιχνεύσιμων πεπτιδίων και έναν δεύτερο χρησιμοποιώντας τον αριθμό των ανιχνεύσιμων πεπτιδίων στην πρωτεόλυση επιφάνειας. Υπολογίσαμε την διορθωμένη τιμή emPAI<sub>SP</sub> στην δεύτερη περίπτωση ως εξής:

$$emPAI_{SP} = 10^{\frac{\#detected\ peptides}{\#surf\ proteolysis\ observable\ peptides}} - 1$$

### 6.3.10. Στατιστική ανάλυση για τον καθορισμό των πρωτεϊνών που η ποσότητα τους έχει μεταβληθεί σημαντικά ανάμεσα στα αγρίου τύπου και SecYEG κύτταρα

#### Αντικατάσταση των μηδενικών τιμών (SC/UP)

Ο αριθμός φασμάτων (SC: Spectral counts) και ο αριθμός των μοναδικών πεπτιδίων (UP: unique peptides) αναλύθηκαν και οι μηδενικές τιμές αντικαταστάθηκαν όπως έχει περιγραφεί παλαιότερα (Zybailov et al, 2006). Υπολογίσαμε την κατανομή που ακολουθούν οι λογαριθμημένες τιμές φασμάτων/μοναδικών πεπτιδίων (SC/UP) (μέσος όρος σε όλες τις τεχνικές επαναλήψεις, ξεχωριστά για αγρίου τύπου και secYEG κύτταρα). Οι κατανομές αυτές ελέγχθηκαν αν ακολουθούν κανονική χρησιμοποιώντας το στατιστικό τεστ κανονικότητας Shapir-Wilk. Εν τέλει οι μηδενικές τιμές αντικαταστάθηκαν με μια τιμή μικρότερη του ενός η οποία δεν επιρεάζει την κανονικότητα των αντίστοιχων κατανομών.

#### Beta-binomial testing (NSAF<sub>SP</sub>)

Ο μέσος αριθμός φασμάτων (SCs) των τεχνικών επαναλήψεων υπολογίστηκε για κάθε βιολογική επανάληψη. Πραγματοποιήσαμε στατιστικό τεστ β-binomial (Pham et al, 2010) από το οποίο καθορίστηκαν οι πρωτεΐνες που η ποσότητα τους έχει μεταβληθεί σημαντικά ανάμεσα στα αγρίου τύπου κύτταρα και σε αυτά όπου το σύστημα SecYEG υπερ-εκφράζεται. Πρωτεΐνες με τιμές p-values μικρότερες από 5% θεωρήθηκαν στατιστικά σημαντικές.

---

*T-test (emPAI<sub>SP</sub>)*

Για να υπολογίσουμε τις πρωτεΐνες που οι ποσότητες τους μεταβάλλονται σημαντικά, με βάση τις αντίστοιχες τιμές emPAI<sub>SP</sub> χρησιμοποιήσαμε t-test (unpaired, two-sided) με την μηδενική υπόθεση ότι οι διασπορές των δύο κατανομών είναι ίσες. Οι τιμές emPAI<sub>SP</sub> υπολογίστηκαν από τον μέσο όρο μοναδικών πεπτιδίων (UPs) από όλες τις τεχνικές επαναλήψεις εντός κάθε βιολογικής επανάληψης. Πρωτεΐνες με τιμές p-values μικρότερες από 5% θεωρήθηκαν στατιστικά σημαντικές.

## Βιβλιογραφία

- Adams H et al (2003) Interactions between phage-shock proteins in *Escherichia coli*. *Journal of Bacteriology* **185**: 1174-1180
- Adelman MR et al (1973) Ribosome-membrane interaction. *The Journal of Cell Biology* **56**: 206-229
- Aivaliotis M et al (2006) Proteomic analysis of chlorosome-depleted membranes of the green sulfur bacterium *Chlorobium tepidum*. *PROTEOMICS* **6**: 217-232
- Aivaliotis M et al (2007) An alternative strategy for the membrane proteome analysis of the green sulfur bacterium *Chlorobium tepidum* using blue native PAGE and 2-D PAGE on purified membranes. *J Proteome Res* **6**: 1048-1058
- Alexandersson E et al (2004) Arabidopsis plasma membrane proteomics identifies components of transport, signal transduction and membrane trafficking. *Plant and Cell Physiology* **45**: 1543-1556
- Aliferis CF et al (2003) HITON: a novel Markov Blanket algorithm for optimal variable selection. *AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium*: 21-25
- Allen TE et al (2003) Genome-Scale Analysis of the Uses of the *Escherichia coli* Genome: Model-Driven Analysis of Heterogeneous Data Sets. *Journal of Bacteriology* **185**: 6392-6399
- Aranda B et al (2010) The IntAct molecular interaction database in 2010. *Nucleic Acids Res* **38**: D525-D531
- Arifuzzaman M et al (2006) Large-scale identification of protein-protein interaction of *Escherichia coli* K-12. *Genome Research* **16**: 686-691
- Baars L et al (2006) Defining the role of the *Escherichia coli* chaperone SecB using comparative proteomics. *J Biol Chem* **281**: 10024-10034
- Badea L et al (2009) Secretion of flagellin by the LEE-encoded type III secretion system of enteropathogenic *Escherichia coli*. *BMC Microbiol* **9**: 30
- Bagos PG et al (2010) Combined prediction of Tat and Sec signal peptides with hidden Markov models. *Bioinformatics* **26**: 2811-2817
- Bankaitis VA et al (1984) Intragenic suppressor mutations that restore export of maltose binding protein with a truncated signal peptide. *Cell* **37**: 243-252
- Barnhart MM et al (2006) Curli biogenesis and function. *Annu Rev Microbiol* **60**: 131-147
- Barrett T et al (2013) NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Research* **41**: D991-D995
- Bassilana M et al (1992) The role of the mature domain of proOmpA in the translocation ATPase reaction. *J Biol Chem* **267**: 25246-25250
- Baud C et al (2002) Allosteric communication between signal peptides and the SecA protein DEAD motor ATPase domain. *Journal of Biological Chemistry* **277**: 13724-13731
- Bednarska NG et al (2013) Protein aggregation in bacteria: the thin boundary between functionality and toxicity. *Microbiology* **159**: 1795-1806
- Bedouelle H et al (1980) Mutations which alter the function of the signal sequence of the maltose binding protein of *Escherichia coli*. *Nature* **285**: 78-81
- Bendezu FO et al (2009a) RodZ (YfgA) is required for proper assembly of the MreB actin cytoskeleton and cell shape in *E. coli*. *EMBO J* **28**: 193-204
- Bendezu FO et al (2009b) RodZ (YfgA) is required for proper assembly of the MreB actin cytoskeleton and cell shape in *E. coli*. *Embo Journal* **28**: 193-204
- Bendtsen JD et al (2005a) Non-classical protein secretion in bacteria. *BMC Microbiol* **5**: 58
- Bendtsen JD et al (2004) Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol* **340**: 783-795
- Bendtsen JD et al (2005b) Prediction of twin-arginine signal peptides. *BMC Bioinformatics* **6**: 167

- Bernsel A et al (2009) Exploring the inner membrane proteome of *Escherichia coli*: which proteins are eluding detection and why? *Trends Microbiol* **17**: 444-449
- Beveridge TJ (1999) Structures of gram-negative cell walls and their derived membrane vesicles. *J Bacteriol* **181**: 4725-4733
- Bilous PT et al (1988) Nucleotide sequence of the *dmsABC* operon encoding the anaerobic dimethylsulphoxide reductase of *Escherichia coli*. *Molecular Microbiology* **2**: 785-795
- Bishop RE (2008) Structural biology of membrane-intrinsic beta-barrel enzymes: sentinels of the bacterial outer membrane. *Biochim Biophys Acta* **1778**: 1881-1896
- Bishop RE et al (2000) Transfer of palmitate from phospholipids to lipid A in outer membranes of gram-negative bacteria. *EMBO J* **19**: 5071-5080
- Blobel G et al (1975) Transfer of proteins across membranes. I. Presence of proteolytically processed and unprocessed nascent immunoglobulin light chains on membrane-bound ribosomes of murine myeloma. *J Cell Biol* **67**: 835-851
- Blocker A et al (2003) Type III secretion systems and bacterial flagella: insights into their function from structural similarities. *Proc Natl Acad Sci U S A* **100**: 3027-3030
- Boel G et al (2004) Is 2-phosphoglycerate-dependent automodification of bacterial enolases implicated in their export? *J Mol Biol* **337**: 485-496
- Bordes P et al (2011) Insights into the extracytoplasmic stress response of *Xanthomonas campestris* pv. *campestris*: Role and regulation of sigma(E)-dependent activity. *Journal of Bacteriology* **193**: 246-264
- Boyd D et al (1990) The role of charged amino acids in the localization of secreted and membrane proteins. *Cell* **62**: 1031-1033
- Braun V (1975) Covalent lipoprotein from the outer membrane of *Escherichia coli*. *Biochim Biophys Acta* **415**: 335-377
- Busch A et al (2012) Chaperone-usher pathways: diversity and pilus assembly mechanism. *Philos Trans R Soc Lond B Biol Sci* **367**: 1112-1122
- Camacho C et al (2009) BLAST+: architecture and applications. *BMC Bioinformatics* **10**: 421
- Cargile BJ et al (2003) Gel based isoelectric focusing of peptides and the utility of isoelectric point in protein identification. *Journal of Proteome Research* **3**: 112-119
- Chandramouli K et al (2009) Proteomics: challenges, techniques and possibilities to overcome biological sample complexity. *Hum Genomics Proteomics* **2009**
- Chang YC et al (2010) Osmolyte-induced folding of an intrinsically disordered protein: folding mechanism in the absence of ligand. *Biochemistry* **49**: 5086-5096
- Chatzi KE et al (2014) SecA-mediated targeting and translocation of secretory proteins. *Biochim Biophys Acta* **1843**: 1466-1474
- Chatzi KE et al (2013) Breaking on through to the other side: protein export through the bacterial Sec system. *Biochem J* **449**: 25-37
- Chen L et al (2013) Substrate-activated conformational switch on chaperones encodes a targeting signal in type III secretion. *Cell Rep* **3**: 709-715
- Cherny I et al (2005) The formation of *Escherichia coli* curli amyloid fibrils is mediated by prion-like peptide repeats. *J Mol Biol* **352**: 245-252
- Choi-Rhee E et al (2003) The biotin carboxylase-biotin carboxyl carrier protein complex of *Escherichia coli* acetyl-CoA carboxylase. *Journal of Biological Chemistry* **278**: 30806-30812
- Choo KH et al (2008a) Flanking signal and mature peptide residues influence signal peptide cleavage. *BMC Bioinformatics* **9 Suppl 12**: S15
- Choo KH et al (2008b) Modeling *Escherichia coli* signal peptidase complex with bound substrate: determinants in the mature peptide influencing signal peptide cleavage. *BMC Bioinformatics* **9 Suppl 1**: S15
- Chou KC (2001) Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins* **43**: 246-255

- Chou KC (2005) Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics* **21**: 10-19
- Choudhury R et al (2012) Engineering RNA endonucleases with customized sequence specificities. *Nat Commun* **3**: 1147
- Collier DN et al (1989) Mutations that improve export of maltose-binding protein in SecB- cells of Escherichia coli. *J Bacteriol* **171**: 4640-4647
- Collinet B et al (2000) RseB binding to the periplasmic domain of RseA modulates the RseA:σE interaction in the cytoplasm and the availability of σE.RNA polymerase. *J Biol Chem* **275**: 33898-33904
- Conchillo-Sole O et al (2007) AGGRESCAN: a server for the prediction and evaluation of "hot spots" of aggregation in polypeptides. *BMC Bioinformatics* **8**
- Consortium TU (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res* **40**: D71-D75
- Corbin RW et al (2003) Toward a protein profile of Escherichia coli: comparison to its transcription profile. *Proc Natl Acad Sci U S A* **100**: 9232-9237
- Cornelis GR (2006) The type III secretion injectisome. *Nat Rev Microbiol* **4**: 811-825
- Covert MW et al (2004) Integrating high-throughput and computational data elucidates bacterial networks. *Nature* **429**: 92-96
- Cowles CE et al (2011) The free and bound forms of Lpp occupy distinct subcellular locations in Escherichia coli. *Mol Microbiol* **79**: 1168-1181
- Cozzarelli NR et al (1965) Growth stasis by accumulated L-alpha-glycerophosphate in Escherichia coli. *J Bacteriol* **90**: 1325-1329
- Crooks GE et al (2004) WebLogo: A sequence logo generator. *Genome Res* **14**: 1188-1190
- Dalbey RE et al (2012) Protein traffic in Gram-negative bacteria--how exported and secreted proteins find their way. *Fems Microbiol Rev* **36**: 1023-1045
- Dalbey RE et al (2014) The membrane insertase YidC. *Biochim Biophys Acta* **1843**: 1489-1496
- Daley DO et al (2005) Global topology analysis of the Escherichia coli inner membrane proteome. *Science* **308**: 1321-1323
- de Boer PA (2010) Advances in understanding E. coli cell fission. *Curr Opin Microbiol* **13**: 730-737
- de Cock H et al (2001) Identification of phospholipids as new components that assist in the in vitro trimerization of a bacterial pore protein. *European Journal of Biochemistry* **268**: 865-875
- De Leeuw E et al (1999) Molecular characterization of Escherichia coli FtsE and FtsX. *Mol Microbiol* **31**: 983-993
- de Sousa Abreu R et al (2009a) Global signatures of protein and mRNA expression levels. *Molecular BioSystems* **5**: 1512-1526
- de Sousa Abreu R et al (2009b) Global signatures of protein and mRNA expression levels. *Mol Biosyst* **5**: 1512-1526
- Dimmer EC et al (2012) The UniProt-GO Annotation database in 2011. *Nucleic Acids Res* **40**: D565-570
- Dorazi R et al (2000) Membrane topology of the N-terminus of the Escherichia coli FtsK division protein. *FEBS Letters* **478**: 13-18
- Dosztanyi Z et al (2005a) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* **21**: 3433-3434
- Dosztanyi Z et al (2005b) The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J Mol Biol* **347**: 827-839
- Douzi B et al (2012) On the path to uncover the bacterial type II secretion system. *Philos Trans R Soc Lond B Biol Sci* **367**: 1059-1072
- Dowhan W et al (2008) Chapter 1 - Functional roles of lipids in membranes. In *Biochemistry of Lipids, Lipoproteins and Membranes (Fifth Edition)*, Dennis EV, Jean EV (eds), pp 1-1. San Diego: Elsevier

- Drummelsmith J et al (2000) Translocation of group 1 capsular polysaccharide to the surface of *Escherichia coli* requires a multimeric complex in the outer membrane. *EMBO J* **19**: 57-66
- Dunker AK et al (2008) Function and structure of inherently disordered proteins. *Curr Opin Struct Biol* **18**: 756-764
- Economou A (1999) Following the leader: bacterial protein export through the Sec pathway. *Trends in Microbiology* **7**: 315-320
- Edgar R et al (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research* **30**: 207-210
- Eichacker LA et al (2004) Hiding behind Hydrophobicity. *Journal of Biological Chemistry* **279**: 50915-50922
- Fekkes P et al (1999) Zinc stabilizes the SecB binding site of SecA. *Biochemistry* **38**: 5111-5116
- Fernandez-Escamilla AM et al (2004) Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat Biotechnol* **22**: 1302-1306
- Fontaine F et al (2011) Membrane localization of small proteins in *Escherichia coli*. *J Biol Chem* **286**: 32464-32474
- Fozo EM et al (2008) Repression of small toxic protein synthesis by the Sib and OhsC small RNAs. *Molecular Microbiology* **70**: 1076-1093
- Francetic O et al (2000) Expression of the endogenous type II secretion pathway in *Escherichia coli* leads to chitinase secretion. *EMBO J* **19**: 6697-6703
- Froderberg L et al (2004) Targeting and translocation of two lipoproteins in *Escherichia coli* via the SRP/Sec/YidC pathway. *J Biol Chem* **279**: 31026-31032
- Fujiki Y et al (1982) Isolation of intracellular membranes by means of sodium carbonate treatment: application to endoplasmic reticulum. *The Journal of Cell Biology* **93**: 97-102
- Gajiwala KS et al (2000) HDEA, a periplasmic protein that supports acid resistance in pathogenic enteric bacteria. *J Mol Biol* **295**: 605-612
- Gardy JL et al (2003) PSORT-B: Improving protein subcellular localization prediction for Gram-negative bacteria. *Nucleic Acids Res* **31**: 3613-3617
- Gibeaut DM et al (1990) Separation of membranes by flotation centrifugation for in vitro synthesis of plant cell wall polysaccharides. *Protoplasma* **156**: 82-93
- Gitai Z et al (2004) An actin-like gene can determine cell polarity in bacteria. *Proceedings of the National Academy of Sciences of the United States of America* **101**: 8643-8648
- GO Consortium (2012) Gene Ontology Annotations and Resources. *Nucleic Acids Res*
- Godlewska R et al (2009) Peptidoglycan-associated lipoprotein (Pal) of Gram-negative bacteria: function, structure, role in pathogenesis and potential application in immunoprophylaxis. *FEMS Microbiol Lett* **298**: 1-11
- Goemans C et al (2013) Folding mechanisms of periplasmic proteins. *Biochim Biophys Acta*
- Goldberg T et al (2014) LocTree3 prediction of localization. *Nucleic Acids Res* **42**: W350-355
- Gonnet P et al (2004) Fine-tuning the prediction of sequences cleaved by signal peptidase II: a curated set of proven and predicted lipoproteins of *Escherichia coli* K-12. *Proteomics* **4**: 1597-1613
- Gouridis G et al (2009) Signal peptides are allosteric activators of the protein translocase. *Nature* **462**: 363-367
- Gray AN et al (2011) Unbalanced charge distribution as a determinant for dependence of a subset of *Escherichia coli* membrane proteins on the membrane insertase YidC. *Mbio* **2**
- Griffin NM et al (2011) Overcoming Key Technological Challenges in Using Mass Spectrometry for Mapping Cell Surfaces in Tissues. *Mol Cell Proteom* **10**
- Grudnik P et al (2009) Protein targeting by the signal recognition particle. *Biol Chem* **390**: 775-782

- Gupta SD et al (1991) Identification and subcellular localization of apolipoprotein N-acyltransferase in *Escherichia coli*. *FEMS Microbiol Lett* **62**: 37-41
- Hiniker A et al (2004) In vivo substrate specificity of periplasmic disulfide oxidoreductases. *J Biol Chem* **279**: 12967-12973
- Hirashima A et al (1974) Cell-free synthesis of a specific lipoprotein of the *Escherichia coli* outer membrane directed by purified messenger RNA. *Proc Natl Acad Sci U S A* **71**: 4149-4153
- Hiss JA et al (2009) Domain organization of long autotransporter signal sequences. *Bioinform Biol Insights* **3**: 189-204
- Horler RS et al (2009) EchoLOCATION: an in silico analysis of the subcellular locations of *Escherichia coli* proteins and comparison with experimentally derived locations. *Bioinformatics* **25**: 163-166
- Horth P et al (2006) Efficient Fractionation and Improved Protein Identification by Peptide OFFGEL Electrophoresis. *Mol Cel Proteom* **5**: 1968-1974
- Hu P et al (2009) Global Functional Atlas of *Escherichia coli* Encompassing Previously Uncharacterized Proteins. *PLoS Biol* **7**: e1000096
- Huang CZ et al (2006) Systematic identification of the subproteome of *Escherichia coli* cell envelope reveals the interaction network of membrane proteins and membrane-associated peripheral proteins. *J Proteome Res* **5**: 3268-3276
- Hunter S et al (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res* **37**: D211-215
- Hussain M et al (1980) Accumulation of glyceride-containing precursor of the outer membrane lipoprotein in the cytoplasmic membrane of *Escherichia coli* treated with globomycin. *J Biol Chem* **255**: 3707-3712
- Huthmacher C et al (2008) A computational analysis of protein interactions in metabolic networks reveals novel enzyme pairs potentially involved in metabolic channeling. *Journal of Theoretical Biology* **252**: 456-464
- Ieva R et al (2008) Incorporation of a polypeptide segment into the beta-domain pore during the assembly of a bacterial autotransporter. *Mol Microbiol* **67**: 188-201
- Imai K et al (2008) SOSUI-GramN: high performance prediction for sub-cellular localization of proteins in gram-negative bacteria. *Bioinformatics* **2**: 417-421
- Hayden JD et al (2008) The extracytoplasmic stress factor, sigma(E), is required to maintain cell envelope integrity in *Escherichia coli*. *Plos One* **3**: e1573
- Headlam MJ et al (2003) The F-G loop region of cytochrome P450<sub>scc</sub> (CYP11A1) interacts with the phospholipid membrane. *Biochim Biophys Acta* **1617**: 96-108
- Hegde RS et al (2006) The surprising complexity of signal sequences. *Trends Biochem Sci* **31**: 563-571
- Hemm MR et al (2008a) Small membrane proteins found by comparative genomics and ribosome binding site models. *Mol Microbiol* **70**: 1487-1501
- Hemm MR et al (2008b) Small membrane proteins found by comparative genomics and ribosome binding site models. *Mol Microbiol* **70**: 1487-1501
- Henderson IR et al (2000) Autotransporter proteins, evolution and redefining protein secretion. *Trends Microbiol* **8**: 529-532
- Handa Y et al (2011) YaeJ is a novel ribosome-associated protein in *Escherichia coli* that can hydrolyze peptidyl-tRNA on stalled ribosomes. *Nucleic Acids Res* **39**: 1739-1748
- Hantke K et al (1973) Covalent binding of lipid to protein. Diglyceride and amide-linked fatty acid at the N-terminal end of the murein-lipoprotein of the *Escherichia coli* outer membrane. *Eur J Biochem* **34**: 284-296
- Hayano T et al (1991) Two distinct forms of peptidylprolyl-cis-trans-isomerase are expressed separately in periplasmic and cytoplasmic compartments of *Escherichia coli* cells. *Biochemistry* **30**: 3041-3048



- Ingerson-Mahar M et al (2010) The metabolic enzyme CTP synthase forms cytoskeletal filaments. *Nat Cell Biol* **12**: 739-746
- Inouye M et al (1974) Discussion paper: biosynthesis and assembly of a structural lipoprotein in the envelope of *Escherichia coli*. *Ann N Y Acad Sci* **235**: 83-90
- Ishihama A (2012) Prokaryotic genome regulation: a revolutionary paradigm. *Proc Jpn Acad Ser B Phys Biol Sci* **88**: 485-508
- Ishihama Y et al (2005) Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Mol Cell Proteomics* **4**: 1265 - 1272
- Ishihama Y et al (2008) Protein abundance profiling of the *Escherichia coli* cytosol. *BMC Genomics* **9**: 102
- Iwasaki M et al (2010) One-Dimensional Capillary Liquid Chromatographic Separation Coupled with Tandem Mass Spectrometry Unveils the *Escherichia coli* Proteome on a Microarray Scale. *Analytical Chemistry* **82**: 2616-2620
- Jeffery CJ (1999) Moonlighting proteins. *Trends in Biochemical Sciences* **24**: 8-11
- Johnson M et al (2008) NCBI BLAST: a better web interface. *Nucleic Acids Res* **36**: W5-W9
- Juncker AS et al (2003) Prediction of lipoprotein signal peptides in Gram-negative bacteria. *Protein Sci* **12**: 1652-1662
- Kadonaga JT et al (1984) The role of the beta-lactamase signal sequence in the secretion of proteins by *Escherichia coli*. *J Biol Chem* **259**: 2149-2154
- Kajava AV et al (2000) The net charge of the first 18 residues of the mature sequence affects protein translocation across the cytoplasmic membrane of gram-negative bacteria. *J Bacteriol* **182**: 2163-2169
- Kajava AV et al (2001) The net charge of the first 18 residues of the mature sequence affects protein translocation across the cytoplasmic membrane of gram-negative bacteria. *J Bacteriol* **182**: 2163-2169
- Kall L et al (2004) A combined transmembrane topology and signal peptide prediction method. *J Mol Biol* **338**: 1027-1036
- Kall L et al (2007) Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server. *Nucleic Acids Res* **35**: W429-432
- Käll L et al (2004) A Combined Transmembrane Topology and Signal Peptide Prediction Method. *Journal of Molecular Biology* **338**: 1027-1036
- Kamio Y et al (1976) Outer membrane of *Salmonella typhimurium*: accessibility of phospholipid head groups to phospholipase c and cyanogen bromide activated dextran in the external medium. *Biochemistry* **15**: 2561-2570
- Karamyshev AL et al (1998) Processing of *Escherichia coli* alkaline phosphatase: role of the primary structure of the signal peptide cleavage region. *J Mol Biol* **277**: 859-870
- Karp PD et al (2007) Multidimensional annotation of the *Escherichia coli* K-12 genome. *Nucleic Acids Research* **35**: 7577-7590
- Kato M et al (1992) In vitro translocation of secretory proteins possessing no charges at the mature domain takes place efficiently in a protonmotive force-dependent manner. *J Biol Chem* **267**: 413-418
- Kerrien S et al (2012) The IntAct molecular interaction database in 2012. *Nucleic Acids Res* **40**: D841-846
- Keseler IM et al (2009) EcoCyc: a comprehensive view of *Escherichia coli* biology. *Nucleic Acids Res* **37**: D464-470
- Keseler IM et al (2011) EcoCyc: a comprehensive database of *Escherichia coli* biology. *Nucleic Acids Res* **39**: D583-D590
- Keseler IM et al (2013) EcoCyc: fusing model organism databases with systems biology. *Nucleic Acids Res* **41**: D605-612
- Khemic V et al (2008) The RNase E of *Escherichia coli* is a membrane-binding protein. *Mol Microbiol* **70**: 799-813
- Khisty VJ et al (1995) Mapping of the binding frame for the chaperone SecB within a natural ligand, galactose-binding protein. *J Biol Chem* **270**: 25920-25927

- Khokhlova OV et al (2003) [Interaction of SecB and SecA with the N-terminal region of mature alkaline phosphatase on its secretion in *Escherichia coli*]. *Mol Biol (Mosk)* **37**: 712-718
- Kim J et al (2000) SecB dependence of an exported protein is a continuum influenced by the characteristics of the signal peptide or early mature region. *J Bacteriol* **182**: 4108-4112
- King GF et al (1999) The dimerization and topological specificity functions of MinE reside in a structurally autonomous C-terminal domain. *Mol Microbiol* **31**: 1161-1169
- Kiraga J et al (2007) The relationships between the isoelectric point and: length of proteins, taxonomy and ecology of organisms. *BMC Genomics* **8**: 163
- Kishii R et al (2007) Structural and functional studies of the HAMP domain of EnvZ, an osmosensing transmembrane histidine kinase in *Escherichia coli*. *J Biol Chem* **282**: 26401-26408
- Klein C et al (2005) The membrane proteome of *Halobacterium salinarum*. *PROTEOMICS* **5**: 180-197
- Kobayashi R et al (2007) *Escherichia coli* phage-shock protein A (PspA) binds to membrane phospholipids and repairs proton leakage of the damaged membranes. *Mol Microbiol* **66**: 100-109
- Koebnik R et al (2000) Structure and function of bacterial outer membrane proteins: barrels in a nutshell. *Mol Microbiol* **37**: 239-253
- Koronakis V et al (2000) Crystal structure of the bacterial membrane protein TolC central to multidrug efflux and protein export. *Nature* **405**: 914-919
- Korotkov KV et al (2012) The type II secretion system: biogenesis, molecular architecture and mechanism. *Nat Rev Microbiol* **10**: 336-351
- Kramer RA et al (2002) Lipopolysaccharide regions involved in the activation of *Escherichia coli* outer membrane protease OmpT. *Eur J Biochem* **269**: 1746-1752
- Krehenbrink M et al (2011) The superoxide dismutase SodA is targeted to the periplasm in a SecA-dependent manner by a novel mechanism. *Mol Microbiol* **82**: 164-179
- Kreibich G et al (1974) Selective release of content from microsomal vesicles without membrane disassembly. *The Journal of Cell Biology* **61**: 789-807
- Krogh A et al (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* **305**: 567-580
- Krug K et al (2013) Deep Coverage of the *Escherichia coli* Proteome Enables the Assessment of False Discovery Rates in Simple Proteogenomic Experiments. *Molecular & Cellular Proteomics : MCP* **12**: 3420-3430
- Kruse T et al (2006) Actin homolog MreB and RNA polymerase interact and are both required for chromosome segregation in *Escherichia coli*. *Gene Dev* **20**: 113-124
- Kudva R et al (2013) Protein translocation across the inner membrane of Gram-negative bacteria: the Sec and Tat dependent protein transport pathways. *Res Microbiol* **164**: 505-534
- Kyte J et al (1982) A simple method for displaying the hydropathic character of a protein. *J Mol Biol* **157**: 105-132
- Laforet GA et al (1989) Signal peptide subsegments are not always functionally interchangeable. M13 procoat hydrophobic core fails to transport alkaline phosphatase in *Escherichia coli*. *J Biol Chem* **264**: 14478-14485
- Lasserre JP et al (2006) A complexomic study of *Escherichia coli* using two-dimensional blue native/SDS polyacrylamide gel electrophoresis. *Electrophoresis* **27**: 3306-3321
- Lee CR et al (2007a) *Escherichia coli* enzyme IIA<sub>Ntr</sub> regulates the K<sup>+</sup> transporter TrkA. *Proc Natl Acad Sci U S A* **104**: 4124-4129
- Lee EY et al (2007b) Global proteomic profiling of native outer membrane vesicles derived from *Escherichia coli*. *Proteomics* **7**: 3143-3153
- Lee PA et al (2006) The bacterial twin-arginine translocation pathway. *Annu Rev Microbiol* **60**: 373-395

- Letain TE et al (1997) TonB protein appears to transduce energy by shuttling between the cytoplasmic membrane and the outer membrane in *Escherichia coli* (vol 24, pg 271, 1997). *Molecular Microbiology* **25**: 617-617
- Li H et al (2012) Alterations of protein complexes and pathways in genetic information flow and response to stimulus contribute to *Escherichia coli* resistance to balofloxacin. *Mol Biosyst* **8**: 2303-2311
- Li P et al (1988) Alteration of the amino terminus of the mature sequence of a periplasmic protein can severely affect protein export in *Escherichia coli*. *Proc Natl Acad Sci U S A* **85**: 7685-7689
- Li Y et al (2010) *Escherichia coli* condensin MukB stimulates topoisomerase IV activity by a direct physical interaction. *Proc Natl Acad Sci U S A* **107**: 18832-18837
- Liechti G et al (2012) Outer membrane biogenesis in *Escherichia coli*, *Neisseria meningitidis*, and *Helicobacter pylori*: paradigm deviations in *H. pylori*. *Front Cell Infect Microbiol* **2**: 29
- Lill R et al (1990) The Atpase Activity of SecA Is Regulated by Acidic Phospholipids, SecY, and the Leader and Mature Domains of Precursor Proteins. *Cell* **60**: 271-280
- Link AJ et al (1997) Comparing the predicted and observed properties of proteins encoded in the genome of *Escherichia coli* K-12. *Electrophoresis* **18**: 1259-1313
- Lomize MA et al (2012) OPM database and PPM web server: resources for positioning of proteins in membranes. *Nucleic Acids Res* **40**: D370-376
- Lopez-Campistrous A et al (2005) Localization, annotation, and comparison of the *Escherichia coli* K-12 proteome under two states of growth. *Mol Cell Proteomics* **4**: 1205-1209
- Lopez-Maury L et al (2008) Tuning gene expression to changing environments: from rapid responses to evolutionary adaptation. *Nat Rev Genet* **9**: 583-593
- Lopian L et al (2003) The BglF sensor recruits the BglG transcription regulator to the membrane and releases it on stimulation. *Proceedings of the National Academy of Sciences* **100**: 7099-7104
- Lu F et al (2013) Membrane association via an amino-terminal amphipathic helix is required for the cellular organization and function of RNase II. *J Biol Chem* **288**: 7241-7251
- Lu P et al (2007) Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat Biotech* **25**: 117-124
- Lutkenhaus J et al (1997) Bacterial cell division and the Z ring. *Annual Review of Biochemistry* **66**: 93-116
- MacIntyre S et al (1990) Export incompatibility of N-terminal basic residues in a mature polypeptide of *Escherichia coli* can be alleviated by optimising the signal peptide. *Mol Gen Genet* **221**: 466-474
- Maddalo G et al (2011) Systematic analysis of native membrane protein complexes in *Escherichia coli*. *J Proteome Res* **10**: 1848-1859
- Makarov A (2000) Electrostatic Axially Harmonic Orbital Trapping: A High-Performance Technique of Mass Analysis. *Analytical Chemistry* **72**: 1156-1162
- Manabe T et al (2013) Flagella proteins contribute to the production of outer membrane vesicles from *Escherichia coli* W3110. *Biochem Biophys Res Commun* **441**: 151-156
- Manadas B et al (2009) Comparative analysis of OFFGel, strong cation exchange with pH gradient, and RP at high pH for first-dimensional separation of peptides from a membrane-enriched protein fraction. *PROTEOMICS* **9**: 1-5
- Martinez-Hackert E et al (2009) Promiscuous substrate recognition in folding and assembly activities of the trigger factor chaperone. *Cell* **138**: 923-934
- Masuda T et al (2009) Unbiased quantitation of *Escherichia coli* membrane proteome using phase transfer surfactants. *Mol Cell Proteomics* **8**: 2770-2777
- Meile J et al (2006) Systematic localisation of proteins fused to the green fluorescent protein in *Bacillus subtilis*: identification of new proteins at the DNA replication factory. *Proteomics* **6**: 2135-2146
- Michel PE et al (2003) Protein fractionation in a multicompartiment device using Off-Gel™ isoelectric focusing. *Electrophoresis* **24**: 3-11

- Miyamoto S et al (2007) Diverse effects of phospholipids on lipoprotein sorting and ATP hydrolysis by the ABC transporter LolCDE complex. *Biochim Biophys Acta* **1768**: 1848-1854
- Moffatt JH et al (2010) Colistin resistance in *Acinetobacter baumannii* is mediated by complete loss of lipopolysaccharide production. *Antimicrob Agents Chemother* **54**: 4971-4977
- Moreda-Piñeiro A et al (2014) A review on preparative and semi-preparative offgel electrophoresis for multidimensional protein/peptide assessment. *Analytica Chimica Acta* **836**: 1-17
- Moreno F et al (1980) A signal sequence is not sufficient to lead beta-galactosidase out of the cytoplasm. *Nature* **286**: 356-359
- Mota LJ et al (2005) The bacterial injection kit: type III secretion systems. *Ann Med* **37**: 234-249
- Murashko ON et al (2012) Membrane binding of *Escherichia coli* RNase E catalytic domain stabilizes protein structure and increases RNA substrate affinity. *Proc Natl Acad Sci U S A* **109**: 7019-7024
- Nakayama H et al (2012) Lipoproteins in bacteria: structures and biosynthetic pathways. *FEBS J* **279**: 4247-4268
- Navarro S et al (2014) Selection against toxic aggregation-prone protein sequences in bacteria. *Biochim Biophys Acta* **1843**: 866-874
- Neilson KA et al (2011) Less label, more free: Approaches in label-free quantitative mass spectrometry. *PROTEOMICS* **11**: 535-553
- Nenninger AA et al (2009) Localized and efficient curli nucleation requires the chaperone-like amyloid assembly protein CsgF. *Proc Natl Acad Sci U S A* **106**: 900-905
- Nesmeyanova MA et al (1997) Positively charged lysine at the N-terminus of the signal peptide of the *Escherichia coli* alkaline phosphatase provides the secretion efficiency and is involved in the interaction with anionic phospholipids. *FEBS Lett* **403**: 203-207
- Neumann-Haefelin C et al (2000) SRP-dependent co-translational targeting and SecA-dependent translocation analyzed as individual steps in the export of a bacterial protein. *EMBO J* **19**: 6419-6426
- Neumann M et al (2009) A periplasmic aldehyde oxidoreductase represents the first molybdopterin cytosine dinucleotide cofactor containing molybdo-flavoenzyme from *Escherichia coli*. *FEBS J* **276**: 2762-2774
- Nilsson I et al (1992) A signal peptide with a proline next to the cleavage site inhibits leader peptidase when present in a sec-independent protein. *FEBS Lett* **299**: 243-246
- Niwa T et al (2009) Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of *Escherichia coli* proteins. *Proc Natl Acad Sci U S A* **106**: 4201-4206
- O'Neill J et al (2011) Role of the MotB linker in the assembly and activation of the bacterial flagellar motor. *Acta Crystallogr D Biol Crystallogr* **67**: 1009-1016
- Oberto J et al (2009) The HU Regulon Is Composed of Genes Responding to Anaerobiosis, Acid Stress, High Osmolarity and SOS Induction *PLoS One* **4**: e4367
- Ochman H et al (2006) The nature and dynamics of bacterial genomes. *Science* **311**: 1730-1733
- Ohlendieck K (2003) Extraction of membrane proteins. In *Protein Purification Protocols*, Cutler P (ed), Vol. 244, 2nd edn, pp 283-290. Totowa, NJ: Humana Press
- Okuda S et al (2011) Lipoprotein sorting in bacteria. *Annu Rev Microbiol* **65**: 239-259
- Orfanoudaki G et al (2014) Proteome-wide subcellular topologies of *E. coli* polypeptides database (STEPdb). *Mol Cell Proteomics* **13**: 3674-3687
- Osborn MJ et al (1972) Mechanism of assembly of the outer membrane of *Salmonella typhimurium*. Site of synthesis of lipopolysaccharide. *J Biol Chem* **247**: 3973-3986
- Ostrovsky de Spicer P et al (1993) PutA protein, a membrane-associated flavin dehydrogenase, acts as a redox-dependent transcriptional regulator. *Proceedings of the National Academy of Sciences* **90**: 4295-4298
- Paetzel M et al (2002) Signal peptidases. *Chem Rev* **102**: 4549-4580

- Pan JY et al (2010) Complexome of Escherichia coli envelope proteins under normal physiological conditions. *J Proteome Res* **9**: 3730-3740
- Papanastasiou M et al (2013 ) The *E. coli* Inner Membrane Peripherome. *Mol Cel Proteom* **12**: 599-610
- Papanastasiou M et al (2013) The Escherichia coli peripheral inner membrane proteome. *Mol Cell Proteomics* **12**: 599-610
- Papanikou E et al (2007) Bacterial protein secretion through the translocase nanomachine. *Nat Rev Microbiol* **5**: 839-851
- Paramasivam N et al (2011) ClubSub-P: Cluster-Based Subcellular Localization Prediction for Gram-Negative Bacteria and Archaea. *Front Microbiol* **2**: 218
- Parlitz R et al (2007) Escherichia coli signal recognition particle receptor FtsY contains an essential and autonomous membrane-binding amphipathic helix. *J Biol Chem* **282**: 32176-32184
- Patel R et al (2014) Protein transport by the bacterial Tat pathway. *Biochim Biophys Acta*
- Patten CL et al (2004a) Microarray analysis of RpoS-mediated gene expression in Escherichia coli K-12. *Mol Genet Genomics* **272**: 580-591
- Patten CL et al (2004b) Microarray analysis of RpoS-mediated gene expression in Escherichia coli K-12. *Mol Genet Genomics* **272**: 580-591
- Perez-Bercoff A et al (2011) Patterns of indirect protein interactions suggest a spatial organization to metabolism. *Molecular BioSystems* **7**: 3056-3064
- Petersen TN et al (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Methods* **8**: 785-786
- Petrushenko ZM et al (2006) Antagonistic interactions of kleisins and DNA with bacterial Condensin MukB. *J Biol Chem* **281**: 34208-34217
- Pettijohn DE (1988) Histone-like proteins and bacterial chromosome structure. *J Biol Chem* **263**: 12793-12796
- Pham TV et al (2010) On the beta-binomial model for analysis of spectral count data in label-free tandem mass spectrometry-based proteomics. *Bioinformatics* **26**: 363-369
- Phoenix DA et al (1990) pH-induced insertion of the amphiphilic alpha-helical anchor of Escherichia coli penicillin-binding protein 5. *Eur J Biochem* **190**: 365-369
- Pieper R et al (2009) Integral and peripheral association of proteins and protein complexes with *Yersinia pestis* inner and outer membranes. *Proteome Sci* **7**: 16
- Polissi A et al (2014) The lipopolysaccharide export pathway in Escherichia coli: structure, organization and regulated assembly of the Lpt machinery. *Mar Drugs* **12**: 1023-1042
- Prehna G et al (2012) A protein export pathway involving Escherichia coli porins. *Structure* **20**: 1154-1166
- Puziss JW et al (1989) Analysis of mutational alterations in the hydrophilic segment of the maltose-binding protein signal peptide. *J Bacteriol* **171**: 2303-2311
- Raffaelli N et al (1999) The *Escherichia coli* NadR regulator Is endowed with nicotinamide mononucleotide adenylyltransferase activity. *Journal of Bacteriology* **181**: 5509-5511
- Rajapandi T et al (1991) The first gene in the Escherichia coli secA operon, gene X, encodes a nonessential secretory protein. *J Bacteriol* **173**: 7092-7097
- Randall LL et al (2002) SecB, one small chaperone in the complex milieu of the cell. *Cell Mol Life Sci* **59**: 1617-1623
- Rauschmeier M et al (2014) New Insights into the Interplay Between the Lysine Transporter LysP and the pH Sensor CadC in Escherichia Coil. *Journal of Molecular Biology* **426**: 215-229
- Reyes-Lamothe R (2012) Use of fluorescently tagged SSB proteins in in vivo localization experiments. *Methods Mol Biol* **922**: 245-253
- Ried G et al (1990) Role of lipopolysaccharide in assembly of Escherichia coli outer membrane proteins OmpA, OmpC, and OmpF. *J Bacteriol* **172**: 6048-6053

- Riley M et al (2005) *Escherichia coli* K-12: a cooperatively developed annotation snapshot—2005. *Nucleic Acids Res* **34**: 1-9
- Robinson LS et al (2006) Secretion of curli fibre subunits is mediated by the outer membrane-localized CsgG protein. *Mol Microbiol* **59**: 870-881
- Rustici G et al (2013) ArrayExpress update—trends in database growth and links to data analysis tools. *Nucleic Acids Research* **41**: D987-D990
- Sabatini DD et al (1971) Ribosome-membrane interaction: Structural aspects and functional implications. *Adv Cytopharmacol* **1**: 119-129
- Safdar A et al (2010) Drug-induced nephrotoxicity caused by amphotericin B lipid complex and liposomal amphotericin B: a review and meta-analysis. *Medicine (Baltimore)* **89**: 236-244
- Salje J et al (2011) Direct membrane binding by bacterial actin MreB. *Mol Cell* **43**: 478-487
- Sanders AN et al (2013) Phenotypic analysis of *Escherichia coli* mutants lacking L,D-transpeptidases. *Microbiology* **159**: 1842-1852
- Sankaran K et al (1994) Lipid modification of bacterial prolipoprotein. Transfer of diacylglycerol moiety from phosphatidylglycerol. *J Biol Chem* **269**: 19701-19706
- Sapay N et al (2006) Prediction of amphipathic in-plane membrane anchors in monotopic proteins using a SVM classifier. *BMC Bioinformatics* **7**: 255
- Schertzer JW et al (2013) Bacterial outer membrane vesicles in trafficking, communication and the host-pathogen interaction. *J Mol Microbiol Biotechnol* **23**: 118-130
- Schryvers A et al (1978) Chemical and functional properties of the native and reconstituted forms of the membrane-bound, aerobic glycerol-3-phosphate dehydrogenase of *Escherichia coli*. *J Biol Chem* **253**: 783-788
- Schulze RJ et al (2014) Membrane protein insertion and proton-motive-force-dependent secretion through the bacterial holo-translocon SecYEG-SecDF-YajC-YidC. *Proc Natl Acad Sci U S A* **111**: 4844-4849
- Schwartz R et al (2001) Whole Proteome pI Values Correlate with Subcellular Localizations of Proteins for Organisms within the Three Domains of Life. *Genome Research* **11**: 703-709
- Sekizawa J et al (1977) Precursors of major outer membrane proteins of *Escherichia coli*. *Biochem Biophys Res Commun* **77**: 1126-1133
- Selkrig J et al (2012) Discovery of an archetypal protein transport system in bacterial outer membranes. *Nat Struct Mol Biol* **19**: 506-510, S501
- Shih YL et al (2011) The N-terminal amphipathic helix of the topological specificity factor MinE is associated with shaping membrane curvature. *PLoS One* **6**: e21425
- Shiomi D et al (2008a) Compensation for the loss of the conserved membrane targeting sequence of FtsA provides new insights into its function. *Mol Microbiol* **67**: 558-569
- Shiomi D et al (2008b) Determination of bacterial rod shape by a novel cytoskeletal membrane protein. *EMBO J* **27**: 3081-3091
- Silhavy TJ et al (2010) The bacterial cell envelope. *Cold Spring Harb Perspect Biol* **2**: a000414
- Singer SJ et al (1972) The fluid mosaic model of the structure of cell membranes. *Science* **175**: 720-731
- Solov'eva TF et al (2012) Biogenesis of beta-barrel integral proteins of bacterial outer membrane. *Biochemistry (Mosc)* **77**: 1221-1236
- Soufo H et al (2010) Bacterial translation elongation factor EF-Tu interacts and colocalizes with actin-like MreB protein. *Proceedings of the National Academy of Sciences* **107**: 3163-3168
- Speers AE et al (2007) Proteomics of Integral Membrane Proteins: Theory and Application. *Chemical Reviews* **107**: 3687-3714
- Spelbrink RE et al (2005) Detection and identification of stable oligomeric protein complexes in *Escherichia coli* inner membranes: a proteomics approach. *J Biol Chem* **280**: 28742-28748

- Statnikov A et al (2005) A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics* **21**: 631-643
- Steck TL (1974) The organization of proteins in the human red blood cell membrane. *The Journal of Cell Biology* **62**: 1-19
- Stenberg F et al (2005) Protein complexes of the Escherichia coli cell envelope. *J Biol Chem* **280**: 34409-34419
- Stymest KH et al (2008) The periplasmic peptidyl prolyl cis-trans isomerases PpiD and SurA have partially overlapping substrate specificities. *FEBS J* **275**: 3470-3479
- Sugimura K et al (1988) Purification, characterization, and primary structure of Escherichia coli protease VII with specificity for paired basic residues: identity of protease VII and OmpT. *J Bacteriol* **170**: 5625-5632
- Summers RG et al (1989) A conservative amino acid substitution, arginine for lysine, abolishes export of a hybrid protein in Escherichia coli. Implications for the mechanism of protein secretion. *J Biol Chem* **264**: 20082-20088
- Sung MT et al (2009) Crystal structure of the membrane-bound bifunctional transglycosylase PBP1b from Escherichia coli. *Proc Natl Acad Sci U S A* **106**: 8824-8829
- Tamaki S et al (1971) Role of lipopolysaccharides in antibiotic resistance and bacteriophage adsorption of Escherichia coli K-12. *J Bacteriol* **105**: 968-975
- Taniguchi Y et al (2010) Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science* **329**: 533-538
- Tian P et al (2009) Identification of a post-targeting step required for efficient cotranslational translocation of proteins across the Escherichia coli inner membrane. *J Biol Chem* **284**: 11396-11404
- Tomba P (2002) Intrinsically unstructured proteins. *Trends in Biochemical Sciences* **27**: 527-533
- Topping TB et al (1994) Determination of the binding frame within a physiological ligand for the chaperone SecB. *Protein Sci* **3**: 730-736
- Tullman-Ercek D et al (2007) Export pathway selectivity of Escherichia coli twin arginine translocation signal peptides. *J Biol Chem* **282**: 8309-8316
- Tyhach RJ et al (1979) Increased synthesis of phosphatidylserine decarboxylase in a strain of Escherichia coli bearing a hybrid plasmid. Altered association of enzyme with the membrane. *J Biol Chem* **254**: 627-633
- van den Ent F et al (2010) Bacterial actin MreB assembles in complex with cell shape protein RodZ. *EMBO J* **29**: 1081-1090
- Van Gerven N et al (2011) Pili and flagella biology, structure, and biotechnological applications. *Prog Mol Biol Transl Sci* **103**: 21-72
- Veit A et al (2007) Global gene expression analysis of glucose overflow metabolism in Escherichia coli and reduction of aerobic acetate formation. *Appl Microbiol Biotechnol* **74**: 406-421
- Vijayendran C et al (2007) The plasticity of global proteome and genome expression analyzed in closely related W3110 and MG1655 strains of a well-studied model organism, Escherichia coli-K12. *Journal of Biotechnology* **128**: 747-761
- Villegas JM et al (2011) Amphipathic C-terminal region of Escherichia coli NADH dehydrogenase-2 mediates membrane localization. *Arch Biochem Biophys* **505**: 155-159
- Vogel C et al (2012) Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat Rev Genet* **13**: 227-232
- Vollmer W (2008) Structural variation in the glycan strands of bacterial peptidoglycan. *Fems Microbiol Rev* **32**: 287-306
- von Heijne G (1983) Patterns of amino acids near signal-sequence cleavage sites. *Eur J Biochem* **133**: 17-21
- von Loeffelholz O et al (2013) Structural basis of signal sequence surveillance and selection by the SRP-FtsY complex. *Nat Struct Mol Biol* **20**: 604-610
- Wagner S et al (2006) Rationalizing membrane protein overexpression. *Trends in Biotechnology* **24**: 364-371

- Walz AC et al (2002) Aerobic sn-glycerol-3-phosphate dehydrogenase from *Escherichia coli* binds to the cytoplasmic membrane through an amphipathic alpha-helix. *Biochem J* **365**: 471-479
- Wang L et al (2005a) luxS-dependent gene regulation in *Escherichia coli* K-12 revealed by genomic expression profiling. *J Bacteriol* **187**: 8350-8360
- Wang L et al (2005b) luxS-Dependent Gene Regulation in *Escherichia coli* K-12 Revealed by Genomic Expression Profiling. *Journal of Bacteriology* **187**: 8350-8360
- Wang YF et al (1997) Channel specificity: structural basis for sugar discrimination and differential flux rates in maltoporin. *J Mol Biol* **272**: 56-63
- Weiner JH et al (2008) Proteome of the *Escherichia coli* envelope and technological challenges in membrane proteome analysis. *Biochim Biophys Acta* **1778**: 1698-1713
- Wilkins MR et al (1999) Protein identification and analysis tools in the ExpASY server. *Methods Mol Biol* **112**: 531-552
- Wimley WC (2003) The versatile beta-barrel membrane protein. *Curr Opin Struct Biol* **13**: 404-411
- Wisniewski JR et al (2009a) Combination of FASP and StageTip-Based Fractionation Allows In-Depth Analysis of the Hippocampal Membrane Proteome. *Journal of Proteome Research* **8**: 5674-5678
- Wisniewski JR et al (2009b) Universal sample preparation method for proteome analysis. *Nat Meth* **6**: 359-362
- Wissenbach U et al (1995) A third periplasmic transport system for L-arginine in *Escherichia coli*: molecular characterization of the artPIQMJ genes, arginine binding and transport. *Mol Microbiol* **17**: 675-686
- Wolff S et al (2008) Complementary Analysis of the Vegetative Membrane Proteome of the Human Pathogen *Staphylococcus aureus*. *Mol Cel Proteom* **7**: 1460-1468
- Wu S et al (2006) Multi-modality of pl distribution in whole proteome. *PROTEOMICS* **6**: 449-455
- Yates JR (1998) Mass spectrometry and the age of the proteome. *Journal of Mass Spectrometry* **33**: 1-19
- Yeats C et al (2003) The BON domain: a putative membrane-binding domain. *Trends Biochem Sci* **28**: 352-355
- Yoon SH et al (2012) Comparative multi-omics systems analysis of *Escherichia coli* strains B and K-12. *Genome Biol* **13**: R37
- Yu F et al (1984) Nucleotide sequence of the *lspA* gene, the structural gene for lipoprotein signal peptidase of *Escherichia coli*. *FEBS Lett* **173**: 264-268
- Yu L et al (2010) SecretP: identifying bacterial secreted proteins by fusing new features into Chou's pseudo-amino acid composition. *J Theor Biol* **267**: 1-6
- Zhang G et al (2013) On the essentiality of lipopolysaccharide to Gram-negative bacteria. *Curr Opin Microbiol* **16**: 779-785
- Zhang R et al (2004) DEG: a database of essential genes. *Nucleic Acids Res* **32**: D271-D272
- Zybailov B et al (2006) Statistical analysis of membrane proteome expression changes in *Saccharomyces cerevisiae*. *J Proteome Res* **5**: 2339-2347







## **ΠΑΡΑΡΤΗΜΑΤΑ**



## ΠΑΡΑΡΤΗΜΑ Α

### Α1 Αναγνώριση προτύπων

Ένα από τα θεμελιώδη προβλήματα της θεωρίας μάθησης είναι το ακόλουθο: ας υποθέσουμε ότι έχουμε δύο κλάσεις από αντικείμενα, στη συνέχεια δεδομένου ότι έχουμε ένα καινούργιο αντικείμενο πρέπει να αποφασίσουμε σε ποια από τις δύο κλάσεις ανήκει. Το πρόβλημα αυτό μπορεί να διατυπωθεί ως εξής:

Ας υποθέσουμε ότι έχουμε ένα σύνολο από εμπειρικά δεδομένα ή αλλιώς **πρότυπα**:

$$(x_1, y_1), \dots, (x_m, y_m) \in X \times \{\pm 1\}$$

όπου το  $X$  είναι ένα μη κενό σύνολο από το οποίο επιλέγονται τα **πρότυπα**  $x_i$  (ή αλλιώς δείγματα, παρατηρήσεις, δεδομένα εκπαίδευσης), το οποίο ονομάζεται και χώρος ή πεδίο εισόδου, και τα  $y_i$  ονομάζονται **στόχοι** (ή έξοδοι). Στην απλή αυτή περίπτωση υπάρχουν μόνο δύο κλάσεις όπου για λόγους μαθηματικής ευκολίας τις αντιστοιχίζουμε στις τιμές -1 και 1. Η εκφυλισμένη αυτή περίπτωση ταξινόμησης ονομάζεται **δυναμική ταξινόμηση**.

Το πρόβλημα λοιπόν της ταξινόμησης είναι ότι δεδομένου ότι έχουμε στην διάθεση μας ένα σύνολο από πρότυπα σημεία, να κατατάξουμε ένα καινούργιο και άγνωστο δεδομένο  $x$  (να προβλέψουμε δηλαδή την αντίστοιχη τιμή  $y \in \{\pm 1\}$ ). Κατά συνέπεια θέλουμε να διαλέξουμε ένα  $y$  ώστε  $(x, y)$  να είναι **όμοιο** με τα πρότυπα (ή δεδομένα εκπαίδευσης). Στο πρόβλημα της δυναμικής ταξινόμησης ο στόχος είναι να εκτιμήσουμε μια συνάρτηση  $f$  η οποία:

$$f : X \rightarrow \{\pm 1\}$$

Η συνάρτηση αυτή ονομάζεται **συνάρτηση απόφασης** (decision function).

### Α2 Μέτρο ομοιότητας δεδομένων

Για να προβλέψουμε την κλάση για το άγνωστο δεδομένο  $x$  θα πρέπει να ορίσουμε ένα μέτρο ομοιότητας μεταξύ του άγνωστου δεδομένου  $x$  και των προτύπων σημείων.

Ένα μέτρο ομοιότητας θα μπορούσε να έχει την μορφή:

$$k : X \times X \rightarrow \mathfrak{R}$$

$$(x, x') \rightarrow k(x, x')$$

δηλαδή μία συνάρτηση η οποία δεδομένου δύο δειγμάτων  $x$  και  $x'$  επιστρέφει ένα πραγματικό αριθμό που αντιστοιχεί στο μέτρο ομοιότητας μεταξύ των δύο δειγμάτων. Η συνάρτηση  $k$  είναι μία συμμετρική συνάρτηση,  $k(x, x') = k(x', x) \forall x, x' \in X$  και ονομάζεται **συνάρτηση πυρήνα (kernel function)**.

Ένα μέτρο ομοιότητας αποτελεί το εσωτερικό γινόμενο (dot product). Δεδομένου δύο διανυσμάτων  $x$  και  $x' \in \mathfrak{R}^N$  το εσωτερικό τους γινόμενο ορίζεται ως εξής:

$$\langle x, x' \rangle := \sum_{i=1}^N [x]_i [x']_i \text{ όπου το } [x]_i \text{ συμβολίζει το } i\text{-στο στοιχείο του διανύσματος } x.$$

Η γεωμετρική απεικόνιση του εσωτερικού γινομένου διανυσμάτων είναι ότι υπολογίζει το συνημίτονο της γωνίας μεταξύ των διανυσμάτων  $x$  και  $x'$  δεδομένου ότι τα διανύσματα είναι κανονικοποιημένα ώστε να έχουν μοναδιαίο μήκος.

$$\|x\| = \sqrt{\langle x, x \rangle}$$

Κατά παρόμοιο τρόπο η απόσταση δύο διανυσμάτων υπολογίζεται ως το μήκος του διανύσματος της διαφοράς τους.

### A3 Χώρος χαρακτηριστικών

Για να μπορέσουμε να χρησιμοποιήσουμε το εσωτερικό γινόμενο ως μέτρο ομοιότητας μεταξύ των δειγμάτων μας θα πρέπει να αναπαραστήσουμε τα δείγματα μας ως διανύσματα σε ένα πεδίο εσωτερικών γινομένων  $H$ .

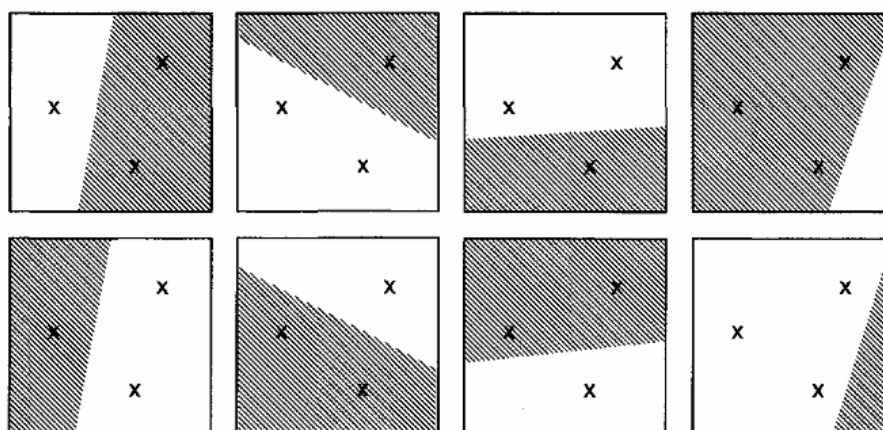
$$\Phi := X \rightarrow H$$

$$x \rightarrow \chi := \Phi(x)$$

Το πεδίο  $H$  ονομάζεται χώρος χαρακτηριστικών. Χρησιμοποιήσαμε το ελληνικό γράμμα  $\chi$  για να συμβολίσουμε την διανυσματική απεικόνιση του δείγματος  $x$  στο πεδίο των χαρακτηριστικών.

#### A4 Διάσταση Vapnik-Chervonenkis (VC)

Μια κλάση από συναρτήσεις διαχωρισμού που περιγράφεται από μια γενική μορφή όπως για παράδειγμα τα υπερ-επίπεδα τύπου  $f(x) = \text{sgn}(\langle w, x \rangle + b)$ , μπορούν να διαχωρίσουν τα δεδομένα μας με συγκεκριμένο τρόπο. Αν υποθέσουμε ότι οι κλάσεις είναι δύο  $y \in \{\pm 1\}$  τότε υπάρχουν  $2^m$  διαφορετικές ταξινομήσεις των  $m$  δειγμάτων. Μια πολύ πλούσια κλάση συναρτήσεων θα μπορούσε να ορίσει και τους  $2^m$  τρόπους διαχωρισμού, να αποκλείσει όλα τα σημεία. Κάποιες άλλες κλάσεις συναρτήσεων θα μπορούσαν να μην έχουν την ίδια ικανότητα διαχωρισμού. Ως διάσταση VC μίας κλάσης συναρτήσεων διαχωρισμού ορίζεται ο μέγιστος αριθμός δειγμάτων  $m$  που κλάση μπορεί να αποκλείσει.



**Εικόνα Α. 1 (Scholkopf & Smola, 2002)** Ένα απλό παράδειγμα διάστασης VC. Υπάρχουν 3 σημεία άρα και  $2^3=8$  τρόποι να ταξινομηθούν τα σημεία. Για τον δισδιάστατο χώρο οι 8 τρόποι διαχωρισμού μπορούν να περιγραφούν με υπερ-επίπεδα διαχωρισμού. Κατά συνέπεια η κλάση συναρτήσεων των υπερ-επίπεδων στο δισδιάστατο χώρο έχει διάσταση VC ίση με τρία.

#### A5 Υπέρ-επίπεδα διαχωρισμού

Ας υποθέσουμε ότι έχουμε μια κλάση συναρτήσεων διαχωρισμού που ορίζεται από όλα τα πιθανά υπερ-επίπεδα σε κάποιο χώρο εσωτερικών γινομένων (dot product space)  $H$ , της μορφής:

$\langle w, x \rangle + b = 0$  όπου  $w \in H, b \in \mathfrak{R}$  (1) όπου το  $w$  ονομάζεται **κανονικό διάνυσμα** και το  $b$  **κατώφλι**.

τότε οι αντίστοιχες συναρτήσεις απόφασης θα είναι της μορφής:

$$f(x) = \text{sgn}(\langle w, x \rangle + b) \quad (2)$$

Τα δεδομένα τα οποία μπορούν να διαχωριστούν με υπερ-επίπεδα ονομάζονται **γραμμικά διαχωρίσιμα**. Ανάμεσα σε όλα τα πιθανά υπερ-επίπεδα υπάρχει ένα **μοναδικό** υπερ-επίπεδο που διαχωρίζει τα δεδομένα με βέλτιστο τρόπο. Το βέλτιστο υπερ-επίπεδο έχει την μέγιστη απόσταση από τα σημεία εκπαίδευσης και μπορεί να υπολογιστεί ως εξής:

$$\max_{w \in H, b \in \mathfrak{R}} \text{imize} = \min \{ \|x - x_i\| \mid x \in H, \langle w, x \rangle + b = 0, i = 1, \dots, m \} \quad (3)$$

Το πρόβλημα της μορφής (3) μπορεί να λυθεί με μεθόδους τετραγωνικού προγραμματισμού (quadratic programming). Συγκεκριμένα για να υπολογιστεί το υπερ-επίπεδο το οποίο μεγιστοποιεί το **περιθώριο** (δες Εικόνα Α.2) θα πρέπει να λυθεί ένα πρόβλημα βελτιστοποίησης με περιορισμούς, της μορφής:

$$\min_{w \in H, b \in \mathfrak{R}} \text{imize} = \tau(w) - \frac{1}{2} \|w\|^2 \quad (4)$$

$$\text{υπό το περιορισμό ότι: } y_i (\langle w, x_i \rangle + b) \geq 1 \quad \forall i = 1, \dots, m \quad (5)$$

Ο περιορισμός (5) εξασφαλίζει ότι η τιμή της συνάρτησης απόφασης  $f(x_i)$  θα είναι +1 όταν  $y_i = +1$  και -1 όταν  $y_i = -1$ . Γενικά μπορεί να αποδειχτεί ότι το κανονικό διάνυσμα  $w$  με το ελάχιστο μήκος είναι αυτό που μεγιστοποιεί και το **περιθώριο** (δες Εικόνα Α.2).





$$L(w, b, a) = \frac{1}{2} \|w\|^2 - \sum_{i=1}^m a_i (y_i (\langle x_i, w \rangle + b) - 1) \quad (6)$$

Ο συνάρτηση Lagrange  $L$  πρέπει να ελαχιστοποιηθεί ως προς τις βασικές μεταβλητές  $w$  και  $b$  και να μεγιστοποιηθεί ως προς τις μεταβλητές Lagrange  $a_i$  (με άλλα λόγια να βρεθεί ένα σημείο καμπής)

Για να καταλάβουμε πως ικανοποιείται ο περιορισμός ανισότητας, ας φανταστούμε ότι παραβιάζεται, δηλαδή  $y_i (\langle w, x_i \rangle + b) - 1 < 0$ . Στην περίπτωση αυτή η συνάρτηση  $L$  αυξάνεται όσο αυξάνονται και οι πολλαπλασιαστές  $a_i \geq 0$  και οι τιμές των μεταβλητών  $w$  και  $b$  πρέπει να μεταβληθούν έτσι ώστε η συνάρτηση  $L$  να μειωθεί.

Η συνάρτηση  $L$  ελαχιστοποιείται στο σημείο όπου μηδενίζονται οι μερικοί παράγωγοι:

$$\frac{\partial}{\partial b} L(w, b, a) = 0 \text{ και } \frac{\partial}{\partial w} L(w, b, a) = 0 \quad (7)$$

Οι λύση αυτών των εξισώσεων δίνει σαν αποτέλεσμα:

$$\sum_{i=1}^m a_i y_i = 0 \text{ και } (8)$$

$$w = \sum_{i=1}^m a_i y_i x_i \quad (9)$$

Η εξίσωση (9) υπονοεί ότι η λύση είναι το υποσύνολο εκείνο των σημείων εκπαίδευσης τα οποία έχουν μη μηδενικές τιμές  $a_i$ . Τα σημεία αυτά ονομάζονται **διανύσματα υποστήριξης (support vectors)**. Τα διανύσματα υποστήριξης αυτά ικανοποιούν την ισότητα  $(y_i (\langle x_i, w \rangle + b) - 1) = 0$  ενώ όλα τα υπόλοιπα σημεία έχουν τιμές μεγαλύτερες  $(y_i (\langle x_i, w \rangle + b) - 1) \geq 0$  και δεν έχουν σχέση με το υπερ-επίπεδο διαχωρισμού και μπορούν να παραλειφθούν. Το υπερ-επίπεδο διαχωρισμού ορίζεται έμμεσα από τα διανύσματα υποστήριξης.

Αντικαθιστώντας τις εξισώσεις (8) και (9) στην συνάρτηση Lagrange (6) οι βασικές μεταβλητές  $w$  και  $b$  εξαλείφονται και το πρόβλημα που προκύπτει **ονομάζεται πρόβλημα διπλής βελτιστοποίησης**.

$$\text{maximize } W(a) = \sum_{i=1}^m a_i - \frac{1}{2} \sum_{i,j=1}^m a_i a_j y_i y_j \langle x_i, x_j \rangle$$

**Υπό τον περιορισμό**  $a_i \geq 0 \forall i = 1, \dots, m$  και  $\sum_{i=1}^m a_i y_i = 0$

Αντικαθιστώντας την εξίσωση (9) στην (2) προκύπτει η συνάρτηση απόφασης

$$f(x) = \text{sgn} \left( \sum_{i=1}^m y_i a_i \langle x, x_i \rangle + b \right) \quad (10)$$

Το πρόβλημα της εύρεσης του βέλτιστου υπερ-επίπεδου διαχωρισμού έχει ανάλογο πρόβλημα στο πεδίο της μηχανικής όπου κάθε διάνυσμα υποστήριξης αντιστοιχεί σε κάθετη δύναμη με μεγέθους  $a_i$  και κατεύθυνσης  $y_j \cdot \frac{w}{\|w\|}$  που ασκείται σε ένα επίπεδο. Η ικανοποίηση της εξίσωσης (8) σημαίνει ότι όλες οι δυνάμεις που ασκούνται στο επίπεδο πρέπει να αθροίζονται στο μηδέν.

### **A6 Ιδιότητες εσωτερικών γινομένων**

Έστω  $X$  ένα υποσύνολο του χώρου  $\mathbb{R}^N$ , ( $N \in \mathbb{N}$ ) στο οποίο ορίζονται εσωτερικά γινόμενα. Ας υποθέσουμε ότι για τα πρότυπα  $x \in X$  η περισσότερη πληροφορία βρίσκεται στα γινόμενα τάξης  $d$  (ή αλλιώς μονώνυμα) των στοιχείων  $[x]_j$  του  $x$

$$[x]_{j_1} \cdot [x]_{j_2} \cdots [x]_{j_d} \text{ όπου } j_1, \dots, j_d \in (1, \dots, N)$$

Ας πάρουμε το απλό παράδειγμα όπου έχουμε δυσδιάστατα πρότυπα, δηλαδή  $X = \mathbb{R}^2$ . Σε αυτήν μπορούμε να ορίσουμε όλα τα μονώνυμα δεύτερης τάξης μέσω μίας μη γραμμικής απεικόνισης:

$$\begin{aligned} \Phi : \mathbb{R}^2 &\rightarrow H = \mathbb{R}^3, \\ ([x]_1, [x]_2) &\mapsto ([x]_1, [x]_2, [x]_1 [x]_2) \end{aligned}$$

Ο υπολογισμός όλων των πιθανών μονώνυμων στην περίπτωση πολυδιάστατων χώρων είναι δύσκολος έως ακατόρθωτος. Συγκεκριμένα για  $N$ -διάστατο χώρο υπάρχουν

$N_H = \binom{d+N-1}{d} = \frac{(d+N-1)!}{d!(N-1)!}$  μονώνυμα τάξης  $d$ , τα οποία ορίζουν ένα χώρο χαρακτηριστήκαν  $H$  διάστασης  $N_H$ .

Παρακάτω θα εξηγήσουμε πώς σε συγκεκριμένες περιπτώσεις μπορούμε να υπολογίσουμε το εσωτερικό γινόμενο των σημείων σε ένα χώρο χαρακτηριστικών μεγαλύτερης διάστασης χωρίς να χρειαστεί να υπολογίσουμε την ακριβή απεικόνιση των σημείων στο χώρο αυτό.

### A7 Το τέχνασμα της συνάρτησης πυρήνα

Για να υπολογίσουμε τα εσωτερικά γινόμενα στον χώρο των χαρακτηριστικών  $\langle \Phi(x), \Phi(x') \rangle$  θα χρησιμοποιήσουμε συναρτήσεις πυρήνα της μορφής:

$$k(x, x') = \langle \Phi(x), \Phi(x') \rangle \quad (13)$$

Η συνάρτηση πυρήνα θα μας βοηθήσει να υπολογίσουμε την τιμή του εσωτερικού γινομένου στο χώρο των χαρακτηριστικών  $H$  χωρίς να χρειαστεί να υπολογίσουμε την ακριβή απεικόνιση των σημείων στον χώρο αυτό (Εικόνα Α.3).

Ας υποθέσουμε ότι έχουμε σημεία δυοδιάστατα και θέλουμε να υπολογίσουμε τα μονώνυμα δεύτερης τάξης, δηλαδή  $N=d=2$ . Τότε για την απεικόνιση:

$$\Phi([x]_1, [x]_2) \mapsto ([x]_1^2, [x]_2^2, [x]_1[x]_2, [x]_2[x]_1)$$

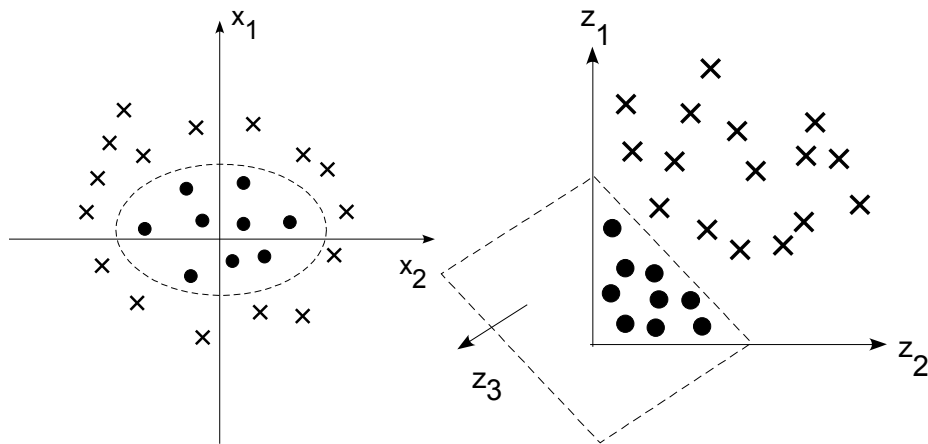
(σημείωση: θεωρούμε αρχικά ότι έχουμε διατεταγμένα μονώνυμα, δηλαδή το  $[x]_1[x]_2$  είναι διαφορετικό από το  $[x]_2[x]_1$ )

το εσωτερικό γινόμενο υπολογίζεται ως:

$$\langle \Phi(x), \Phi(x') \rangle = [x]_1^2 [x']_1^2 + [x]_2^2 [x']_2^2 - 2[x]_1 [x']_1 [x]_2 [x']_2 = \langle x, x' \rangle^2$$

Με άλλα λόγια το εσωτερικό γινόμενο στο χώρο των χαρακτηριστικών  $H$  είναι ίσο με το τετράγωνο του εσωτερικού γινομένου στο χώρο εισόδου  $X$ . Το ίδιο ισχύει και για την γενίκευση των  $N$  και  $d \in \mathbb{N}$ .

$$\langle \Phi_d(x), \Phi_d(x') \rangle = \langle x, x' \rangle^d$$



**Εικόνα Α. 3 (Scholkopf & Smola, 2002)** Παράδειγμα δυαδική ταξινόμησης και απεικόνισης των σημείων εισόδου στον χώρο των χαρακτηριστικών όπου και είναι γραμμικά διαχωρίσιμα. Στο συγκεκριμένο παράδειγμα υποθέτουμε ότι τα σημεία εκπαίδευσης δύο κλάσεων (κύκλοι και σταυροί) βρίσκονται στον δυσδιάστατο χώρο και συνάρτηση διαχωρισμού τους είναι μια έλλειψη. Όταν απεικονίζουμε τα σημεία εκπαίδευσης στο μη γραμμικό χώρο  $\Phi_2(x) = \Phi_2(z_1, z_2, z_3) = ([x]_1^2, [x]_2^2, \sqrt{2}[x]_1[x]_2)$  η συνάρτηση διαχωρισμού του είναι ένα υπερ-επίπεδο.

### A8 Ταξινομητές διανυσμάτων υποστήριξης

Στις ενότητες A2 και A3 ορίσαμε τον χώρο των χαρακτηριστικών ως ένα χώρο εσωτερικών γινομένων στον οποίο απεικονίζουμε τα πρότυπα σημεία μέσω μίας συνάρτησης,

$$\begin{aligned} \Phi &:= X \rightarrow H \\ x &\rightarrow \chi := \Phi(x) \end{aligned} \quad (11)$$

αλλά και την συνάρτηση πυρήνα ως ένα μέτρο ομοιότητας μεταξύ δύο σημείων, ενώ στην ενότητα A7 δείξαμε πως η συνάρτηση πυρήνα μπορεί να χρησιμοποιηθεί για τον υπολογισμό των εσωτερικών γινομένων στον χώρο των χαρακτηριστικών.

$$k(x, x') = \langle \Phi(x), \Phi(x') \rangle \quad (12)$$

Στην παρούσα ενότητα έχουμε όλους τους απαραίτητους ορισμούς και μπορούμε να περιγράψουμε τις μηχανές διανυσμάτων υποστήριξης. Το πρόβλημα του γραμμικού διαχωρισμού περιγράφηκε στην ενότητα Α5 με την χρήση εσωτερικών γινομένων, μπορούμε λοιπόν να αντικαταστήσουμε την συνάρτηση πυρήνα η οποία υπολογίζει εσωτερικά γινόμενα, στην θέση του εσωτερικού γινομένου της συνάρτησης απόφασης (10), τότε έχουμε:

$$f(x) = \text{sgn} \left( \sum_{i=1}^m y_i a_i \langle \Phi(x), \Phi(x_i) \rangle + b \right) = \text{sgn} \left( \sum_{i=1}^m y_i a_i k(x, x_i) + b \right) \quad (13)$$

Το αντίστοιχο πρόβλημα βελτιστοποίησης είναι της μορφής:

$$\text{maximize } W(a) = \sum_{i=1}^m a_i - \frac{1}{2} \sum_{i,j=1}^m a_i a_j y_i y_j k(x, x_i)$$

$$\text{Υπό τον περιορισμό } a_i \geq 0 \quad \forall i = 1, \dots, m \text{ και } \sum_{i=1}^m a_i y_i = 0$$

### A9 Ταξινομητές με χαλαρό περιθώριο

Συχνά στην πράξη δεν υπάρχει πάντα κατάλληλο υπερ-επίπεδο διαχωρισμού λόγω ύπαρξης θορύβου στα πρότυπα σημεία το οποίο οδηγεί σε επικάλυψη των σημείων των δύο κλάσεων. Για να επιτρέψουμε στο πρόβλημα της βελτιστοποίησης να παραβιάσει ως ένα βαθμό τους περιορισμούς (ανισότητα (5)) εισάγουμε βοηθητικές μεταβλητές

$$\xi_i \geq 0 \quad \forall i = 1, \dots, m$$

για να χαλαρώσουμε τον περιορισμό (5)

$$y_i (\langle w, x_i \rangle + b) \geq 1 - \xi_i \quad (14)$$

Σε συγκεκριμένη περίπτωση ο κατάλληλος ταξινομητής μπορεί να υπολογιστεί ελαχιστοποιώντας την συνάρτηση εκτίμησης :

$$\tau(w, \xi) = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^m \xi_i \quad (15)$$

---

**A10 Παραδείγματα συναρτήσεων πυρήνα**

Πολυωνυμικός

$$k(x, x') = \langle x, x' \rangle^d$$

Gaussian

$$k(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right)$$

Σιγμοειδής

$$k(x, x') = \tanh(\kappa \langle x, x' \rangle + \Theta)$$

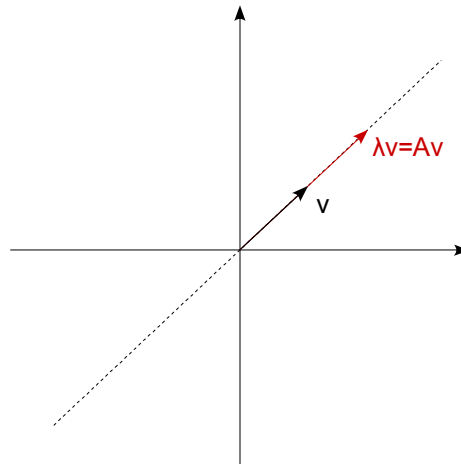
## ΠΑΡΑΡΤΗΜΑ Β

### Β1 Ιδιοτιμές και ιδιοδιανύσματα

Έστω  $A$  ένας  $N \times N$  πίνακας με πραγματικά στοιχεία. Ο πραγματικός ή μιγαδικός αριθμός  $\lambda$  καλείται **ιδιοτιμή** του πίνακα  $A$  εάν υπάρχει μη μηδενικό διάνυσμα  $v$  με πραγματικά ή μιγαδικά στοιχεία τέτοιο ώστε:

$$Av = \lambda v$$

Το μη μηδενικό διάνυσμα  $v$  καλείται **ιδιοδιάνυσμα** του πίνακα  $A$  για πού αντιστοιχεί στην ιδιοτιμή  $\lambda$ .



**Εικόνα Β. 1** Ιδιοδιανύσματα ενός πίνακα  $A$  είναι τα διανύσματα τα οποία όταν πολλαπλασιαστούν με τον πίνακα  $A$  δεν αλλάζουν κατεύθυνση αλλά μόνο το μήκος.

Για να εξηγήσουμε τι σημαίνουν οι ιδιοτιμές θα πρέπει πρώτα να εξηγήσουμε τι σημαίνουν τα ιδιοδιανύσματα. Σχεδόν όλα τα διανύσματα αλλάζουν κατεύθυνση εάν πολλαπλασιαστούν με τον πίνακα  $A$  (μετασχηματιστούν). Συγκριμένα ξεχωριστά διανύσματα  $x$  παραμένουν στην ίδια κατεύθυνση όταν πολλαπλασιαστούν με τον πίνακα  $A$ , έχουν την ίδια κατεύθυνση με το διάνυσμα  $Ax$ . **Αυτά τα διανύσματα τα ονομάζουμε ιδιοδιανύσματα.** Πολλαπλασιάζοντας τον πίνακα  $A$  με το ιδιοδιάνυσμα του τότε είναι σαν να πολλαπλασιάζεις το ιδιοδιάνυσμα  $x$  με μια σταθερή τιμή  $\lambda$  η οποία στην ουσία μεταβάλλει μόνο το μήκος του ιδιοδιανύσματος  $x$ .



Έστω ένας πίνακας  $\mathbf{A}$   $N \times N$  με πραγματικά στοιχεία. Ο πραγματικός ή μιγαδικός αριθμός  $\lambda$  είναι ιδιοτιμή του πίνακα  $\mathbf{A}$  εάν και μόνο αν

$$\det(\mathbf{A} - \lambda \mathbf{I}) = 0$$

### Απόδειξη

Ένας αριθμός  $\lambda$  είναι ιδιοτιμή του  $\mathbf{A}$  εάν και μόνο εάν

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v},$$

όπου το  $\mathbf{v}$  είναι μη μηδενικό διάνυσμα, οπότε έχουμε τις ισοδυναμίες

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v} \Leftrightarrow \mathbf{A}\mathbf{v} - \lambda\mathbf{v} = \mathbf{0} \Leftrightarrow (\mathbf{A} - \lambda\mathbf{I})\mathbf{v} = \mathbf{0}$$

Δηλαδή το σύστημα:

$$\mathbf{A} - \lambda\mathbf{I} = \begin{bmatrix} \alpha_{11} - \lambda & \alpha_{12} & \dots & \alpha_{1N} \\ \alpha_{12} & \alpha_{22} - \lambda & \dots & \alpha_{2N} \\ \dots & \dots & \dots & \dots \\ \alpha_{N1} & \alpha_{N2} & \dots & \alpha_{NN} - \lambda \end{bmatrix}$$

αυτό σημαίνει ότι η ορίζουσα του πίνακα είναι μηδενική, δηλαδή οι ρίζες του  $\det(\mathbf{A} - \lambda\mathbf{I}) = 0$  είναι οι ιδιοτιμές του πίνακα μιας και αυτές είναι οι μοναδικές τιμές που μηδενίζουν την ορίζουσα του συστήματος.

Το πολυώνυμο  $p(\lambda) = (-1)^N \det(\mathbf{A} - \lambda\mathbf{I}) = \lambda^N + b_{N-1}\lambda^{N-1} + \dots + b_1\lambda + b_0$  ονομάζεται **χαρακτηριστικό πολυώνυμο** και η εξίσωση  $\det(\mathbf{A} - \lambda\mathbf{I}) = 0$  **χαρακτηριστική εξίσωση**. Η χαρακτηριστική εξίσωση έχει  $N$  στο σύνολο πραγματικές ή μιγαδικές λύσεις οι οποίες είναι οι ρίζες του χαρακτηριστικού πολυώνυμου. Ένας συμμετρικός πίνακας έχει ορθογώνια ιδιοδιανύσματα.

## ΠΑΡΑΡΤΗΜΑ C

### C1 Εκτίμηση της ανά ζεύγος ενέργειας αλληλεπίδρασης χρησιμοποιώντας την αμινοξική σύσταση μιας αλληλουχίας

Η ενέργεια αλληλεπίδρασης μεταξύ των αμινοξέων μιας πρωτεΐνης είναι συνάρτηση της της αμινοξικής σύστασης αλλά της τελικής διαμόρφωσης της. Η συνολική ενέργεια της πρωτεΐνης μπορεί να υπολογιστεί αν πολλαπλασιάσουμε τον συνολικό αριθμό των αλληλεπιδράσεων με την ενέργεια αλληλεπίδρασης.

Η ενέργεια αλληλεπίδρασης αν ζεύγος αμινοξέων έχει υπολογιστεί και συνοψίζεται σε ένα πίνακα 20x20 (Εικόνα C.1). Βασιζόμενοι στον πίνακα αυτόν η συνολική ενέργεια μιας αλληλουχίας μπορεί να υπολογιστεί ως:

$$E = \sum_{ij}^{20} M_{ij} C_{ij} ,$$

όπου  $M_{ij}$  είναι η ενέργεια αλληλεπίδρασης ανάμεσα στα αμινοξέα τύπου  $i$  και  $j$  και  $C_{ij}$  είναι ο αριθμός των αλληλεπιδράσεων ανάμεσα σε αμινοξέα τύπου  $i$  και  $j$  σύμφωνα με την διαμόρφωση της πρωτεΐνης. Τότε μπορούμε να προσεγγίσουμε την ενέργεια ανά αμινοξύ  $E / L$  μέσω της αμινοξικής σύστασης.

Table 1. M matrix

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	-0.20	-0.44	0.16	0.26	-0.46	-0.26	0.50	-0.57	0.10	-0.36	-0.22	0.07	0.14	0.01	0.20	-0.09	-0.05	-0.42	0.05	-0.50
C	-0.44	-2.99	0.21	0.19	-0.88	-0.34	-1.11	-0.36	-0.09	-0.53	-0.43	-0.52	-0.14	-0.43	-0.24	0.13	-0.22	-0.62	0.24	-0.79
D	0.16	0.21	0.17	0.55	0.38	0.35	-0.23	0.44	-0.39	0.28	0.35	-0.02	1.03	0.49	-0.37	0.19	-0.12	0.69	0.04	0.43
E	0.26	0.19	0.55	0.60	0.55	0.65	0.18	0.37	-0.47	0.33	0.29	0.01	0.69	0.04	-0.52	0.18	0.37	0.39	0.03	0.17
F	-0.46	-0.88	0.38	0.55	-0.94	0.17	-0.40	-0.88	0.01	-1.08	-0.78	0.22	0.20	0.26	-0.19	-0.22	0.02	-1.15	-0.60	-0.88
G	-0.26	-0.34	0.35	0.65	0.17	-0.12	0.18	0.24	0.19	0.24	0.02	-0.04	0.60	0.46	0.50	0.28	0.28	0.27	0.51	-0.35
H	0.50	-1.11	-0.23	0.18	-0.40	0.18	0.42	-0.00	0.79	-0.24	-0.07	0.20	0.25	0.69	0.24	0.21	0.11	0.16	-0.85	-0.26
I	-0.57	-0.36	0.44	0.37	-0.88	0.24	-0.00	-1.16	0.15	-1.25	-0.58	-0.09	0.36	-0.08	0.14	0.32	-0.27	-1.06	-0.68	-0.85
K	0.10	-0.09	-0.39	-0.47	0.01	0.19	0.79	0.15	0.42	0.13	0.48	0.26	0.50	0.15	0.53	0.10	-0.19	0.10	0.10	0.04
L	-0.36	-0.53	0.28	0.33	-1.08	0.24	-0.24	-1.25	0.13	-1.10	-0.50	0.21	0.42	-0.01	-0.07	0.17	0.07	-0.97	-0.95	-0.63
M	-0.22	-0.43	0.35	0.29	-0.78	0.02	-0.07	-0.58	0.48	-0.50	-0.74	0.32	0.01	0.26	0.15	0.48	0.16	-0.73	-0.56	-1.02
N	0.07	-0.52	-0.02	0.01	0.22	-0.04	0.20	-0.09	0.26	0.21	0.32	0.14	0.27	0.37	0.13	0.15	0.10	0.40	-0.12	0.32
P	0.14	-0.14	1.03	0.69	0.20	0.60	0.25	0.36	0.50	0.42	0.01	0.27	0.27	1.02	0.47	0.54	0.88	-0.02	-0.37	-0.12
Q	0.01	-0.43	0.49	0.04	0.26	0.46	0.69	-0.08	0.15	-0.01	0.26	0.37	1.02	-0.12	0.24	0.29	0.04	-0.11	0.18	0.11
R	0.20	-0.24	-0.37	-0.52	-0.19	0.50	0.24	0.14	0.53	-0.07	0.15	0.13	0.47	0.24	0.17	0.27	0.45	0.01	-0.73	0.01
S	-0.09	0.13	0.19	0.18	-0.22	0.28	0.21	0.32	0.10	0.17	0.48	0.15	0.54	0.29	0.27	-0.06	0.08	0.12	-0.22	-0.14
T	-0.05	-0.22	-0.12	0.37	0.02	0.28	0.11	-0.27	-0.19	0.07	0.16	0.10	0.88	0.04	0.45	0.08	-0.03	-0.01	0.11	-0.32
V	-0.42	-0.62	0.69	0.39	-1.15	0.27	0.16	-1.06	0.10	-0.97	-0.73	0.40	-0.02	-0.11	0.01	0.12	-0.01	-0.89	-0.56	-0.71
W	0.05	0.24	0.04	0.03	-0.60	0.51	-0.85	-0.68	0.10	-0.95	-0.56	-0.12	-0.37	0.18	-0.73	-0.22	0.11	-0.56	-0.05	-1.41
Y	-0.50	-0.79	0.43	0.17	-0.88	-0.35	-0.26	-0.85	0.04	-0.63	-1.02	0.32	-0.12	0.11	0.01	-0.14	-0.32	-0.71	-1.41	-0.76

Contact potential derived from 785 proteins using the approach of Thomas & Dill.<sup>20</sup>

Εικόνα C.1 – Πίνακας πιθανότητας αλληλεπιδράσεων (Dosztanyi et al, 2005)

Υποθέτουμε ότι η ενεργειακή συνεισφορά κάθε αμινοξέος δεν εξαρτάται μόνο από τον τύπο του αλλά και από τα πιθανά αμινοξέα με τα οποία μπορεί να αλληλεπιδράσει μέσα στην

αλληλουχία. Κατ' επέκταση αν μια αλληλουχία περιέχει αμινοξέα που μπορούν να σχηματίσουν ευνοϊκές αλληλεπιδράσεις με το αντίστοιχο αμινοξύ τότε και η ενεργειακή συνεισφορά της αλληλεπίδρασης είναι πιο ευνοϊκή. Έστω ότι  $n_i$  η συχνότητα ενός αμινοξέος στην αλληλουχία, δηλαδή  $n_i = \frac{N_i}{L}$  (όπου  $L$  το μήκος της αλληλουχίας). Τότε η συσχέτιση μεταξύ της συχνότητας το αμινοξέων και της συνολικής ενέργεια αλληλεπίδρασης μπορεί να περιγραφεί ως μια τετραγωνική εξίσωση της μορφής:

$$\frac{E_{estimated}}{L} = \sum_{ij} n_i P_{ij} n_j$$

όπου  $P$  είναι ο πίνακας πρόβλεψης ενέργειας (energy predictor matrix). Κάθε στοιχείο  $P_{ij}$  του πίνακα λέει πως η ενέργεια του αμινοξέος  $i$  εξαρτάται από την αμινοξική σύσταση της αλληλουχίας σε αμινοξέα τύπου  $j$ . Οι τιμές του πίνακα  $P$  υπολογίστηκαν με την μέθοδο ελάχιστων τετραγώνων και χρησιμοποιώντας λυμένες δομές σφαιρικών μορίων (Dosztanyi et al, 2005) για την εκτίμηση της συχνότητας των αλληλεπιδράσεων.

Με βάση την αθροιστική ιδιότητα της ενέργειας μπορούμε να αναλύσουμε την συνολική ενέργεια της πρωτεΐνης  $k$  στις ενεργειακές συνεισφορές κάθε είδους αμινοξέος  $E_k = \sum_i e_i^k$ , όπου  $e_i^k$  είναι η ενέργεια όλων των αμινοξέων τύπου  $i$  που αλληλεπιδρούν με άλλα κατάλοιπα της αλληλουχίας. Η τιμή της ενέργειας  $e_i^k$  εξαρτάται από τον αριθμό των αλληλεπιδράσεων που δημιουργεί με άλλα κατάλοιπα τύπου  $j$ :

$$e_i^k (calculated) = \sum_j M_{ij} C_{ij}^k$$

και μπορεί να προσεγγιστεί ως εξής:

$$e_i^k (estimated) = N_i^k \sum_{j=1}^{20} P_{ij} n_j^k$$

Η τιμές για κάθε σειρά του πίνακα  $P$  υπολογίζονται εάν ελαχιστοποιήσουμε την συνάρτηση:

$$Z_i = \sum_k \left( e_i^k - N_i^k \sum_{j=1}^{20} P_{ij} n_j^k \right)^2$$

Η λύση του προβλήματος βρίσκεται στην Εικόνα C.2.

Table 2. P energy predictor matrix

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	-1.65	-2.83	1.16	1.80	-3.73	-0.41	1.90	-3.69	0.49	-3.01	-2.08	0.66	1.54	1.20	0.98	-0.08	0.46	-2.31	0.32	-4.62
C	-2.83	-39.58	-0.82	-0.53	-3.07	-2.96	-4.98	0.34	-1.38	-2.15	1.43	-4.18	-2.13	-2.91	-0.41	-2.33	-1.84	-0.16	4.26	-4.46
D	1.16	-0.82	0.84	1.97	-0.92	0.88	-1.07	0.68	-1.93	0.23	0.61	0.32	3.31	2.67	-2.02	0.91	-0.65	0.94	-0.71	0.90
E	1.80	-0.53	1.97	1.45	0.94	1.31	0.61	1.30	-2.51	1.14	2.53	0.20	1.44	0.10	-3.13	0.81	1.54	0.12	-1.07	1.29
F	-3.73	-3.07	-0.92	0.94	-11.25	0.35	-3.57	-5.88	-0.82	-8.59	-5.34	0.73	0.32	0.77	-0.40	-2.22	0.11	-7.05	-7.09	-8.80
G	-0.41	-2.96	0.88	1.31	0.35	-0.20	1.09	-0.65	-0.16	-0.55	-0.52	-0.32	2.25	1.11	0.84	0.71	0.59	-0.38	1.69	-1.90
H	1.90	-4.98	-1.07	0.61	-3.57	1.09	1.97	-0.71	2.89	-0.86	-0.75	1.84	0.35	2.64	2.05	0.82	-0.01	0.27	-7.58	-3.20
I	-3.69	0.34	0.68	1.30	-5.88	-0.65	-0.71	-6.74	-0.01	-9.01	-3.62	-0.07	0.12	-0.18	0.19	-0.15	0.63	-6.54	-3.78	-5.26
K	0.49	-1.38	-1.93	-2.51	-0.82	-0.16	2.89	-0.01	1.24	0.49	1.61	1.12	0.51	0.43	2.34	0.19	-1.11	0.19	0.02	-1.19
L	-3.01	-2.15	0.23	1.14	-8.59	-0.55	-0.86	-9.01	0.49	-6.37	-2.88	0.97	1.81	-0.58	-0.60	-0.41	0.72	-5.43	-8.31	-4.90
M	-2.08	1.43	0.61	2.53	-5.34	-0.52	-0.75	-3.62	1.61	-2.88	-6.49	0.21	0.75	1.90	2.09	1.39	0.63	-2.59	-6.88	-9.73
N	0.66	-4.18	0.32	0.20	0.73	-0.32	1.84	-0.07	1.12	0.97	0.21	0.61	1.15	1.28	1.08	0.29	0.46	0.93	-0.74	0.93
P	1.54	-2.13	3.31	1.44	0.32	2.25	0.35	0.12	0.51	1.81	0.75	1.15	-0.42	2.97	1.06	1.12	1.65	0.38	-2.06	-2.09
Q	1.20	-2.91	2.67	0.10	0.77	1.11	2.64	-0.18	0.43	-0.58	1.90	1.28	2.97	-1.54	0.91	0.85	-0.07	-1.91	-0.76	0.01
R	0.98	-0.41	-2.02	-3.13	-0.40	0.84	2.05	0.19	2.34	-0.60	2.09	1.08	1.06	0.91	0.21	0.95	0.98	0.08	-5.89	0.36
S	-0.08	-2.33	0.91	0.81	-2.22	0.71	0.82	-0.15	0.19	-0.41	1.39	0.29	1.12	0.85	0.95	-0.48	-0.06	0.13	-3.03	-0.82
T	0.46	-1.84	-0.65	1.54	0.11	0.59	-0.01	0.63	-1.11	0.72	0.63	0.46	1.65	-0.07	0.98	-0.06	-0.96	1.14	-0.65	-0.37
V	-2.31	-0.16	0.94	0.12	-7.05	-0.38	0.27	-6.54	0.19	-5.43	-2.59	0.93	0.38	-1.91	0.08	0.13	1.14	-4.82	-2.13	-3.59
W	0.32	4.26	-0.71	-1.07	-7.09	1.69	-7.58	-3.78	0.02	-8.31	-6.88	-0.74	-2.06	-0.76	-5.89	-3.03	-0.65	-2.13	-1.73	-12.39
Y	-4.62	-4.46	0.90	1.29	-8.80	-1.90	-3.20	-5.26	-1.19	-4.90	-9.73	0.93	-2.09	0.01	0.36	-0.82	-0.37	-3.59	-12.39	-2.68

The pairwise energy per amino acid is estimated as a quadratic form in the amino acid composition vector using the elements of this matrix.

Εικόνα C.2 – Πίνακας P πρόβλεψης ενέργειας αλληλεπιδράσεων (Dosztanyi et al, 2005)