**ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ
ΤΜΗΜΑ ΙΑΤΡΙΚΗΣ
ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ ΣΤΙΣ
ΒΑΣΙΚΕΣ ΙΑΤΡΙΚΕΣ ΕΠΙΣΤΗΜΕΣ**

**Στατιστική Ανακάλυψη και πειραματική
επιβεβαίωση μεταγραφικών παραγόντων που
ελέγχουν την ενεργοποίηση των λεμφοκυττάρων
σε άνοσες καταστάσεις**

**Statistical Discovery and Experimental validation
of transcription factors controlling lymphocyte
activation in immune-related states**

**Χρήστος Αργυρόπουλος M.D. MSc
Διδακτορική Διατριβή**

Πάτρα, Φεβρουάριος 2005

**ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ
ΤΜΗΜΑ ΙΑΤΡΙΚΗΣ
ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ ΣΤΙΣ
ΒΑΣΙΚΕΣ ΙΑΤΡΙΚΕΣ ΕΠΙΣΤΗΜΕΣ**

**Στατιστική Ανακάλυψη και πειραματική
επιβεβαίωση μεταγραφικών παραγόντων που
ελέγχουν την ενεργοποίηση των λεμφοκυττάρων
σε άνοσες καταστάσεις**

**Statistical Discovery and Experimental validation
of transcription factors controlling lymphocyte
activation in immune-related states**

**Χρήστος Αργυρόπουλος M.D. MSc
Διδακτορική Διατριβή**

Πάτρα, Φεβρουάριος 2005

Τριμελής Συμβουλευτική Επιτροπή:

1. Γεώργιος Νικηφορίδης, Καθηγητής Ιατρικής Φυσικής, επιβλέπων
2. Ιωάννης Βαράκης, Καθηγητής Ανατομίας, μέλος συμβουλευτικής επιτροπής
3. Αθανασία Μουζάκη, Αν. Καθηγήτρια Αιματολογίας, μέλος συμβουλευτικής επιτροπής

Επταμελής Εξεταστική Επιτροπή:

1. Γεώργιος Νικηφορίδης, Καθηγητής Ιατρικής Φυσικής, επιβλέπων
2. Ιωάννης Βαράκης, Καθηγητής Ανατομίας, μέλος συμβουλευτικής επιτροπής
3. Αθανασία Μουζάκη, Αν. Καθηγήτρια Αιματολογίας, μέλος συμβουλευτικής επιτροπής
4. Μιχάλης Βραχάτης, Καθηγητής Τμήματος Μαθηματικών, μέλος εξεταστικής επιτροπής
5. Αναστάσιος Μπεζεριάνος, Καθηγητής Ιατρικής Φυσικής, μέλος εξεταστικής επιτροπής
6. Αθανάσιος Παπαβασιλείου, Καθηγητής Βιολογικής Χημείας, μέλος εξεταστικής επιτροπής
7. Χριστόδουλος Φλωρδέλλης, Καθηγητής Φαρμακολογίας, μέλος εξεταστικής επιτροπής

Περιεχόμενα

ΠΕΡΙΕΧΟΜΕΝΑ.....	V
ΠΕΡΙΛΗΨΗ	VIII
ΕΥΧΑΡΙΣΤΙΕΣ	IX

1 ΕΙΣΑΓΩΓΗ 1

2 ΤΟ ΒΙΟΛΟΓΙΚΟ ΣΥΣΤΗΜΑ: ΜΟΡΙΑΚΟΙ ΜΗΧΑΝΙΣΜΟΙ ΤΗΣ ΕΝΕΡΓΟΠΟΙΗΣΗΣ ΤΩΝ T - ΚΥΤΤΑΡΩΝ & ΕΛΕΓΧΟΣ ΤΟΥ ΙΟΥ HIV3

2.1 ΜΗΧΑΝΙΣΜΟΙ ΕΝΕΡΓΟΠΟΙΗΣΗΣ Τ ΛΕΜΦΟΚΥΤΤΑΡΩΝ.....	3
2.2 ΜΕΤΑΓΡΑΦΙΚΟΣ ΕΛΕΓΧΟΣ ΤΟΥ ΓΟΝΙΔΙΟΥ ΤΗΣ ΙΝΤΕΡΛΕΥΚΙΝΗΣ – 2.....	3
2.3 ΜΕΤΑΓΡΑΦΙΚΟΣ ΕΛΕΓΧΟΣ ΤΟΥ ΙΟΥ HIV ΣΤΑ ΠΑΡΘΕΝΙΚΑ Τ ΛΕΜΦΟΚΥΤΤΑΡΑ	8
2.4 ΣΥΝΟΨΙΣ ΤΟΥ ΒΙΟΛΟΓΙΚΟΥ ΕΡΩΤΗΜΑΤΟΣ	10

3 ΜΙΑ ΤΥΠΙΚΗ ΜΕΘΟΔΟΛΟΓΙΑ ΓΙΑ ΤΗΝ ΑΝΑΚΑΛΥΨΗ ΜΕΤΑΓΡΑΦΙΚΩΝ ΚΑΤΑΣΤΟΛΕΩΝ ΣΤΑΘΕΡΑΣ ΚΑΤΑΣΤΑΣΗΣ (STEADY STATE TRANSCRIPTIONAL REPRESSORS)..... 11

3.1 ΠΕΙΡΑΜΑΤΙΚΗ ΠΡΟΣΕΓΓΙΣΗ ΣΤΗ ΜΕΛΕΤΗ ΜΕΤΑΓΡΑΦΙΚΩΝ ΠΑΡΑΓΟΝΤΩΝ	11
ΤΕΧΝΙΚΕΣ ΜΕΛΕΤΗΣ ΕΝΕΡΓΟΤΗΤΑΣ ΔΕΣΜΕΥΣΗΣ ΜΕΤΑΓΡΑΦΙΚΩΝ ΠΑΡΑΓΟΝΤΩΝ.....	12
ΜΕΛΕΤΗ ΕΠΙΔΡΑΣΗΣ ΜΕΤΑΓΡΑΦΙΚΟΥ ΠΑΡΑΓΟΝΤΑ ΣΤΟ ΡΥΘΜΟ ΜΕΤΑΓΡΑΦΗΣ	12
ΠΟΣΟΤΙΚΟΠΟΙΗΣΗ ΕΠΙΠΕΔΩΝ ΜΕΤΑΓΡΑΦΙΚΩΝ ΠΑΡΑΓΟΝΤΩΝ	13
3.2 ΠΡΟΣΕΓΓΙΣΕΙΣ ΣΤΗΝ ΕΡΜΗΝΕΙΑ ΠΙΘΑΝΟΤΗΤΩΝ.....	15
Η ΣΥΧΝΟΤΙΚΗ ΕΡΜΗΝΕΙΑ ΠΙΘΑΝΟΤΗΤΑΣ	15
Η ΒΑΥΕΣΙΑΝ ΠΙΘΑΝΟΘΕΩΡΗΤΙΚΗ ΣΧΟΛΗ	16
3.3 ΒΑΥΕΣΙΑΝ ΠΕΙΡΑΜΑΤΙΚΟΣ ΣΧΕΔΙΑΣΜΟΣ.....	17
3.4 ΠΕΙΡΑΜΑΤΙΚΗ ΠΡΟΣΕΓΓΙΣΗ ΣΤΗΝ ΑΝΙΧΝΕΥΣΗ ΜΕΤΑΓΡΑΦΙΚΩΝ ΚΑΤΑΣΤΟΛΕΩΝ	19
ΠΑΡΑΓΟΝΤΑΣ X	20
3.5 ΔΕΝΤΡΟΕΙΔΗΣ ΔΟΜΗ ΜΙΑΣ ΣΕΙΡΙΑΚΗΣ ΠΕΙΡΑΜΑΤΙΚΗΣ ΔΙΑΔΙΑΚΑΣΙΑΣ ΓΙΑ	
ΑΝΙΧΝΕΥΣΗ ΜΕΤΑΓΡΑΦΙΚΩΝ ΠΑΡΑΓΟΝΤΩΝ	23
3.6 ΤΟ “ΔΙΑΦΟΡΙΚΟ” ΠΕΙΡΑΜΑ EMSA	26
3.7 ΑΝΙΧΝΕΥΣΗ ΜΕΤΑΓΡΑΦΙΚΩΝ ΚΑΤΑΣΤΟΛΕΩΝ ΜΕΤΑΞΥ ΠΕΡΙΣΣΟΤΕΡΩΝ ΤΩΝ	
ΔΥΟ ΚΑΤΑΣΤΑΣΕΩΝ.	27
3.8 ΣΥΝΑΡΤΗΣΕΙΣ ΩΦΕΛΙΜΟΤΗΤΑΣ ΚΑΙ ΚΑΤΩΦΛΙΑ ΥΠΟΕΚΦΡΑΣΙΜΟΤΗΤΑΣ.....	28
3.9 Ο ΡΟΛΟΣ ΤΩΝ ΜΙΚΡΟΣΥΣΤΟΙΧΕΙΩΝ (MICROARRAYS) ΣΤΟ ΣΧΕΔΙΑΣΜΟ	
ΠΕΙΡΑΜΑΤΩΝ ΑΝΙΧΝΕΥΣΗΣ ΜΕΤΑΓΡΑΦΙΚΩΝ ΚΑΤΑΣΤΟΛΕΩΝ.....	31

4 ΤΟ ΠΡΟΒΛΗΜΑ BEHRENS FISHER.....33

4.1 ΑΝΑΛΥΣΗ ΤΟΥ ΚΑΝΟΝΙΚΟΥ ΜΟΝΤΕΛΟΥ ΠΙΘΑΝΟΤΗΤΑΣ	34
4.2 ΤΟ ΠΡΟΒΛΗΜΑ ΤΩΝ «ΔΥΟ ΜΕΣΩΝ ΟΡΩΝ».....	35
ΣΥΧΝΟΤΙΚΗ ΛΥΣΗ ΣΤΟ ΠΡΟΒΛΗΜΑ BEHRENS-FISHER.....	36
BAYESIAN ΛΥΣΗ ΣΤΟ ΠΡΟΒΛΗΜΑ BEHRENS-FISHER	36
4.3 ΥΠΟΛΟΓΙΣΜΟΣ ΟΛΟΚΛΗΡΩΜΑΤΩΝ ΤΗΣ ΚΑΤΑΝΟΜΗΣ BEHRENS FISHER	38

ΣΧΕΣΗ ΤΗΣ ΚΑΤΑΝΟΜΗΣ BEHRENS FISHER ΜΕ ΤΗ ΘΕΩΡΙΑ ΤΩΝ ΥΠΕΡΓΕΩΜΕΤΡΙΚΩΝ ΣΥΝΑΡΤΗΣΕΩΝ	38
ΕΚΦΡΑΣΗ ΤΗΣ ΚΑΤΑΝΟΜΗΣ BEHRENS FISHER ΜΕ ΟΛΟΚΛΗΡΩΜΑΤΑ ΤΥΠΟΥ EULER	39
ΙΣΟΔΥΝΑΜΕΣ ΕΚΦΡΑΣΕΙΣ ΤΗΣ ΚΑΤΑΝΟΜΗΣ BEHRENS FISHER.....	41
4.4 ΑΣΥΜΠΤΩΤΙΚΗ ΑΝΑΛΥΣΗ ΤΩΝ ΟΛΟΚΛΗΡΩΜΑΤΩΝ ΤΗΣ ΚΑΤΑΝΟΜΗΣ BEHRENS FISHER	44
ΑΝΑΛΥΣΗ ΤΗΣ ΕΚΦΡΑΣΗΣ 1	44
ΑΝΑΛΥΣΗ ΤΗΣ ΕΚΦΡΑΣΗΣ 2	46
ΑΝΑΛΥΣΗ ΑΣΥΜΠΤΩΤΙΚΗΣ ΣΥΜΠΕΡΙΦΟΡΑΣ ΣΥΝΑΡΤΗΣΕΩΝ LAURICELLA	47
4.5 ΑΡΙΘΜΗΤΙΚΟΣ ΥΠΟΛΟΓΙΣΜΟΣ ΤΗΣ ΚΑΤΑΝΟΜΗΣ BEHRENS FISHER ΜΕ ΤΗ ΜΕΘΟΔΟ ΤΟΥ ΔΙΠΛΟΥ ΕΚΘΕΤΙΚΟΥ ΜΕΤΑΣΧΗΜΑΤΙΣΜΟΥ	49
ΕΦΑΡΜΟΓΗ ΤΟΥ ΔΙΠΛΟΥ ΕΚΘΕΤΙΚΟΥ ΜΕΤΑΣΧΗΜΑΤΙΣΜΟΥ ΣΤΗ ΚΑΤΑΝΟΜΗ BEHRENS FISHER	49
ΣΥΓΚΡΙΤΙΚΗ ΑΝΑΛΥΣΗ ΜΕΘΟΔΩΝ ΑΡΙΘΜΗΤΙΚΗΣ ΟΛΟΚΛΗΡΩΣΗΣ ΓΙΑ ΤΟΝ ΥΠΟΛΟΓΙΣΜΟ ΠΙΘΑΝΟΤΗΤΩΝ ΒΑΣΙΣΜΕΝΩΝ ΣΤΗΝ ΚΑΤΑΝΟΜΗ BEHRENS-FISHER.....	50
ΣΧΕΔΙΑΣΜΟΣ ΚΑΙ ΥΛΟΠΟΙΗΣΗ ΛΟΓΙΣΜΙΚΟΥ ΓΙΑ ΤΟΝ ΥΠΟΛΟΓΙΣΜΟ ΤΗΣ ΚΑΤΑΝΟΜΗΣ BEHRENS FISHER	62

5 ΑΝΑΛΥΣΗ ΠΕΙΡΑΜΑΤΩΝ ΜΙΚΡΟΣΥΣΤΟΙΧΕΙΩΝ ΓΙΑ ΤΗΝ ΑΝΑΚΑΛΥΨΗ ΜΕΤΑΓΡΑΦΙΚΩΝ ΚΑΤΑΣΤΟΛΕΩΝ67

5.1 ΠΗΓΕΣ ΔΕΔΟΜΕΝΩΝ	67
ΑΝΤΙΣΤΟΙΧΗΣΗ ΙΧΝΗΘΕΤΩΝ ΣΕ ΓΟΝΙΔΙΑ	68
ΕΠΕΞΕΡΓΑΣΙΑ ΕΙΚΟΝΩΝ ΜΙΚΡΟΣΥΣΤΟΙΧΕΙΩΝ	69
ΚΑΝΟΝΙΚΟΠΟΙΗΣΗ ΜΕΤΡΗΣΕΩΝ	69
ΟΜΑΔΟΠΟΙΗΣΗ ΜΕΤΡΗΣΕΩΝ ΣΕ ΑΝΤΑΛΛΑΞΙΜΑ ΔΕΙΓΜΑΤΑ	70
ΣΥΝΔΥΑΣΜΟΣ ΠΕΠΟΙΘΗΣΕΩΝ ΜΕΤΡΗΣΕΩΝ ΜΙΚΡΟΣΥΣΤΟΙΧΕΙΩΝ	71
ΕΠΙΛΟΓΗ ΚΑΤΩΦΛΙΟΥ ΠΙΘΑΝΟΤΗΤΑΣ ΥΠΟΕΚΦΡΑΣΗΣ ΜΕΣΩ ΑΝΑΛΥΣΗΣ ΚΑΜΠΥΛΩΝ ROC	72
ΜΕΘΟΔΟΛΟΓΙΑ ΚΑΙ ΥΛΟΠΟΙΗΣΗ ΑΝΑΛΥΣΗΣ ΚΑΜΠΥΛΩΝ ROC	72
ΣΧΕΔΙΑΣΜΟΣ ΚΑΙ ΥΛΟΠΟΙΗΣΗ ΤΟΥ ΛΟΓΙΣΜΙΚΟΥ ΑΝΑΛΥΣΗΣ ΤΩΝ ΔΕΔΟΜΕΝΩΝ ΜΙΚΡΟΣΥΣΤΟΙΧΕΙΩΝ	74
5.2 ΑΠΟΤΕΛΕΣΜΑΤΑ ΑΝΑΛΥΣΕΩΝ	75
ΕΠΙΛΟΓΗ ΚΑΤΩΦΛΙΟΥ ΥΠΟΕΚΦΡΑΣΙΜΟΤΗΤΑΣ.....	75
ΑΠΟΤΕΛΕΣΜΑΤΑ ΠΕΙΡΑΜΑΤΩΝ ΔΙΑΦΟΡΙΚΗΣ ΕΚΦΡΑΣΗΣ.....	79

6 ΕΚΤΕΛΕΣΗ ΚΑΙ ΑΠΟΤΕΛΕΣΜΑΤΑ ΕΠΙΒΕΒΑΙΩΤΙΚΩΝ ΠΕΙΡΑΜΑΤΩΝ87

6.1 ΜΕΘΟΔΟΛΟΓΙΑ ΠΕΙΡΑΜΑΤΩΝ	87
ΚΥΤΤΑΡΟΚΑΛΛΙΕΡΓΕΙΕΣ.....	87
ΑΠΟΜΟΝΩΣΗ ΠΥΡΗΝΙΚΩΝ ΕΚΧΥΛΙΣΜΑΤΩΝ ΑΠΟ Τ ΛΕΜΦΟΚΥΤΤΑΡΑ.....	87
ΠΕΙΡΑΜΑΤΑ EMSA	88
DNA ΚΛΩΝΟΙ	88
ΔΙΑΜΟΛΥΝΣΕΙΣ.....	88
ΔΟΚΙΜΑΣΙΕΣ ΛΟΥΣΙΦΕΡΑΣΗΣ ΚΑΙ ΑΚΕΤΥΛΑΣΗΣ ΤΗΣ ΧΛΩΡΑΜΦΕΝΙΚΟΛΗΣ	89
ΑΝΤΙΔΡΑΣΕΙΣ ΑΛΥΣΣΙΔΩΤΗΣ ΑΝΤΙΔΡΑΣΗΣ ΠΟΛΥΜΕΡΑΣΗΣ.....	89
6.2 ΠΕΙΡΑΜΑΤΙΚΗ ΕΠΙΒΕΒΑΙΩΣΗ ΤΗΣ ΕΙΔΙΚΗΣ ΕΚΦΡΑΣΗΣ ΤΟΥ ETS-2 ΣΤΑ ΠΑΡΘΕΝΙΚΑ Τ ΛΕΜΦΟΚΥΤΤΑΡΑ.....	89

6.3	Ο ETS-2 ΔΡΑ ΩΣ ΚΑΤΑΣΤΟΛΕΑΣ ΤΟΥ ΥΠΟΚΙΝΗΤΗ ΤΗΣ ΙΝΤΕΡΛΕΥΚΙΝΗΣ –2 ...	91
	ΑΠΟΤΕΛΕΣΜΑΤΑ ΠΕΙΡΑΜΑΤΩΝ EMSA	91
	ΑΠΟΤΕΛΕΣΜΑΤΑ ΠΕΙΡΑΜΑΤΩΝ ΔΙΑΜΟΛΥΝΣΗΣ	92
	ΠΟΣΟΤΙΚΟΠΟΙΗΣΗ ΤΗΣ ΣΧΕΣΗΣ ΙΝΤΕΡΛΕΥΚΙΝΗΣ 2 ΚΑΙ ETS-2 ΣΕ ΦΥΣΙΟΛΟΓΙΚΑ Τ	
	ΛΕΜΦΟΚΥΤΤΑΡΑ	95
6.4	Ο ETS-2 ΔΡΑ ΩΣ ΚΑΤΑΣΤΟΛΕΑΣ ΤΟΥ HIV-LTR.....	96
<u>7</u>	<u>ΣΥΖΗΤΗΣΗ ΚΑΙ ΕΠΙΛΟΓΟΣ.....</u>	<u>100</u>
7.1	ΜΕΘΟΔΟΛΟΓΙΚΟ ΠΛΑΙΣΙΟ ΑΝΙΧΝΕΥΣΗΣ ΜΕΤΑΓΡΑΦΙΚΩΝ ΚΑΤΑΣΤΟΛΕΩΝ....	100
7.2	ΤΟ ΠΡΟΒΛΗΜΑ ΚΑΙ Η ΚΑΤΑΝΟΜΗ BEHRENS FISHER.....	103
7.3	ΠΡΟΕΚΤΑΣΕΙΣ ΤΩΝ ΒΙΟΛΟΓΙΚΩΝ ΕΥΡΗΜΑΤΩΝ	106
<u>8</u>	<u>ΒΙΒΛΙΟΓΡΑΦΙΑ</u>	<u>108</u>

Περίληψη

Σκοπός της παρούσης διατριβής είναι να προτείνει ένα *τυπικό* πλαίσιο για μια τέτοια μεθοδολογία ανάλυσης που να ενσωματώνει ποσοτικά, και λειτουργικά δεδομένα και στοχεύει στον σχεδιασμό πειραμάτων για την επιβεβαίωση («αγνώστων») μεταγραφικών παραγόντων που δεσμεύονται σε ένα λειτουργικά ενεργό μοτίβο στον υποκινητή γονιδίων. Το πλαίσιο αυτό που βασίζεται στη Bayesian πιθανοκρατική θεωρία όπως αυτή θεμελιώνεται στη θεωρία λήψης αποφάσεων και εφαρμόζεται σε ένα πρόβλημα από το ερευνητικό πεδίο της ανοσοβιολογίας, δηλαδή την ανίχνευση μεταγραφικών παραγόντων που ενέχονται στην αρνητική ρύθμιση της γονιδιακής έκφρασης κατά τη διαδικασία ενεργοποίηση των T λεμφοκυττάρων. Η υιοθέτηση του πλαισίου αυτού σμιλεύει μια αυστηρή διαδοχή διεξαγωγής *in vitro* και *in silico* πειραμάτων που ξεκινά από τεχνικές μαζικής ανάλυσης γονιδιακής ρύθμισης, περνά μέσα από βάσεις μεταγραφικών παραγόντων, και μέσω πειραμάτων ηλεκτροφορητικής κινητικότητας (Electromobility Shift Assays) στοχεύεται στην επιβεβαίωση που προσφέρουν τα πειράματα διαμόλυνσης (transfection and reporter assays).

Κατά την τυποποίηση της λογικοφανούς αυτής προσέγγισης ανακύπτει ένα από τα γνωστότερα προβλήματα της εφαρμοσμένης στατιστικής, το πρόβλημα των «δύο μέσων όρων» ή πρόβλημα Behrens-Fisher. Για τη λύση αυτού του προβλήματος προτείνονται νέα τεχνικά μαθηματικά εργαλεία τα οποία αξιοποιήθηκαν για την κατασκευή του αντίστοιχου λογισμικού. Με την εφαρμογή των εργαλείων αυτών στο εφαρμοσμένο ανοσοβιολογικό πρόβλημα προέκυψε μια μη αναμενόμενη σύνδεση μεταξύ δύο φαινομενικά μη συνδεδεμένων συστημάτων: των γονιδίων των κυτταροκινών και του ιού HIV.

Ευχαριστίες

Η παρούσα εργασία δε θα μπορούσε να πραγματοποιηθεί χωρίς τη ενεργό συμμετοχή των ακόλουθων ατόμων τους οποίους ευχαριστώ: τον επιβλέποντα καθηγητή Γ. Νικηφορίδη που πρώτος με εισήγαγε στη Bayesian στατιστική από το επίπεδο του προπτυχιακού φοιτητή και έπειτα, την αναπληρώτρια καθηγήτρια Α. Μουζάκη για τις ενδιαφέρουσες συζητήσεις πάνω στο ανοσοβιολογικό πρόβλημα και την επίβλεψη της στα αρχικά βήματα της εργασίας αυτής (σε επίπεδο διπλωματικής ειδίκευσης), το μεταδιδακτορικό ερευνητή Τάσσο Γεωργακόπουλο και το διδακτορικό φοιτητή Παύλο Αδαμόπουλο για τη διεξαγωγή των επιβεβαιωτικών πειραμάτων στα πλαίσια της συνεργατικής προσπάθειας για τον εντοπισμό μεταγραφικών καταστολέων του υποκινητή της ιντερλευκίνης 2 και του LTR του ιού HIV.

Στη διάρκεια των 3 χρόνων που κράτησε η εργασία αυτή στηρίχθηκα σε ψυχολογικό και επιστημονικό επίπεδο από το στενό μου φίλο Χριστόφορο Καρυσιανό, τον οποίο ευχαριστώ ολόψυχα για τις ζωνχές μας συζητήσεις. Επιπλέον ευχαριστώ τους γονείς μου για την συμπαράσταση τους στη διάρκεια της τελευταίας τριετίας.

1 Εισαγωγή

Οι γενετιστές και οι μοριακοί βιολόγοι εδώ και πολλά χρόνια επέδειξαν ενδιαφέρον για την ποσοτικοποίηση των γονιδίων και των προϊόντων τους ¹. Η υβριδοποίηση νουκλεϊκών οξέων και οι κινητικές μέθοδοι ανάλυσης (Cot and Rot curve analysis) σύντομα αποδείχθηκαν ένα ανεκτίμητο εργαλείο και χρησιμοποιήθηκαν ευρέως για να διαπιστώσουν τα στοιχεία που σήμερα παίρνουμε για δεδομένα. Μέθοδοι μεταφοράς πηκτωμάτων ακολούθησαν αμέσως μετά, για να χρησιμοποιηθούν ως μέσα σύγκρισης του αριθμού και των προϊόντων των γονιδίων. Τεχνικές Northern Blotting και δοκιμασίες S1 νουκλεάσης γρήγορα αναδείχθηκαν στη πρώτη γραμμή έρευνας και ακόμα και σήμερα χρησιμεύουν σαν χρυσά πρότυπα έναντι στις οποίες συγκρίνονται νέες τεχνικές. Ένα σημαντικό άλμα προς τα εμπρός ήταν η εισαγωγή της PCR, η οποία επέτρεψε την ενίσχυση και τον προσδιορισμό της ποσότητας ασθενώς εκφρασμένων γονιδίων. Τα τέλη της δεκατίας του '80 και οι αρχές δεκαετίας του '90 γνωρίζει μια έκρηξη τεχνικών που σχεδιάζονται με σκοπό την ποσοτικοποίηση **περισσότερων** γονιδίων με **τη μεγαλύτερες ακρίβεια** και **πιστότητα** και με **λιγότερο** αρχικό υλικό (DNA ή RNA).

Αλλά η επανάσταση επρόκειτο ακόμα να έρθει το 1995 υπό μορφή μιας δημοσίευσης στο *Science* που καθιέρωσαν τις τεχνικές που απαιτήθηκαν για την αποτύπωση (profiling) της γονιδιακής έκφρασης σε γονιδιωματική κλίμακα ². Η ρομποτική, τεχνολογίες υβριδοποίησης, οπτικές συσκευές ανάλυσης και υπολογιστές χρησιμοποιήθηκαν μαζί για να ελέγξουν ταυτόχρονα την έκφραση χιλιάδων γονιδίων και οι μικροσυστοιχίες (**microarrays**) έγιναν μέρος του οπλοστασίου της μοριακής βιολογίας. Η βιολογία είχε μάθει τελικά να μετρά κατά χιλιάδες, και έγινε γρήγορα έμπειρη στο «άθλημα» αυτό. Ο αρχικός ενθουσιασμός έδωσε τόπο στον υγιή σκεπτικισμό: πώς θα μπορούσαν να χρησιμοποιηθούν τα επίπεδα γονιδίων που μετρώταν τώρα κατά χιλιάδες; Υποστηρίχθηκε ότι μια αλλαγή παραδείγματος είχε επέλθει, και ότι η «προσανατολισμένη προς την υπόθεση» βιολογική επιστήμη πρέπει τώρα να υιοθετήσει μια διερευνητική προσέγγιση. Αλλά τι συνέβαινε εάν ο ερευνητής είχε ήδη την υπόθεση; θα μπορούσε να χρησιμοποιήσει τις δημόσια διαθέσιμες και ογκώδεις βάσεις δεδομένων γονιδιακής έκφρασης για να απαντήσει στο ερώτημα του; Είναι εφικτό *θεωρητικά* και *τεχνικά* να χρησιμοποιήσει τους προαναφερθέντες πόρους για μια post-hoc προσέγγιση;

Στις επόμενες μερικές σελίδες ελπίζουμε να δώσουμε μια θετική απάντηση σε αυτές τις ερωτήσεις εξετάζοντας ένα πρόβλημα από την περιοχή του ανοσοβιολογίας. Πριν από μερικά χρόνια πειραματικά στοιχεία υπέδειξαν τη σημασία μεταγραφικών καταστολέων ³ στην αρνητική ρύθμιση της γονιδιακής έκφρασης κατά τη διαδικασία ενεργοποίησης των T λεμφοκυττάρων. Ένα μοτίβο στη ρυθμιστική περιοχή γονιδίων κυτταροκινών και του ιού HIV προτάθηκε ⁴ σαν θέση δέσμευσης ενός υποθετικού καταστολέα αλλά η σχετική πρωτεΐνη αντιστάθηκε σε όλες τις προσπάθειες του βιοχημικού χαρακτηρισμού της.

Η υπόθεσή μας είναι πραγματικά πολύ απλή. Έχουμε μια περιοχή δέσμευσης, ξέρουμε το χρονικό πρότυπο της έκφρασης της αντίστοιχης πρωτεΐνης αλλά δεν ξέρουμε ποια είναι. Εάν είχαμε πρόσβαση στα χρονικά πρότυπα έκφρασης δυνητικά όλων των γονιδίων θα μπορούσαμε ίσως να εστιάσουμε την αναζήτηση σε μερικούς υποψηφίους που έχουν μια υψηλή πιθανότητα να είναι η πρωτεΐνη που ψάχνουμε; Με ποιο τρόπο θα έπρεπε να αντιστοιχήσουμε την πιθανότητα αυτή, η ακόμα γενικότερα ποιο θα ήταν το πλαίσιο που θα μας επέτρεπε να εντάξουμε δεδομένα μαζικής γονιδιακής έκφρασης, στο πειραματικό σχεδιασμό καθοδηγούμενης από υπόθεσης βιολογικής έρευνας;

Σκοπός της παρούσης διατριβής είναι να προτείνει ένα τυπικό πλαίσιο για μια τέτοια μεθοδολογία ανάλυσης που να ενσωματώνει ποσοτικά, και λειτουργικά δεδομένα και στοχεύει στον σχεδιασμό πειραμάτων για την επιβεβαίωση της ταυτότητας («αγνώστων») μεταγραφικών παραγόντων που δεσμεύονται σε ένα λειτουργικά ενεργό μοτίβο στον υποκινητή γονιδίων. Το πλαίσιο αυτό που βασίζεται στη Bayesian πιθανοκρατική θεωρία όπως αυτή θεμελιώνεται στη θεωρία λήψης αποφάσεων, τυποποιεί μαθηματικά τα τρία βασικά χαρακτηριστικά που πρέπει να ικανοποιεί ο υποθετικός καταστολέας X που δρώντας μέσω ενός μοτίβου Y καταστέλλει ένα ή περισσότερα γονίδια σε μια κατάσταση Z ως προς καταστάσεις $\Xi_1, \Xi_2 \dots \Xi_n$. Συγκεκριμένα ζητούμε από το γονίδιο του X να εκφράζεται στην κατάσταση Z και όχι στην κατάσταση Ξ_i από την πρωτεΐνη X να δεσμεύεται στο Y και τέλος η δέσμευση αυτή να έχει ένα μετρήσιμο λειτουργικό αποτέλεσμα είτε ενεργοποίησης (ενεργοποιητές) είτε καταστολής (καταστολείς) του γονιδίου στον οποίο βρίσκεται το μοτίβο Y . Η υιοθέτηση του πλαισίου αυτού σμιλεύει μια αυστηρή διαδοχή διεξαγωγής *in vitro* και *in silico* πειραμάτων που ξεκινά από τεχνικές μαζικής ανάλυσης γονιδιακής ρύθμισης, περνά μέσα από βάσεις μεταγραφικών παραγόντων, και μέσω πειραμάτων ηλεκτροφορητικής κινητικότητας (Electromobility Shift Assays) στοχεύεται στην επιβεβαίωση που προσφέρουν τα πειράματα διαμόλυνσης (transfection and reporter assays). Κατά την τυποποίηση της λογικοφανούς αυτής προσέγγισης ανακύπτει ένα από τα γνωστότερα προβλήματα της εφαρμοσμένης στατιστικής, το πρόβλημα των «δύο μέσων όρων» ή πρόβλημα Behrens-Fisher. Για τους σκοπούς του πονήματος αυτού υιοθετούμε την προσέγγιση του προβλήματος αυτού από το πρίσμα των κανονικών μοντέλων πεποιθήσεων (normal probability belief models) που οδηγεί στην χρήση της ομώνυμης κατανομής. Το τεχνικό επίτευγμα της διατριβής αυτής είναι η έκφραση πιθανοτήτων βασισμένων στην κατανομή Behrens-Fisher με τη μορφή ολοκληρωμάτων τύπου Euler γνωστών από τη θεωρία των υπεργεωμετρικών συναρτήσεων. Δείχνουμε ότι για τα ολοκληρώματα αυτά υπάρχουν βέλτιστες μέθοδοι αριθμητικές ολοκλήρωσης βασισμένες στο διπλό εκθετικό μετασχηματισμό, που επιτρέπουν τον υπολογισμό χιλιάδων τέτοιων ολοκληρωμάτων εντός δευτερολέπτων καθιστώντας πρακτική την χρήση της προσέγγισης αυτής σε δεδομένα μικροσυστοιχειών.

Τέλος η προσέγγιση αυτή εφαρμόζεται στο αρχικό βασικό πρόβλημα ανοσολογίας σαν παράδειγμα. Δύο βασικά ερωτήματα, αυτό της ύπαρξης καταστολέα του γονιδίου της interleukin-2 και του άγνωστου καταστολέα του LTR του γονιδιώματος του ιού HIV απαντώνται με τη χρήση του προαναφερθέντος πλαισίου και σύμφωνα με την αρχική υπόθεση δείχνεται ότι η ίδια πρωτεΐνη εμπλέκεται και στα δύο φαινόμενα. Τα δεδομένα για την εφαρμογή του προτεινόμενου πλαισίου, προήλθαν από τη βάση δεδομένων μικροσυστοιχειών του Stanford (Stanford Microarray Database) από ένα πείραμα που σχεδιάστηκε, εκτελέστηκε και δημοσιεύτηκε στο τέλος της περασμένης χιλιετίας και το οποίο στα σημεία αναφοράς (control experiment) περιλαμβάνονταν τα δεδομένα αποτύπωσης των καταστάσεων που μελετούσαμε^{5,6}. Παρά τον μικρό αριθμό πειραματικών σημείων (3-10 σύνολα μαζικής αποτύπωσης ανά κατάσταση) και την μη κανονικότητα των μετρήσεων οι Bayesian τεχνικές κατάφεραν να εντοπίσουν τον υποψήφιο με το μικρότερο αριθμό ψευδώς θετικών. Πιστεύουμε ότι η επιτυχής αυτή εφαρμογή ανοίγει τον δρόμο για εφαρμογή αντίστοιχων πλαισίων και σε άλλα συστήματα (όχι κατά ανάγκη σχετιζόμενα με το μεταγραφικό έλεγχο). Υπό την προϋπόθεση ότι το ερευνητικό ζήτημα έχει να κάνει με το βασικό ερώτημα της διαφοράς μεταξύ δύο ή περισσότερων καταστάσεων, το πρόβλημα (και η κατανομή) Behrens-Fisher θα έχουν κεντρικό ρόλο στην επίλυση του πρώτου.

2 Το βιολογικό σύστημα: Μοριακοί μηχανισμοί της ενεργοποίησης των T - κυττάρων & έλεγχος του ιού HIV

2.1 Μηχανισμοί ενεργοποίησης T λεμφοκυττάρων

Τα T λεμφοκύτταρα ωριμάζουν, αποκτούν τα λειτουργικά τους χαρακτηριστικά, και μαθαίνουν την έννοια του «ιδίου» στο θύμο αδένα. Στο θύμο αδένα γίνεται η διαδικασία της θυμικής επιλογής μετά την οποία τα ώριμα λεμφοκύτταρα αφήνουν το θύμο για να κυκλοφορήσουν στην περιφέρεια και στους ιστούς του λεμφικού συστήματος. Τα κύτταρα T εκφράζουν τα μόρια επιφάνειας CD4 ή CD8 κατά τρόπο αμοιβαία αποκλειστικό. Τα CD4+ T-κύτταρα παίζουν σημαντικό ρόλο στην υγιή ανοσοαπόκριση και ο ρόλος τους είναι κεντρικός στην παθογένεση των αυτοάνοσων ασθενειών. Η τροποποίηση της λειτουργίας των κυττάρων αυτών από *ανοσοκατασταλτικά φάρμακα* είναι σημαντική στη θεραπευτική των μεταμοσχεύσεων. Από την άλλη η μείωση του αριθμού των βοηθητικών λεμφοκυττάρων κατά τη διάρκεια της μόλυνσης από τον HIV οδηγεί στη δυσλειτουργία του ανοσοποιητικού συστήματος που εκδηλώνεται με ευκαιριακές λοιμώξεις, κακοήθειες και αυτοάνοσα φαινόμενα. Η λειτουργική σημασία των T λεμφοκυττάρων γίνεται εμφανής μετά την ενεργοποίησή τους ⁷⁻⁹ η οποία συνοδεύεται από την επαγωγή της έκφρασης πολλών γονιδίων σημαντικότερα εκ των οποίων είναι τα γονίδια των κυτταροκινών και των υποδοχέων τους. Η έκφραση της λεμφοτροπικής κυττοκίνης IL-2 παίζει κεντρικό ρόλο στον πολλαπλασιασμό των T λεμφοκυττάρων μέσω αυτοκρινών μηχανισμών. Λόγω των πλειοτροπικών δράσεων της IL-2 στην επιβίωση, διαφοροποίηση και λειτουργία των T λεμφοκυττάρων, η ρύθμιση του γονιδίου της στα T κύτταρα έχει αποτελέσει αντικείμενο έντονης έρευνας. Ένα γενικά παραδεκτό μοντέλο της ενεργοποίησης του γονιδίου της ιντερλευκίνης 2 (META) την ενεργοποίηση των T λεμφοκυττάρων δίνεται στην Εικόνα 1 (μεμβρανικά συμβάντα κατά την ενεργοποίηση των T κυττάρων) και Εικόνα 2 (μετα-μεμβρανικά συμβάντα). Κομβικό ρόλο στην μεταγωγή των σημάτων από την επιφάνεια του κυττάρου στον πυρήνα ¹⁰ παίζει η αύξηση του ενδοκυττάρου Ca^{2+} που οδηγεί στην αύξηση της ενεργότητας της φωσφατάσης καλσινευρίνης ¹¹ (η φωσφατάση 2B). Η καλσινευρίνη αποφωσφορυλώνει αμινοξικά κατάλοιπα σερίνης/θρεονίνης σε ένα μεγάλο αριθμό μεταγραφικών παραγόντων, ένας εκ των οποίων ο πυρηνικός παράγοντας των ενεργοποιημένων T-κυττάρων (NFAT) παίζει κομβικό ρόλο στην ενεργοποίηση του γονιδίου της ιντερλευκίνης 2. Επιπλέον της ιντερλευκίνης 2 ο NFAT συμμετέχει στο μεταγραφικό έλεγχο και άλλων γονιδίων κυττοκινών όπως των ιντερλευκινών 3,4,12 ¹²⁻¹⁹ φλεγμονωδών διαμεσολαβητών (TNF-alpha) ^{20;21}, και αυξητικών παραγόντων (e.g. GMCSF) ^{22;23}.

2.2 Μεταγραφικός έλεγχος του γονιδίου της ιντερλευκίνης – 2

Ο υποκινητής του γονιδίου της ιντερλευκίνης 2, που κάτω από φυσιολογικές συνθήκες αποτελεί τον κύριο αυξητικό παράγοντα των T κυττάρων, καλύπτει μια περιοχή 275 bp στο 5' άκρο του γονιδίου. Στη βραχεία αυτή περιοχή υπάρχουν πιθανές θέσεις δέσμευσης για ποικιλία μεταγραφικών παραγόντων (Εικόνα 3), με πλειοτροπικές δράσεις πάνω στην ενεργότητα του υποκινητή ²⁴.

Ο κύριος θετικός μεταγραφικός παράγοντας (ενεργοποιητής) του γονιδίου της ιντερλευκίνης -2 κατά την ενεργοποίηση των T λεμφοκυττάρων είναι ο NFAT ^{25;26}.

Ο παράγοντας αυτός βρίσκεται σε ανενεργή μορφή στο κυτταρόπλασμα των παρθενικών T λεμφοκυττάρων · η αύξηση του ενδοκυττάριου ασβεστίου σαν αποτέλεσμα της ενεργοποίησης του TCR οδηγεί στην αποφωσφορυλίωση αμινοξέων σερίνης/θρεονίνης στον NFAT. Η μετα-μεταφραστική τροποποίηση του μορίου έχει

Εικόνα 1 Δομή του υποδοχέα των T κυττάρων και σχηματική αναπαράσταση των πρώιμων συμβάντων κατά την ενεργοποίηση των T λεμφοκυττάρων ²⁷. Ο υποδοχέας των T κυττάρων (TCR) αποτελείται από ειδικές προς αντιγόνο αλυσίδες (α και β) και μη ειδικές αλυσίδες (CD3γ, CD3δ, CD3ε and ζ). Η ιδιότητα των μορίων ζ CD3 να μεταγάγουν μεμβρανικά σήματα οφείλεται σε ένα κοινό δομικό μοτίβο γνωστό σαν ITAM (Immunoreceptor Tyrosine-based Activation Motif) που βρίσκεται στο κυτταροπλασματικό τους άκρο (μπλέ οβάλ). Όταν τα κύτταρα βρίσκονται σε ηρεμία οι περιοχές ITAMs είναι είτε ατελώς φωσφορυλιωμένες είτε μη-φωσφορυλιωμένες. Μια πρωτεΐνη με ενεργότητα πρωτεϊνικής κινάσης τυροσίνης, η Lck δεσμεύεται στο κυτταροπλασματικό άκρο του CD4 μορίου. Τα T- κύτταρα ενεργοποιούνται όταν ένα αντιγονοπαρουσιαστικό κύτταρο που φέρει αντιγόνο στην επιφάνεια του HLA μορίου αλληλεπιδρά με τον TCR. Σαν αποτέλεσμα της αλληλεπίδρασης αυτής τα μόρια TCR και CD4 συσσωματώνονται και οι περιοχές ITAMs εντός των CD3 και ζ chains φωσφορυλιώνονται από τις κινάσες Lck και Fyn. Η κινάση ZAP-70 (ή η σχετιζόμενη πρωτεΐνη, Syk) δεσμεύεται ειδικά στις περιοχές ITAMs στις οποίες και τα δύο αμινοξικά κατάλοιπα τυροσίνης έχουν φωσφορυλιωθεί (κόκκινοι κύκλοι). Ακολούθως η ZAP-70 ενεργοποιείται από την Lck, και μέσω αυτής μια πληθώρα άλλων μορίων που ελέγχουν καταρράκτες δεύτερων μηνυμάτων στρατολογούνται στο μεμβρανικό σύμπλοκο του TCR και μέσω αυτών ενεργοποιείται ο μεταγραφικός μηχανισμός. Τροποποιημένο από ²⁸

ως αποτέλεσμα την μετατόπιση του στον πυρήνα των T λεμφοκυττάρων όπου σε συνεργασία με άλλους μεταγραφικούς παράγοντες (c-fos/jun) ενεργοποιεί τα γονίδια στόχους ²⁹⁻³⁴. Η δράση του NFAT διαμεσολαβείται μέσω της δέσμευσης του σε δύο στοιχεία του υποκινητή της ιντερλευκίνης-2 γνωστά σαν ARRE-1/ARRE-2 (Antigen Receptor Response Element 1, 2) περιέχουν τη βασική αλληλουχία GGAAAA και αλληλοεπικαλύπτονται με τα Pu-b_p, Pu-b_d (proximal and distal Purine Box ή PRRE: Purine Rich Response Element)³⁵. Η δέσμευση αυτή σε συνεργασία με τους παράγοντες fos/jun και Octamer ³⁶ οδηγεί στην μεταγραφή του γονιδίου της ιντερλευκίνης-2 κατά την ενεργοποίηση των T κυττάρων. Μέχρι σήμερα έχουν κλωνοποιηθεί τα cDNA τεσσάρων παραγόντων NFAT ^{37,38} από τα οποία μόνο το NFATp (το οποίο ονομάζεται και NFAT-1) και το NFATc φαίνεται να έχουν συγκεκριμένο ρόλο στην επαγωγή γονιδίων λεμφοκινών κατά την ενεργοποίηση των

Τ-λεμφοκυττάρων³⁹.

Εικόνα 2 Η μεταγωγή σήματος μέσω του TCR ενεργοποιεί (a) το μονοπάτι ασβέστιο/calcieneurin/NF-AT και (b) το μονοπάτι MAPK που με τη σειρά του ενεργοποιεί τις κινάσες ERK JNK. Η ενεργοποίηση του TCR αυξάνει την ενεργότητα των Src πρωτεϊνικών κινασών Lck and Fyn, που με τη σειρά τους ενεργοποιούν την Syk πρωτεϊνική κινάση ZAP-70. Η ZAP-70 ενεργοποιεί την φωσφολιπάση C-g (PLC-g) μέσω της οποίας διαμεσολαβείται η αύξηση του ενδοκυττάριου ασβεστίου ($[Ca^{2+}]_i$). Η αύξηση αυτή μέσω της καλμοδουλίνης ωθεί την καλσινευρίνη (πρωτεΐνη με ενεργότητα φωσφατάσης) να αποφωσφορυλιώσει τον μεταγραφικό παράγοντα NF-AT, μεταβολή που οδηγεί στη μετακίνηση του στον πυρήνα. Εκεί, ο NF-AT δεσμεύεται στον υποκινητή της IL-2 και επάγει την μεταγραφή του γονιδίου. Η ZAP-70 φωσφορυλιώνει επίσης την Shc που σχηματίζει ένα μεμβρανικό σύμπλοκο με την Grb-2 και τον παράγοντα ανταλλαγή νουκλεοτιδίων γουανίνης SOS. Η SOS καταλύει την ανταλλαγή GDP για GTP στην p21ras που οδηγεί στην ενεργοποίηση του κλασσικού μονοπατιού MAPK που μέσω των κινασών Raf, MEK-1/2 και ERK-1/2 αυξάνει τη μεταγραφική δραστηριότητα των γονιδίων cFos και JunB genes, τα προϊόντα των οποίων ετεροδιμερίζονται για να φτιάξουν το σύμπλοκο AP-1 που διεγείρει την μεταγραφή της IL-2. Η εναλλακτική οδός της MAPK ενεργοποιείται επίσης κατά τη διέγερση των T κυττάρων. Κατά την οδό αυτή ο TCR ενεργοποιεί την Grb2 που με τη σειρά της αυξάνει την ενεργότητα της p95vav, μια πρωτεΐνη ανταλλαγής νουκλεοτιδίων για τη Rac. Η ενεργοποιημένη Rac-GTP ευοδώνει την σηματοδότηση μέσω MEKK-1, SEK-1/2 και JNK-1/2. Η JNK-1/2 φωσφορυλιώνει τους μεταγραφικούς παράγοντες cJun και ATF-2 αυξάνοντας έτσι την μεταγραφική τους δραστηριότητα και οδηγώντας σε μεταγραφή του cJun μέσω του υποκινητή του γονιδίου cJun.

Στο μοντέλο του θετικού μοντέλου ρύθμισης της ενεργοποίησης του γονιδίου της ιντερλευκίνης-2 όμως δεν μπορεί να καλύψει μια σειρά παρατηρήσεων:

1. πειράματα EMSA σε φυσιολογικά T λεμφοκύτταρα και Ipr CD4-/CD8- θυμοκύτταρα (κανένα από τα οποία δεν εκφράζει φυσιολογικά IL-2) ανέδειξαν ένα ταχέως μετακινούμενο σύμπλοκο πρωτεΐνης – NFAT ιχνηθέτη⁴⁰. Κατά την ενεργοποίηση των φυσιολογικών κυττάρων, το σύμπλοκο αυτό εξαφανίζεται για να αντικατασταθεί από το σύμπλοκο πρωτεΐνης NFAT-NFAT ιχνηθέτη, αλλαγή που συμπίπτει με αύξηση της μεταγραφής του γονιδίου της ιντερλευκίνης –2. Τα ενεργοποιημένα Ipr CD4-/CD8- (που χαρακτηρίζονται από μειωμένη ικανότητα παραγωγής IL-2) δεν αυξάνουν ποτέ το σύμπλοκο NFAT-NFAT ιχνηθέτη στα επίπεδα των φυσιολογικών T κυττάρων και πάντα εκφράζουν το ταχέως μετακινούμενο σύμπλοκο άγνωστης πρωτεΐνης-NFAT ιχνηθέτη. Η εύλογη ερμηνεία της παρατήρησης αυτής είναι ότι η πρωτεΐνη που μετέχει στο ταχέως μετακινούμενο σύμπλοκο καταστέλλει την μεταγραφική δραστηριότητα της ιντερλευκίνης-2 στα Ipr CD4-/CD8- και κατά επέκταση στα φυσιολογικά T κύτταρα.

Εικόνα 3 Σχηματική αναπαράσταση των πιθανών θέσεων δέσμευσης μεταγραφικών παραγόντων στον υποκινητή του γονιδίου της ιντερλευκίνης –2 . Συντομογραφίες : Pu-b_d Pu-b_p : distal/proximal Purine Box, UPS : Upstream Promoter Site, TRE: TPA Response Element, TCE_p TCE_d : proximal, distal T Cell response Element, US1 , US2/3 : Upstream Site 1 2/3, ARRE-1/2: Antigen Receptor response Element ½, Oct-1,2 : Octamer –1,2 , CD28RE : CD28 Response Element, NF-AT: Nuclear Factor of Activated T cells. Σχήμα από : ⁴¹

2. Μια σειρά πειραμάτων στο σύστημα του ωοκυττάρου του βατράχου *X.laavis* έδειξε ότι τα εν ηρεμία παρθενικά T λεμφοκύτταρα περιέχουν ένα μεταγραφικό καταστολέα που μειώνει το βασικό ρυθμό μεταγραφής του γονιδίου της ιντερλευκίνης 2⁴². Στο σύστημα αυτό στον πυρήνα του ωοκυττάρου εγχύονται γονίδια-ανταποκριτές ή έκφραση των οποίων ελέγχεται από υποκινητές (φυσικούς ή με μεταλλάξεις) και στη συνέχεια ελέγχεται η επίδραση πυρηνικών ή κυτταροπλασματικών εκχυλισμάτων από διαφορετικά είδη κυττάρων . Όταν στα πειράματα αυτά χρησιμοποιήθηκαν πυρηνικά εκχυλίσματα παρθενικών T κυττάρων διαπιστώθηκε στα τελευταία υπήρχε μια άγνωστη πρωτεΐνη που κατέστειλε την έκφραση του υποκινητή της IL-2, ιδιότητα διαμεσολαβούμενη πλήρως από το ARRE-2¹. Όταν χρησιμοποιούνταν εκχυλίσματα ενεργοποιημένων T κυττάρων δεν

¹ Για το σκοπό αυτό εγχύθηκαν σε δύο σειρές πειραμάτων ολόκληρος ο υποκινητής της IL-2 με και χωρίς το ARRE-2. Στην πρώτη περίπτωση η ενεργότητα του ενζύμου ανταποκριτή ήταν μικρότερη από τη δεύτερη περίπτωση.

παρατηρούταν (όπως αναμένεται) καταστολή. Πειράματα EMSA με εκχυλίσματα παρθενικών T κυττάρων ανέδειξαν επίσης ένα ταχέως μετακινούμενο σύμπλοκο να δεσμεύεται στο ARRE-2 σε συμφωνία με την παρατήρηση στα lpr CD4-/CD8- λεμφοκύτταρα του ποντικού.

Εικόνα 4 Αλληλουχία του υποκινητή του γονιδίου της ιντερλευκίνης -2 και των αντίστοιχων πιθανών θέσεων δέσμευσης μεταγραφικών παραγόντων. Συνοτομογραφίες : Pu-b_d Pu-b_p : distal/proximal Purine Box, UPS : Upstream Promoter Site, TRE: TPA Response Element, TCE_p TCE_d : proximal, distal T Cell response Element, US1 , US2/3 : Upstream Site 1 2/3, ARRE-1/2: Antigen Receptor response Element 1/2, Oct-1,2 : Octamer -1,2 , CD28RE : CD28 Response Element, NF-AT: Nuclear Factor of Activated T cells. Σχήμα από ⁴³

Οι παραπάνω παρατηρήσεις είναι συμβατές με ένα μοντέλο ρύθμισης του γονιδίου της ιντερλευκίνης-2 κατά το οποίο ένας καταστολέας (πρωτεΐνη με δυνατότητα ειδικής δέσμευσης στο DNA) κρατά το γονίδιο της ιντερλευκίνης-2 κλειστό. Όταν τα κύτταρα ενεργοποιούνται, ο καταστολέας εξαφανίζεται και ένας ενεργοποιητής συντίθεται *de novo*. Όταν τα ενεργοποιημένα κύτταρα μετάπεσουν σε μνημονικά και εξετασθούν για την ύπαρξη του καταστολέα, τότε αυτός δεν αναδεικνύεται. Αντίθετα ο ενεργοποιητής, βρίσκεται σε μια λιγότερο ενεργή μορφή στο κυτταρόπλασμα των μνημονικών κυττάρων και μετά την ενεργοποίηση των τελευταίων μετατοπίζεται στον πυρήνα όπου και ενεργοποιεί την μεταγραφή του γονιδίου της ιντερλευκίνης 2. Μια σειρά πειραμάτων EMSA με αντισώματα εναντίον διαφόρων μεταγραφικών παραγόντων έδειξε ότι το σύμπλοκο του ενεργοποιητή περιέχει τους μεταγραφικούς παράγοντες FRA-2/ JUN-D ⁴⁴ ενώ αντίθετα κανένας από τους παράγοντες αυτούς δεν συμμετείχε στο σύμπλοκο του καταστολέα. Λόγω της παρουσίας μια πιθανής θέσης αναγνώρισης ETS-1² στη θέση ARRE-2, προτάθηκε ότι ο μεταγραφικός αυτός παράγοντας δρα με κατασταλτικό τρόπο δεσμευόμενος σε αυτή τη θέση. Υπερέκφραση antisense ETS-1 mRNA σε κύτταρα Jurkat ⁴⁵ είχε ως αποτέλεσμα την μείωση των επιπέδων της ενδογενούς πρωτεΐνης ETS-1 και την αύξηση της παραγωγής της ιντερλευκίνης-2, φαινόμενο που διαμεσολαβείται από την περιοχή της πρωτεΐνης που δεσμεύεται στο DNA. Τα πειράματα αυτά δε φαίνεται να έχουν

² Οι μεταγραφικοί παράγοντες της οικογένειας ETS (winged helix loop) δεσμεύονται σε θέσεις στον υποκινητή διαφόρων γονιδίων με τη γενική αλληλουχία GGAA

απόλυτη με τη φυσιολογική ρύθμιση της ρύθμισης του γονιδίου της ιντερλευκίνη-2 μια και πειράματα EMSA με αντισώματα εναντίον του παράγοντα αυτού δεν οδήγησαν σε υπερμετακίνηση (bandshift) του συμπλόκου καταστολέα ARRE-2 ιχνηθέτη ⁴⁶.

2.3 Μεταγραφικός έλεγχος του ιού HIV στα παρθενικά T λεμφοκύτταρα

Κύτταρα που φέρουν το CD4 υποδοχέα όπως τα T λεμφοκύτταρα και τα μακροφάγα μολύνονται ειδικά από τον ιό HIV-1. Μετά την εισαγωγή του στο εσωτερικό του κυττάρου το ιικό γονιδίωμα υφίσταται αντίστροφη μεταγραφή σε cDNA. Στα ενεργά T κύτταρα, το προ-ιικό DNA παραμένει στο κυτταρόπλασμα σε γραμμική μορφή · όταν τα κύτταρα διαιρούνται, τότε εισέρχεται στον πυρήνα όπου και ενσωματώνεται στο γονιδίωμα του ξενιστή. Το ενσωματωμένο προ-ιικό DNA μπορεί να παραμείνει σιωπηλό για μήνες ή και χρόνια και τότε η λοίμωξη χαρακτηρίζεται σαν λανθάνουσα. Στα CD4 κύτταρα τα μη ενεργοποιημένα παρθενικά λεμφοκύτταρα (CD45RA) εκφράζουν τον HIV-1 πολύ λιγότερο αποτελεσματικά από τα μνημονικά T κύτταρα (CD45RO), διαδικασία που σχετίζεται με το μειωμένο ρυθμό μεταγραφής από τον ιικό υποκινητή ^{47;48}.

Η γονιδιακή μεταγραφή του HIV-I ρυθμίζεται μέσω της συνεργασίας ιικών και κυτταρικών μεταγραφικών παραγόντων οι οποίοι αλληλεπιδρούν με συγκεκριμένες αλληλουχίες του ιικού γονιδιώματος (long terminal repeats, LTRs) ⁴⁹⁻⁵¹. Το 5' LTR περιέχει περιοχές που παρουσιάζουν ομολογία με τα στοιχεία του υποκινητή του γονιδίου της ιντερλευκίνης-2 (IL-2) καθώς και με στοιχεία του υποκινητή άλλων γονιδίων όπως της ιντερφερόνης-γ (IFN-γ) και του υποδοχέα της IL-2 (IL-2R). Στο 5' LTR του HIV-I έχουν ταυτοποιηθεί οι θέσεις δέσμησης (Εικόνα 5) για αρκετούς μεταγραφικούς παράγοντες (COUP, USF, TCF-1α, SP-1, UBP-1/ LBP-1, UBP2, CTF/NF1) και επιπλέον έχουν ταυτοποιηθεί αλληλουχίες όπου δεσμεύονται μεταγραφικοί παράγοντες οι οποίοι είναι ειδικοί των λεμφοκυττάρων, όπως οι NF-κB, AP-1 και αυτοί της ομάδας NFAT ⁵²⁻⁵⁴. Αλληλεπίδραση αυτών των παραγόντων με πρωτείνες του ιού (όπως Tat, Rev, Nef) ρυθμίζουν την γονιδιακή έκφραση και πολλαπλασιασμό του ^{55;56}. Επιπλέον έχειδειχτεί ότι μία υπομονάδα του NFAT ονομαζόμενη NFATc δεσμεύεται στην περιοχή των κB ρυθμιστικών στοιχείων και συνεργάζεται με τον NF-κB στη μεταγραφική ενεργοποίηση του HIV-I ^{57;58}, συνεργασία που μπορεί να άρει την καταστολή της έκφρασης του ιού που παρατηρείται στα μνημονικά T κύτταρα. Εντός του υποκινητή του ιού HIV το επωνομαζόμενο NRE (Negative Regulatory Enhancer) located που βρίσκεται μεταξύ των βάσεων -340 to -185 ⁵⁹ εμπεριέχει μια περιοχή (-279 to -250) με πολύ μεγάλη ομολογία με το στοιχείο ARRE-2/Pu-bd της IL-2 PRRE ⁶⁰. Αν και αρχικά θεωρήθηκε σαν μοτίβο δέσμησης NFAT, το στοιχείο στο οποίο πραγματικά δεσμεύεται ο NFAT βρίσκεται στις θέσεις -255 and -216. Στον υποκινητή του HIV το στοιχείο PRRE και η θέση δέσμησης του NFAT δεν υπερτίθενται όπως στον υποκινητή της IL-2 αλλά βρίσκονται το ένα δίπλα στο άλλο, και επιπλέον δεν εμφανίζουν ομολογία (Εικόνα 6)

Εικόνα 5 Σχηματική αναπαράσταση του υποκινητή του ιού HIV, που δείχνει τη θέση των ρυθμιστικών στοιχείων σε σχέση με την θέση έναρξης της μεταγραφής. RATS: Repressor Activator Target Sequence (αντιστοιχεί στο HIV- PRRE). Από ⁶¹

HIV-1 RATS -279 AGGCCAATGAAGGAGAGAACAACAGCTTGT -250
 IL-2 PRRE -292 AAGAAAGGAGGAAAAAAGT-GT -273
 HIV-1 NFAT -252 TGTTCACCCCTATFAGCCTGCATGGGATGGAGGACGC -216
 HIV-1 NFκB -108 ACAAGGGACTTTCCGCTGGGGACTTTTCCA - 80

Εικόνα 6 Αλληλουχίες στοιχείων δέσμευσης μεταγραφικών παραγόντων στους υποκινητές του ιού HIV και της IL-2. PRRE : Purine Riche Response Element, NFAT : Nuclear Factor of Activated T cells, NFκB : Nuclear Factor of kappa B, RATS : Repressor Activator Target Sequence.

Η ανταπόκριση της μεταγραφικής δραστηριότητας του ιού HIV στην ενεργοποίηση του κυττάρου ξενιστή, εμφανίζει εντυπωσιακές ομοιότητες με το γονίδιο της ιντερλευκίνης-2. Λόγω της ομοιότητας σε επίπεδο αλληλουχίας με τα αντίστοιχα στοιχεία στον υποκινητή της ιντερλευκίνης -2, οι θέσεις HIV-1 PRRE, όπως και το γειτονικό μοτίβο NFAT θεωρούνται σαν τις πιθανές θέσεις μέσω τις οποίες ο HIV «αντιλαμβάνεται» την κατάσταση ενεργοποίησης των T κυττάρων. Προσπάθειες να αποδώσουν λειτουργική σημασία σε αυτές τις δύο θέσεις κατέλειξαν σε αντικρουόμενα αποτελέσματα. Σε σταθερά διαμολυνσμένες κυτταρικές σειρές T το LTR του HIV-1 ανταποκρίνεται σε μια πληθώρα σημάτων που ενεργοποιούν την πρωτεϊνική κινάση C ⁶². Διαγραφή του στοιχείου PRRE από το LTR είχε ως αποτέλεσμα αυξημένο ρυθμό μεταγραφής σε μια σειρά ανθρώπινου μελανώματος , ενώ ένθεση ενός αντιγράφου του PRRE στ πλαίσιο ενός υβριδικού υποκινητή είχε ως αποτέλεσμα την αύξηση του ρυθμού μεταγραφής, δείχνοντας έτσι ότι το στοιχείο PRRE έχει διττό ρόλο, τόσο θετικό όσο και αρνητικό ⁶⁰. Λόγω των αντικρουόμενων αυτών ευρημάτων, η ρύθμιση του HIV-1 LTR εξετάσθηκε στο σύστημα των ωοκυττάρων του *X. laevis* ⁶³ χρησιμοποιώντας εκχυλίσματα από κύτταρα περιφερικού αίματος ή ομφάλιου λώρου (πηγή πλούσια σε παρθενικά T κύτταρα). Πλασμίδια που περιείχαν ολόκληρο το HIV-1 LTR ή διάφορους υβριδικούς υποκινητές με ενσωματωμένα τα μοτίβα δέσμευσης PRRE, NFAT ή NFκB motifs εγχύονταν στα ωοκύτταρα του *Xenopus*, μαζί με τα προαναφερθέντα κυτταρικά εκχυλίσματα T λεμφοκυττάρων ή χρησιμοποιήθηκαν σε πειράματα διαμόλυνσης σε εν ηρεμία παρθενικά T κύτταρα. Τα αποτελέσματα αυτών των πειραμάτων έδειξαν ότι ο υποκινητής του HIV-1 ελέγχεται με τρόπο παρόμοιο με τον υποκινητή της IL-2. Έτσι πρωτεΐνες από εν ηρεμία παρθενικά T κύτταρα κατέστειλαν τον βασικό ρυθμό μεταγραφής από τον LTR· πρωτεΐνες από τα διεγερμένα κύτταρα μπόρεσαν να άρουν την καταστολή αυτή. Το στοιχείο του LTR που μοιάζει με το IL-2 PRRE έδρασε σαν ενεργοποιητής και καταστολέας και μετονομάστηκε σε RATS (Repressor-Activator Target Sequence).

Πλασμίδιο Εκχυλίσματα	HIV-LTR	HIV-LTR ΔPRRE	Υποκινητής IL- 2	Υποκινητής IL- 2 ΔPRRE
Παρθενικά T κύτταρα	0	+	0	+
Ενεργοποιημένα T κύτταρα	++	++	++	++

Εικόνα 7 Αποτελέσματα πειραμάτων δια-ενεργοποίησης (transactivation) σε ωοκύτταρα βατράχου. Εκχυλίσματα T κυττάρων (παρθενικών ή ενεργοποιημένων) συγχεχύθηκαν με πλασμίδια που έθεταν το ένζυμο ακετυλάση της χλωραμφαινικόλης υπό τον έλεγχο συνθετικών υποκινητών. HIV-LTR : υποκινητής HIV, HIV-1 LTR ΔPRRE: υποκινητής HIV χωρίς το PRRE στοιχείο. Ο ρυθμός μεταγραφής του ενζύμου ανταποκριτή μειώνεται από το στοιχείο PRRE όταν αυτό συνδιαμολύνεται με εκχυλίσματα παρθενικών κυττάρων, υποδεικνύοντας την ύπαρξη μιας πρωτεΐνης με κατασταλτικό ρόλο στη μεταγραφή των υποκινητών που το φέρουν.

2.4 Συνόψιση του βιολογικού ερωτήματος

Ανακεφαλαιώνοντας οι διαθέσιμες πειραματικές ενδείξεις υποδεικνύουν ότι:

1. Ο μεταγραφικός έλεγχος του γονιδίου της ιντερλευκίνης-2 στα παρθενικά T λεμφοκύτταρα εμπεριέχει ένα ενεργό κατασταλτικό μηχανισμό. Ο μηχανισμός αυτός «κρατά» το γονίδιο σιωπηλό στα παρθενικά κύτταρα, και μόνο όταν τα κύτταρα ενεργοποιηθούν άρεται η καταστολή αυτή.
2. Σε επίπεδο υποκινητή η καταστολή αυτή εξαρτάται από το στοιχείο ARRE-2 γνωστό και σαν «άπω» θέση δέσμησης του NFAT ή σαν IL-2 PRRE
3. Η καταστολή διαμεσολαβείται από μια πρωτεΐνη 30-50 kDA που εκφράζεται μόνο στα T παρθενικά κύτταρα ⁶⁴. Η πρωτεΐνη αυτή δεν εκφράζεται στα ενεργοποιημένα T κύτταρα ούτε στα B κύτταρα.
4. Αντίστοιχα ο χαμηλός ρυθμός μεταγραφής του ιού HIV στα παρθενικά T κύτταρα φαίνεται ότι οφείλεται σε μια ενεργό διαδικασία καταστολής που εξαρτάται από την παρουσία ενός στοιχείου PRRE (Purine Rich Response Element) στον υποκινητή του ιού. Το ίδιο στοιχείο, χρησιμεύει σαν ενεργοποιητής μετά τη διέγερση των λεμφοκυττάρων (RATS – Repressor Activator Target Sequence)
5. Η καταστολή της μεταγραφικής δραστηριότητας μέσω του στοιχείου RATS οφείλεται στη δέσμηση μιας μικρής πρωτεΐνης 30-50kDA. Η έκφραση της πρωτεΐνης αυτής περιορίζεται στα παρθενικά T κύτταρα. Συγκεκριμένα δεν εκφράζεται στα ενεργοποιημένα T, παρθενικά B και ενεργοποιημένα B.
6. Οι αλληλουχίες RATS και ARRE-2 φαίνονται ότι δεσμεύονται από την ίδια πρωτεΐνη επειδή :
 - ο τα σύμπλοκα που εμφανίζονται στα πειράματα EMSA με τους δύο ιχνηθέτες είναι πανομοιότυπα
 - ο σε πειράματα ανταγωνισμού (hot – cold competition EMSA) ο κάθε ένας από τους δύο ιχνηθέτες μπορεί να μετατοπίσει τον άλλο όταν χρησιμοποιείται σε περίσσεια.
 - ο Σε πειράματα δια-ενεργοποίησης, η συμπεριφορά των δύο υποκινητών είναι παρόμοια

Επομένως το βασικό ερώτημα που έπρεπε να απαντηθεί ήταν ποιος είναι ο άγνωστος καταστολέας που δεσμευόμενος στα δύο στοιχεία, προκαλεί την καταστολή των;

Για την απάντηση του ερωτήματος αυτού προτείνεται μια μεθοδολογία ανίχνευσης μεταγραφικών καταστολέων που συνδιάζει πειράματα μεταγραφικού ελέγχου καθώς και τεχνικές αποτύπωσης της γονιδιακής έκφρασης (gene expression profiling). Η μεθοδολογία αυτή προκύπτει από την σκοπιά τυπικών μεθόδων (formal methods) πειραματικού σχεδιασμού (experimental design) της Bayesian θεωρίας πιθανοτήτων. Η ανάπτυξη της μεθοδολογίας αυτής, και των απαραίτητων τεχνικών εργαλείων δίνεται στα επόμενα δύο κεφάλαια. Έχοντας αναπτύξει τα απαραίτητα θεωρητικά και τεχνικά εργαλεία, εφαρμόζουμε τη θεωρία σε ένα δημόσια διαθέσιμο σύνολο δεδομένων γονιδιακής αποτύπωσης και καταλήγουμε να προτείνουμε έναν υποψήφιο ο οποίος ελέγχεται με τα κλασσικά πειράματα μοριακής έρευνας που είχαν προβλεφτεί στο προδιαγεγραμμένο πλαίσιο.

3 Μια τυπική μεθοδολογία για την ανακάλυψη μεταγραφικών καταστολέων σταθεράς κατάστασης (steady state transcriptional repressors)

Σε προηγούμενο κεφάλαιο περιγράφηκε το βιολογικό ερώτημα που αποτέλεσε το έναυσμα για την παρούσα εργασία, δηλαδή η αναζήτηση ενός άγνωστου μεταγραφικού καταστολέα στα παρθενικά T λεμφοκύτταρα. Στο κεφάλαιο αυτό θα περιγραφεί μια τυπική μεθοδολογία για την ανακάλυψη μεταγραφικών καταστολέων που συνδιάζει τεχνικές τόσο κλασσικής μοριακής έρευνας μεταγραφικού έλεγχου όσο και τεχνικές ποσοτικοποίησης γονιδιακής έκφρασης. Η μεθοδολογία αυτή προκύπτει από την σκοπιά του σχεδιασμού ενός βέλτιστου πειράματος, πρόβλημα που θεωρείται εντός της Bayesian σχολής της θεωρίας πιθανοτήτων. Προτού προχωρήσουμε στην περιγραφή της μεθοδολογίας κρίνεται σκόπιμη μια γενική ανασκόπηση των συστατικών της μεθοδολογίας αυτής, και συγκεκριμένα των πειραματικών τεχνικών που χρησιμοποιούνται για τη μελέτη μεταγραφικών παραγόντων και των μεθόδων μαζικής γονιδιακής αποτύπωσης με τη βοήθεια μικροσυστοιχειών (microarrays).

3.1 Πειραματική προσέγγιση στη μελέτη μεταγραφικών παραγόντων

Οι μεταγραφικοί παράγοντες είναι πρωτεΐνες με τις ακόλουθες ιδιότητες^{65,66}:

1. δεσμεύονται με ειδικό τρόπο σε βραχείες αλληλουχίες DNA που βρίσκονται στους υποκινητές των γονιδίων, στις περιοχές δηλαδή πριν την κωδικοποιούσα περιοχή ενός γονιδίου που ελέγχει το ρυθμό μεταγραφής του DNA σε mRNA
2. αλληλεπιδρούν άμεσα ή έμμεσα με το μεταγραφικό μηχανισμό, ρυθμίζοντας έτσι το μεταγραφικό επίπεδο των γονιδίων στόχων
3. οι δύο λειτουργίες, δηλαδή της δέσμευσης στο DNA και του ελέγχου του ρυθμού μεταγραφής διαμεσολαβούνται από διαφορετικές περιοχές της πρωτεΐνης
4. η δέσμευση ενός μεταγραφικού παράγοντα στη θέση αναγνώρισης του, μπορεί να καταλήξει είτε σε αύξηση του ρυθμού μεταγραφής του γονιδίου στόχου (ενεργοποιητής) είτε σε ελάττωση του (καταστολέας)
5. ο ίδιος παράγοντας μπορεί να λειτουργεί είτε ως ενεργοποιητής είτε ως καταστολέας σε διαφορετικά γονίδια και σε διαφορετικά είδη κυττάρων. Αυτή η ιδιότητα είναι άμεση συνέπεια του διαχωρισμού μεταξύ των λειτουργιών δέσμευσης στο DNA και αλληλεπίδρασης με το βασικό μεταγραφικό μηχανισμό αλλά και με άλλους μεταγραφικούς παράγοντες, οι θέσεις δέσμευσης των οποίων γειτνιάζουν με τη θέση αναγνώρισης του υπό εξέταση μεταγραφικού παράγοντα. Συνεπώς η πλήρης κατανόηση της θετικής ή αρνητικής επίδρασης ενός μεταγραφικού παράγοντα επιβάλλει τη θεώρηση του *πλαισίου* στο οποίο συμβαίνει το βασικό μηχανιστικό γεγονός της δέσμευσης της πρωτεΐνης αυτής πάνω στο DNA.

Η πλασιοεξαρτώμενη θεώρηση της λειτουργίας των μεταγραφικών παραγόντων προϋποθέτει τη μελέτη των ακόλουθων ιδιοτήτων των:

1. την ενεργότητα δέσμευσης στο DNA
2. το αποτέλεσμα της δέσμευσης στο ρυθμό μεταγραφής των γονιδίων
3. το επίπεδο έκφρασης του μεταγραφικού παράγοντα σε διαφορετικά είδη κυττάρων ή σε διαφορετικές καταστάσεις του ίδιου κυττάρου

Τεχνικές μελέτης ενεργότητας δέσμευσης μεταγραφικών παραγόντων

Οι τεχνικές αυτές αποσκοπούν στο να απαντήσουν στα ακόλουθα ερωτήματα:

1. αν σε δεδομένο είδος κυττάρου *υπάρχει* πρωτεΐνη³ που να δεσμεύεται σε δεδομένη βραχεία αλληλουχία DNA
2. εφόσον η απάντηση στο πρώτο ερώτημα είναι καταφατική *ποια* είναι η άγνωστη αυτή πρωτεΐνη

Η βασική τεχνική μελέτη της ύπαρξης ενεργότητας δέσμευσης σε δεδομένη βραχεία αλληλουχία DNA είναι η δοκιμασία ηλεκτροφορητικής κινητικότητας (EMSA – Electrophoretic Mobility Shift Assay)^{67,68}. Στη δοκιμασία αυτή κυτταρικά εκχυλίσματα επωάζονται με ραδιενεργά σημασμένο δίκλωνο ολιγονουκλεοτιδικό ιχνηθέτη (probe), ο οποίος υποθέτουμε ότι αντιστοιχεί σε θέση δέσμευσης ενός μεταγραφικού παράγοντα. Τα δείγματα στη συνέχεια υποβάλλονται σε ηλεκτροφόρηση σε γέλη πολυ-ακρυλαμίδιου· εφόσον στο δείγμα υπάρχει μια πρωτεΐνη που να αναγνωρίζει την αλληλουχία του ιχνηθέτη, τότε στη γέλη θα εμφανισθούν δύο ζώνες (bands) : μία που αντιστοιχεί στο ελεύθερο ιχνηθέτη και μία ζώνη με μικρότερη κινητικότητα που αντιστοιχεί στο σύμπλοκο πρωτεΐνης – DNA. Αν σε ένα δείγμα διαπιστωθεί ότι υπάρχει μια τέτοια πρωτεΐνη, τότε επανάληψη της αντίδρασης με ειδικό αντίσωμα για ένα συγκεκριμένο υποψήφιο οδηγεί στην εμφάνιση μιας τρίτης ζώνης που αντιστοιχεί στο σύμπλοκο αντίσωμα-πρωτεΐνη-ιχνηθέτη (φαινόμενο υπερμετακίνησης supershift) εφόσον αυτός ο υποψήφιος δεσμεύεται στον ιχνηθέτη. Η ειδικότητα της δοκιμασίας ηλεκτροφορητικής κινητικότητας μπορεί να ελεγχθεί με χρήση περίσσειας μη ραδιενεργού ιχνηθέτη σε ένα εύρος συγκεντρώσεων, και ποσοτικοποίηση (σε cpm) της ελάττωσης της έντασης της ζώνης ιχνηθέτη – πρωτεΐνης · με αυτό τον τρόπο μπορεί να υπολογισθεί κατά προσέγγιση η σταθερά δέσμευσης της πρωτεΐνης στην ακολουθία του DNA. Μια τροποποίηση του πρωτοκόλλου αυτού γνωστή σαν «εξάλειψη» (depletion)⁶⁹, προ-επωάζει το μείγμα της αντίδρασης με το αντίσωμα εναντίον του μεταγραφικού παράγοντα και στη συνέχεια τον απομακρύνει με τη βοήθεια στηλών protein-A sepharose. Στην περίπτωση αυτή, κατά την ηλεκτροφόρηση παρατηρούμε εξάλειψη (ή μείωση της έντασης) της ζώνης που αντιστοιχεί στο σύμπλοκο πρωτεΐνης-ιχνηθέτη. Η μείωση της έντασης της ζώνης ιχνηθέτη-πρωτεΐνης σε πειράματα εξάλειψης επιδέχεται της ίδιας ερμηνείας με το φαινόμενο της υπερμετακίνησης · και στις δύο περιπτώσεις μπορούμε να ταυτοποιήσουμε την άγνωστη πρωτεΐνη που δεσμεύεται στον ιχνηθέτη με την πρωτεΐνη που αναγνωρίζεται από το αντίσωμα που χρησιμοποιήθηκε.

Μελέτη επίδρασης μεταγραφικού παράγοντα στο ρυθμό μεταγραφής

Για τη μελέτη της επίδρασης του μεταγραφικού παράγοντα στο ρυθμό μεταγραφής ενός γονιδίου δεν μπορούμε να βασισθούμε σε πειράματα δέσμευσης λόγω του διαχωρισμού των ενεργοτήτων δέσμευσης στο DNA και αλληλεπίδρασης με το μεταγραφικό μηχανισμό σε επίπεδο περιοχών (domains) των μεταγραφικών παραγόντων. Η εξακρίβωση του αν ένας παράγοντας δρα σαν ενεργοποιητής ή σαν καταστολέας μέσω της δέσμευσης σε μια βραχεία αλληλουχία προϋποθέτει μια τεχνική που να μπορεί να ποσοτικοποιήσει το ρυθμό μεταγραφής ενός γονιδίου στον υποκινητή του οποίου βρίσκεται η αλληλουχία αυτή με έναν εύκολο και ποσοτικό

³ Στο σημείο αυτό χρησιμοποιούμε τον όρο πρωτεΐνη αντί για τον όρο μεταγραφικός παράγοντας, γιατί αυτά τα πειράματα αναφέρονται στην ανίχνευση πρωτεϊνών με την ιδιότητα να δεσμεύονται στο DNA. Εκτός από τους μεταγραφικούς παράγοντες, υπάρχουν και άλλες πρωτεΐνες με αυτή την ιδιότητα π.χ. ιστονικές πρωτεΐνες ή πρωτεΐνες HMG (High Mobility Group)

τρόπο. Για το σκοπό αυτό καταφεύγουμε στις δοκιμασίες γονιδίων ανταπόκρισης (reporter gene assays)^{70;71} που βασίζονται στην ένθεση της θέσης αναγνώρισης του μεταγραφικού παράγοντα στον υποκινητή ενός ενζύμου, την ενεργότητα του οποίου μπορούμε να μετρήσουμε μέσω της αντίδρασης που καταλύει. Οι μεταβολές της ενζυμικής ενεργότητας αντιστοιχούν σε μεταβολή της ποσότητας του ενζύμου που κάτω από συνθήκες σταθεράς κατάστασης αντικατοπτρίζουν μεταβολές στην ποσότητα του αντίστοιχου mRNA · η τελευταία, κάτω από συγκεκριμένες προϋποθέσεις, είναι ευθέως ανάλογη με το ρυθμό μεταγραφής. Για να πληρωθεί η τελευταία συνθήκη πρέπει ο ρυθμός καταστροφής του mRNA του ενζύμου να είναι σταθερός, όπως συμβαίνει π.χ αν χρησιμοποιηθεί ένα ετερόλογο ένζυμο που δεν προέρχεται από τον υπό εξέταση οργανισμό. Συνήθως χρησιμοποιούνται τα ένζυμα της ακετυλάσης της χλωραμφαινικόλης ή της λουσιφεράσης · και τα δύο είναι ένζυμα που δεν προέρχονται από κύτταρα θηλαστικών και καταλύουν αντιδράσεις που μπορούν να ποσοτικοποιηθούν με τη βοήθεια χρωματογραφίας λεπτής στοιβάδας (διαχωρισμός ακετυλιωμένων ραδιοσημασμένων παραγώγων χλωραμφαινικόλης) ή φωτομετρίας (αντίδραση λουσιφεράσης). Αν ο μεταγραφικός παράγοντας δρα σαν ενεργοποιητής, τότε η ένθεση του στοιχείου δέσμευσης του στον υποκινητή του ενζύμου-ανταποκριτή οδηγεί σε αύξηση της ενζυμικής ενεργότητας. Αντίθετα αν ο παράγοντας δρα σαν καταστολέας τότε η ένθεση του στοιχείου αναγνώρισης του θα οδηγεί σε μείωση της ενζυμικής ενεργότητας⁴

Ποσοτικοποίηση επιπέδων μεταγραφικών παραγόντων

Η ποσοτικοποίηση της έκφρασης των γονιδίων μεταγραφικών παραγραφικών παραγόντων μπορεί να γίνει με οποιαδήποτε από τις μεθόδους ανίχνευσης mRNA. Τόσο κλασσικές τεχνικές ποσοτικοποίησης (RT-PCR, Northern blotting) όσο και μαζικές τεχνικές γονιδιακής αποτύπωσης (high throughput gene profiling) μπορούν να παράσχουν δεδομένα για τη διαφορική έκφραση των γονιδίων των μεταγραφικών παραγόντων σε μια ποικιλία καταστάσεων. Γνωρίζοντας ότι ένας παράγοντας εκφράζεται σε ένα δεδομένο είδος κυττάρου μπορούμε να υποθέσουμε ότι συμμετέχει στον έλεγχο της έκφρασης ενός ή περισσότερων γονιδίων, που με τη σειρά τους καθορίζουν το κυτταρικό φαινότυπο. Η αποτύπωση της έκφρασης όλων των μεταγραφικών παραγόντων που υπάρχουν στο γονιδίωμα θα μπορούσε να παράσχει ισχυρές πληροφορίες για τα δίκτυα γονιδιακής ρύθμισης που συμμετέχουν στην διατήρηση ενός δεδομένου φαινότυπου. Τη δυνατότητα της παράλληλης ποσοτικοποίησης μεγάλου αριθμού γονιδίων παρέχει η τεχνολογία των DNA μικροσυστοιχειών (DNA microarrays). Η τεχνική βασίζεται στην υβριδοποίησης συμπληρωματικών μορίων DNA, ένα εκ των οποίων είναι ακινητοποιημένο πάνω σε ένα σταθερό υπόστρωμα (συνήθως γυαλί)^{72;73}. Λόγω του τρόπου κατασκευής των μικροσυστοιχειών σε κάθε θέση της συστοιχείας έχει εκτυπωθεί μόριο στόχος με γνωστή αλληλουχία · με την υβριδοποίηση βιολογικών δειγμάτων τα οποία έχουν σημειωθεί με φθορίζουσες χρωστικές, οι θέσεις αυτές γενούν φθορίζοντα σήματα η ένταση των οποίων είναι ανάλογη με τη συγκέντρωση του αντίστοιχου mRNA στο

⁴ Για τη διαπίστωση του κατασταλτικού ρόλου μιας ακολουθίας που βρίσκεται στον υποκινητή ενός συγκεκριμένου γονιδίου συνήθίζεται να γίνονται δύο τέτοια πειράματα: πρώτα τοποθετείται ολόκληρος ο υποκινητής μπροστά από το γονίδιο ανταπόκρισης και στη συνέχεια γίνεται το πείραμα με τον υποκινητή χωρίς την ολιγονουκλεοτιδική αλληλουχία που εξετάζεται. Επειδή ο παράγοντας που αναγνωρίζει την τελευταία δεν δεσμεύεται στην δεύτερη περίπτωση θα παρατηρούμε αύξηση της ενεργότητας του ενζύμου ανταπόκρισης συγκριτικά με την πρώτη περίπτωση αν ο μεταγραφικός παράγοντας δρα σαν καταστολέας.

βιολογικό δείγμα. Τα σήματα αυτά εξάγονται από την εικόνα που προκύπτει από τη σάρωση της μικροσυστοιχείας με μέθοδο που διαφέρει ανάλογα με την τεχνική σήμανσης⁷⁴⁻⁷⁷. Στη συνέχεια οι μετρήσεις αυτές κανονικοποιούνται,⁷⁸⁻⁸⁵ μια διαδικασία που αποσκοπεί στη μείωση της μεταβλητότητας των μετρήσεων. Οι κανονικοποιημένες μετρήσεις εισαγάγονται σε βάσεις δεδομένων και αξιοποιούνται για την διατύπωση και έλεγχο υποθέσεων σχετικά με τα επίπεδα του mRNA των γονιδίων, μεμονωμένα είτε σε ομάδες. Η όλη διαδικασία διεξαγωγής πειραμάτων μικροσυστοιχειών δίνεται στο ακόλουθο σχήμα:

Εικόνα 8 Ο κύκλος ζωής ενός πειράματος μικροσυστοιχειών

Λόγω της ποσοτικής φύσης των μετρήσεων των επιπέδων mRNA με μικροσυστοιχείες, οι τελευταίες μπορούν να χρησιμοποιηθούν σαν μέθοδος screening για μεταβολές των επιπέδων των πρώτων σε διάφορες καταστάσεις. Μετρήσεις που κρίνονται ενδιαφέροντες (βάση πιθανοκρατικών στατιστικών μοντέλων, τεχνικών μηχανικής ευφυΐας) επιβεβαιώνονται στη συνέχεια με κλασσικές τεχνικές ποσοτικοποίησης^{86;87} και/ή χρησιμεύουν για το σχεδιασμό περαιτέρω in vitro ή in

silico πειραμάτων⁸⁸⁻⁹⁴. Αυτός είναι και ο σκοπός της παρούσας δηλαδή η πρόταση ένα τυπικού (formal) πλαισίου πειραματικού σχεδιασμού, που «χτίζεται» γύρω από μετρήσεις μικροσυστοιχειών και στοχεύει στην ανίχνευση μεταγραφικών καταστολών.

Η περιγραφή του πλαισίου αυτού βασίζεται στη «γλώσσα» της θεωρίας των πιθανοτήτων, ένα μαθηματικό εργαλείο που επιδέχεται διαφορετικές ερμηνείες. Εν συντομία θα περιγράψουμε τις δύο διαφορετικές ερμηνείες (δηλαδή τη συχνοτική και την Bayesian) που χρησιμοποιούνται συνήθως και το ρόλο τους στον πειραματικό σχεδιασμό.

3.2 Προσεγγίσεις στην ερμηνεία πιθανότητων

Η συχνοτική ερμηνεία πιθανότητας

Κλασσικά η πιθανότητα ενός συμβάντος ταυτίζεται με τη *σχετική συχνότητα* του πρώτου είτε σε μια ακολουθία επαναλαμβανόμενων πειραμάτων, είτε σε μια ομάδα (ensemble) «παρόμοιων συστημάτων». Αυτή η θεώρηση είναι γνωστή σαν συχνοτική (frequentist), «ορθόδοξη» ή «δειγματική» πιθανότητα. Ο έλεγχος υποθέσεων και η συμπερασματολογία εντός της συχνοτικής θεωρίας σχετίζεται με το ακόλουθο πρόβλημα:

Έστω μια άγνωστη παράμετρος ενδιαφέροντος δ που ανήκει σε ένα σύνολο Ψ και έστω οι δύο αμοιβαία αποκλειόμενες υποθέσεις αναφορικά με το υποσύνολο του Ψ που ανήκει η δ :

$$\delta \in \Psi_0 \text{ ή } \delta \in \Psi_1$$

όπου με τη συμβολολογία της συνολοθεωρίας έχουμε $\Psi_0 \cup \Psi_1 = \Psi$, $\Psi_0 \cap \Psi_1 = \emptyset$. Θεωρώντας το πρόβλημα της εξαγωγής συμπερασμάτων για το μέσο (average) επίπεδο έκφρασης ενός είδους mRNA σε δύο καταστάσεις, μια πιθανή Ψ_0 μπορούσε να είναι η $\theta < 0$ (δηλαδή ότι το γονίδιο υποεκφράζεται όταν τα λεμφοκύτταρα ενεργοποιούνται) και συνεπώς η «συμπληρωματική» της υπόθεση Ψ_1 δηλώνει ότι $\theta \geq 0$. Στη γλώσσα της κλασσικής στατιστικής θεωρίας αναφερόμαστε στη δήλωση:

$$H_0 : \delta \in \Psi_0 \text{ σαν τη μηδενική υπόθεση}$$

και στην

$$H_1 : \delta \in \Psi_1 \text{ σαν την εναλλακτική υπόθεση}$$

Συνηθίζεται να λέγεται ότι αν αποφασίσουμε να απορρίψουμε την H_0 όταν αυτή είναι αληθής ότι έχουμε κάνει **σφάλμα τύπου I**. Αντίστοιχα αν αποφασίσουμε να δεχτούμε την H_0 όταν αυτή είναι ψευδής τότε έχουμε κάνει **σφάλμα τύπου II**. Η έκβαση μιας στατιστικής δοκιμασίας αποφασίζεται από την *περιοχή απόρριψης* R που ορίζεται σαν το σύνολο όλων αυτών των πειραματικών δεδομένων που εφόσον παρατηρηθούν οδηγούν στην απόρριψη της μηδενικής υπόθεσης. Η επιλογή της περιοχής απόρριψης αντιπροσωπεύει κατά μία έννοια ένα συμβιβασμό μεταξύ των δύο τύπων σφαλμάτων. Μία από τις παρατηρήσεις για τη διαδικασία της στατιστικής συμπερασματολογίας από τη σκοπιά της συχνοτικής θεωρίας είναι ότι αυτή βασίζεται στην πιθανότητα (για διάφορες τιμές της δ) του συνόλου R στο οποίο οι πειραματικές μετρήσεις μπορεί ή όχι να ανήκουν. Η απόρριψη της μηδενικής υπόθεσης βασίζεται όχι μόνο στα δεδομένα (μετρήσεις) που έγιναν αλλά και σε άλλα δεδομένα που ποτέ δεν παρατηρήθηκαν. Αυτή η άποψη της κλασσικής διαδικασίας συμπερασματολογίας έρχεται σε αντίφαση με τις προσδοκίες του χρήστη της θεωρίας εξαγωγής

συμπερασμάτων, ότι δηλαδή τα συμπεράσματα πρέπει να βασίζονται στα διαθέσιμα δεδομένα και μόνο σε αυτά.

Η ερμηνεία των επιπέδων σημαντικότητας και των διαστημάτων εμπιστοσύνης έχει την ίδια αντι-δισαισθητική υφή. Η απόρριψη της μηδενικής υπόθεσης σε επίπεδο σημαντικότητας 5% δεν σημαίνει ότι η υπόθεση έχει πιθανότητα 5% να είναι αληθής. Το μόνο συμπέρασμα που μπορούμε να εξάγουμε είναι ότι αν επαναλαμβάναμε το πείραμα πολλές φορές θα απορρίπταμε την μηδενική υπόθεση μόνο στο 5% όλων των δοκιμασιών. Παρά τα μειονεκτήματα αυτά, η «δειγματική» στατιστική θεωρία αποτελεί την «κλασσική» στατιστική προσέγγιση · ο κάθε ερευνητής μπορεί να διεξαγάγει την έρευνα του με τα εργαλεία της θεωρίας αυτής και μπορεί να αναμένει ότι τα συμπεράσματα θα είναι σωστά τις περισσότερες φορές. Υπάρχουν όμως και περιπτώσεις όπου τα εργαλεία της θεωρίας αυτής αποτυγχάνουν όπως π.χ. αν η συμπερασματολογία πρέπει να διεξαχθεί βάσει πολύ μικρών δειγμάτων ή σε περίπτωση προβλημάτων που δεν μπορούν να ανευρεθούν στατιστικές δοκιμασίες (pivotal statistics) που συναρτώνται από δειγματικές ποσότητες και μόνο. Στις περιπτώσεις αυτές, τα εργαλεία της συχνοτικής θεωρίας θα αποτύχουν κάνοντας επιτακτική την ανάγκη θεώρησης διαφορετικών σχολών πιθανοτήτων, τα αποτελέσματα των οποίων και «σωστά» είναι και επιδέχονται μιας διαισθητικά προφανούς ερμηνείας.

Η Bayesian πιθανοθεωρητική σχολή

Η Bayesian⁵ θεωρία πιθανοτήτων is founded on a much more general definition of probability. Η BPT αντιμετωπίζει την πιθανότητα σαν ένα ποσοτικό μέτρο της πεποίθησης ότι μια συγκεκριμένη δήλωση για τον εξωτερικό κόσμο είναι αληθής, όταν ελλειπή δεδομένα δεν μας επιτρέπουν να καθορίσουμε την αλήθεια της πρώτης με απόλυτη βεβαιότητα. Το μέτρο αυτό ορίζεται με κλίμακα μέτρησης το διάστημα των πραγματικών αριθμών $[0,1]$ όπου το 1 αντιπροσωπεύει την απόλυτη βεβαιότητα για την αλήθεια της πρότασης, ενώ το 0 αντιπροσωπεύει απόλυτη βεβαιότητα για το ψεύδος της. Εντός του Bayesian πλαισίου, το μαθηματικό εργαλείο της θεωρίας των πιθανοτήτων χρησιμοποιείται για την αναγωγή της κοινής λογικής σε αριθμούς και η πιθανότητα μιας πρότασης αντιπροσωπεύει την πεποίθηση του χρήστη της θεωρίας ότι η τελευταία είναι αληθής. Αυτός ο ορισμός έχει μια προφανή σύνδεση με την χρήση του όρου «πιθανότητα» στην καθημερινή, μη-τεχνική ορολογία. Για τη σχολή αυτή, η θεωρία πιθανοτήτων (η οποία μπορεί να οριστεί ανεξάρτητα από την ερμηνεία της) δεν είναι παρά μια μορφή “ποσοτικής επιστημολογίας”, μια αριθμητική κωδικοποίηση της γνώσης του χρήστη της θεωρίας⁹⁵.

Για να διεξαγάγει κανείς επαγωγικό συμπέρασμα εντός της Bayesian θεωρίας, δηλαδή να καταλήξει σε εύλογα συμπεράσματα σε καταστάσεις αβεβαιότητας οφείλει να προχωρήσει σε μια διαδικασία τριών βημάτων: (1) Να καθορίσει της υποθέσεις, τα μοντέλα⁶ τις διαθέσιμες πληροφορίες πριν τη διεξαγωγή του πειράματος καθώς και τα πειραματικά δεδομένα που θα συλλεχθούν. (2) Να χρησιμοποιήσει την γλώσσα των πιθανοτήτων για να αποδώσει a priori πεποιθήσεις στις υπό διερεύνηση υποθέσεις. (3) Να χρησιμοποιήσει τα τεχνικά (μαθηματικά) εργαλεία της θεωρίας πιθανοτήτων για να υπολογίσει τις αριθμητικές τιμές της αλήθειας των υπό διερεύνηση υποθέσεων βάσει των διαθέσιμων δεδομένων. Αυτές αντιπροσωπεύουν τις a posteriori πιθανότητες για τις υποθέσεις αυτές

⁵ Ο όρος Bayesian πρόέρχεται από το όνομα του μαθηματικού και κληρικού Bayes ο οποίος το 18^ο αιώνα διατύπωσε το ομώνυμο θεώρημα.

⁶ Ο ρόλος των μοντέλων στη Bayesian θεωρία είναι αντίστοιχος με το ρόλο των μετρητικών συσκευών στην διεξαγωγή πειραματικών μετρήσεων

Το μαθηματικό εργαλείο για την τροποποίηση της αλήθειας μια υπόθεσης (H) δωθέντων διαθέσιμων δεδομένων (E) και προϋπάρχουσας γνώσης (δηλαδή το πλαίσιο I) δίνεται από το θεώρημα του Bayes

$$P(H | E, I) = \frac{P(H | I) P(E | H, I)}{P(E | I)}$$

Ο όρος στο αριστερό σκέλος της εξίσωσης, $P(H | E, I)$ ονομάζεται a posteriori (ή posterior) πιθανότητα, και εκφράζει την πεποίθηση για την αλήθεια της υπόθεσης έχοντας εξετάσει τις πειραματικές ενδείξεις E εντός του πλαισίου I. Ο όρος $P(H | I)$ είναι η a priori πιθανότητα της H με δεδομένο το πλαίσιο I και μόνο, δηλαδή η πεποίθηση για την αλήθεια της H πριν να εξετάσουμε τα πειραματικά δεδομένα E. Ο όρος $P(E | I)$ ονομάζεται πιθανοφάνεια, και εκφράζει την πιθανότητα για τις τιμές των πειραματικών ενδείξεων θεωρώντας ότι τόσο η υπόθεση H όσο και το πλαίσιο είναι αληθή. Ο παρανομαστής είναι ανεξάρτητος του H, και μπορεί να θεωρηθεί σαν μια σταθερά κανονικοποίησης. Η πλαισιακή πληροφορία I είναι η λογική σύζευξη (τουλάχιστον) όλων των δηλώσεων και δεδομένων που σχετίζονται με τον καθορισμό των $P(H | I)$ και $P(E | I)$. Για λόγους οικονομίας συμβόλων και όταν το πλαίσιο είναι προφανές το I παραλείπεται από τις εξισώσεις.

Εντός της θεωρίας, όλες οι πιθανότητες είναι δεσμευμένες ή εξαρτημένες· εκφράζουν την πεποίθηση μας για την αλήθεια μιας ή περισσότερων υποθέσεων κάνοντας την παραδοχή ότι κάποιες άλλες υποθέσεις είναι αληθείς. Η σωστή εφαρμογή της θεωρίας απαιτεί την δέσμευση των πεποιθήσεων από όλες τις πληροφορίες που είτε άμεσα είτε έμμεσα καθορίζουν την αντίστοιχη πιθανότητα. Η παραβίαση της αρχής αυτής, οδηγεί σε παραδοξότητες και βρίσκεται πίσω από τα διαφορετικά συμπεράσματα που καταλήγουν εναλλακτικές στατιστικές θεωρίες. Το κύριο αντικείμενο της Bayesian ανάλυσης είναι η a posteriori πιθανότητα, η οποία εμπεριέχει όλες τις πληροφορίες που χρειάζεται κανείς για να μιλήσει για μία υπόθεση..

Η Bayesian προσέγγιση στον έλεγχο υποθέσεων είναι πιο άμεση από την συχνοτική. Το μόνο που απαιτείται είναι ο υπολογισμός των a posteriori πιθανοτήτων με βάση το θεώρημα του Bayes. Έτσι στην περίπτωση που μας ενδιαφέρει το αν ένα γονίδιο υποεκφράζεται βάσει δεδομένων τεχνικών ποσοτικοποίησης της γονιδιακής έκφρασης θα έπρεπε να υπολογίσουμε τις δύο πιθανότητες:

- $P(H_0 | E, I)$ (πιθανότητα ένα γονίδιο να υποεκφράζεται βάσει μετρήσεων μικροσυστοιχειών, RT-PCR, Northern Blot κλπ) και
- $P(H_1 | E, I)$ (πιθανότητα ένα γονίδιο να μην υποεκφράζεται βάσει μετρήσεων μικροσυστοιχειών, RT-PCR, Northern Blot κλπ)

Η πηγή των βιολογικών δεδομένων, προϋπάρχοντα πειραματικά δεδομένα, πρωτόκολλα συλλογής δεδομένων αποτελούν την πλαισιακή πληροφορία και καθορίζουν επίσης και τις μαθηματικές μορφές των μοντέλων που χρησιμοποιούνται. Η δήλωση ότι $P(H_0 | E, I) < 5\%$ έχει μια άμεση και διαισθητικά προφανή ερμηνεία εντός της Bayesian διαδικασίας συμπερασμού: Δωθέντων των πειραματικών ενδείξεων και των πλαισιακών (προϋπάρχοντων) δεδομένων η μηδενική υπόθεση έχει μικρότερη από 5% πιθανότητα να είναι αληθής. Εν αντιθέσει το «κλασσικό» (συχνοτικό) πλαίσιο εκφράζει την (αντίστοιχη αλλά όχι ισοδύναμη) δήλωση ότι επαναλαμβάναμε το πείραμα άπειρες φορές θα απορρίπταμε εσφαλμένα την μηδενική υπόθεση σε 5% των υποθετικών αυτών πειραμάτων.

3.3 Bayesian πειραματικός σχεδιασμός

Κατά σχεδιασμό ενός πειράματος συγκεκριμένες αποφάσεις πρέπει να ληφθούν πριν τη συλλογή δεδομένων και η συλλογή αυτών περιορίζεται από τη διαθεσιμότητα

πόρων (οικονομικών, χρονικών, εξοπλισμού κλπ). Επειδή προϋπάρχουσες πληροφορίες είναι συνήθως διαθέσιμες πριν από τη διεξαγωγή του πειράματος και πράγματι συχνά παρακινούν την εκτέλεση του, οι Bayesian μέθοδοι που αποδίδουν κομβικό ρόλο στις πληροφορίες αυτές είναι ιδανικές για την τυποποίηση του πειραματικού σχεδιασμού. Μια επίσημη παρουσίαση του Bayesian πειραματικού σχεδιασμού από την άποψη της θεωρίας λήψης απόφασης δίνεται στη δημιουργική εργασία των Bernardo και Smith ⁹⁶ και μια περίληψη των κύριων σημείων αναπαράγεται κατωτέρω.

Η θεωρία ενδιαφέρεται για την επιλογή του καλύτερου σχεδίου δράσης (δηλ. πείραμα) μεταξύ διάφορων εναλλακτικών λύσεων, μέσω μια διαδικασία που περιλαμβάνει διαδοχικές, αλληλοεξαρτώμενες αποφάσεις.

Πρώτα *επιλέγουμε* ένα πείραμα e και λαμβάνοντας υπόψη τα στοιχεία D που παράγει, προβαίνουμε σε μια *δράση* (action) a , η οποία σχετίζεται με μια κλάση εξαντλητικών (exhaustive) και αμοιβαία ανεξάρτητων *ενδεχομένων* (mutually independent events). Καθένα από αυτά τα ενδεχόμενα οδηγεί στη συνέχεια σε συγκεκριμένες *συνέπειες* (consequences) κάθε μια από τις οποίες έχει τη δικιά της *χρησιμότητα ή ωφελιμότητα* (utility). Σε καθεστώς πλήρους γενικότητας:

1. τα πιθανά σύνολα δεδομένων εξαρτώνται από διενεργηθέν πείραμα,
2. το σύνολο των διαθέσιμων (πιθανών) δράσεων εξαρτάται όχι μόνο από τα πιθανά πείραμα αλλά και από τα αποτελέσματα των
3. οι συνέπειες και οι αντίστοιχες χρησιμότητες συναρτώνται από το συνδυασμό επιλογών πειράματος και δράσης.

Η ύπαρξη ενός *μηδενικού* πειράματος (e_0) δηλ της επιλογής μια δράσης χωρίς τη *φυσική* εκτέλεση οποιουδήποτε πειράματος ολοκληρώνει την προδιαγραφή του προβλήματος. Οι προτιμήσεις για τα αβέβαια σενάρια, ως αποτέλεσμα των επιλογών της δράσης, εξαρτώνται από τη στάση απέναντι στις σχετικές συνέπειες και τις αξιολογήσεις των αντίστοιχων αβεβαιοτήτων. Οι τελευταίες εκφράζονται με όρους πιθανοτήτων των αβέβαιων γεγονότων, ενώ η στάση απέναντι στις συνέπειες κωδικοποιείται με μια *συνάρτηση ωφελιμότητας* (utility function) που παίρνει τις τιμές από μηδέν (αντίστοιχος στη χειρότερη συνέπεια) και το ένα (η καλύτερη συνέπεια).

Μέσα σε αυτό το πλαίσιο το πρόβλημα του πειραματικού σχεδιασμού αναδιατυπώνεται ως ένα πρόβλημα σειριακής λήψης αποφάσεων (sequential decision making problem). Γραφικά αντιπροσωπεύουμε τέτοια προβλήματα υπό τη μορφή ενός δέντρου *απόφασης* που περιλαμβάνει κόμβους απόφασης (όπου η επιλογή μιας δράσης απαιτείται) και κόμβους αβεβαιότητας όπου η έκβαση είναι πέρα από τον έλεγχό μας (Εικόνα 9). Τα περισσότερα πρακτικά προβλήματα πειραματικού σχεδιασμού περιλαμβάνουν πολλαπλές διαδοχικές εκτιμήσεις του είδους απόφαση-γεγονός-συνέπεια, και συνεπώς μπορούν να αναχθούν στην επαναλαμβανόμενη ανάλυση της απλής δομής που παρουσιάζεται στην Εικόνα 9. Μέσα σε ένα τέτοιο πρόβλημα, η βέλτιστη επιλογή της πρώτης απόφασης εξαρτάται από τις βέλτιστες επιλογές στους επόμενους κόμβους. «Σε όρους καθομιλουμένης, δεν μπορούμε χαρακτηριστικά να αποφασίσουμε τι να κάνουμε σήμερα χωρίς να σκεφτούμε πρώτα αυτό που θα κάνουμε αύριο, και αυτό φυσικά συνήθως εξαρτάται από τις πιθανές συνέπειες των σημερινών ενεργειών» ⁹⁶ (σελίδα 59)

Μια τεχνική γνωστή ως *αντίστροφη επαγωγή* (backward induction) που ξεκινά από το τελευταίο στάδιο του προβλήματος και προχωρά αντίστροφα στο πρώτο έχει καθιερωθεί ως απαραίτητο και ικανοποιητικό εργαλείο για τη βέλτιστη λύση προβλημάτων διαδοχικής λήψης αποφάσεων. Κατά τη διάρκεια της αντίστροφης επαγωγής, αρχίζουμε από το τελικό (νιοστό στάδιο) τη *μεγιστοποίηση της*

αναμενόμενης ωφελιμότητας (expected utility): κατόπιν προχωρούμε στη μεγιστοποίηση της αναμενόμενης χρησιμότητας του προηγούμενου βήματος με τη, δωθείσας της βέλτιστης απόφασης του νιοστού βήματος. Στη συνέχεια προχωράμε να λύσουμε το $n-2$ βήματος δωθείσας της βέλτιστης απόφασης του $n-1$ βήματος έως ότου υπολογίζουμε τη βέλτιστη επιλογή στο πρώτο βήμα, μια διαδικασία που είναι επίσης γνωστή σαν δυναμικός προγραμματισμός. Μια τυπική μαθηματική περιγραφή αυτής της διαδικασίας για ένα μονοβηματικό και ένα πολυβηματικό πρόβλημα απόφασης δίνεται στο επόμενο τμήμα, με την πρόθεση να εφαρμοστούν αυτές οι αρχές στο σχέδιο ενός σύνθετου πειράματος για τον προσδιορισμό μεταγραφικών καταστολέων

Εικόνα 9. Ένα δέντρο λήψης αποφάσεων που απεικονίζει κόμβους απόφασης (τετράγωνα) και κόμβους αβεβαιότητας (κύκλοι). Αφότου έχει επιλεγεί μια δράση και ένα συγκεκριμένο αβέβαιο γεγονός έχει εμφανισθεί, ο κλάδος που συνδέει ένα κόμβο απόφασης με ένα κόμβο αβεβαιότητας μας οδηγεί σε μια αντίστοιχη συνέπεια. Ο μεμονωμένος ιθύνων προλογίζει επάγει μια διάταξη των ενεργειών με τον ορισμό μιας σειράς προτιμήσεων μεταξύ των συνεπειών όπως αυτή εκφράζεται μέσω των συναρτήσεων χρησιμότητας.

Εικόνα 10 Δέντρο απόφασης με πολλαπλούς κόμβους απόφασης

3.4 Πειραματική προσέγγιση στην ανίχνευση μεταγραφικών καταστολέων

Στις ακόλουθες παραγράφους θα εφαρμόσουμε τη γενική θεωρία του στο συγκεκριμένο πρόβλημα σχεδίασης ενός πειράματος που στοχεύει στον προσδιορισμό του μεταγραφικού παράγοντα που ρυθμίζει μέσω δέσμευσης σε ένα δεδομένο *cis-acting* μοτίβο την έκφραση του αντίστοιχου υποκινητή βασισμένο στα εξής:

Τυπικά το πρόβλημα ορίζεται ως εξής:

Λαμβάνοντας υπόψη δύο ή περισσότερες καταστάσεις όπου ένα δεδομένο στοιχείο είναι υπεύθυνο για την καταστολή και ενεργοποίηση ενός υποκινητή, σχεδιάστε μια πειραματική διαδικασία για τον εντοπισμό των παραγόντων μεταγραφής αρμόδιων για την καταστολή (ή/και την ενεργοποίηση).⁷

Οι απαραίτητες βασικές πληροφορίες είναι ότι μια τέτοια περιοχή πρέπει να υπάρχει και η θέση της εντός του υποκινητή να μπορεί να ορισθεί με έναν ορισμένο βαθμό εμπιστοσύνης. Οριζόμενο υπό αυτήν τη μορφή το πρόβλημα είναι το αντίστροφο από αυτό που προσπαθούμε να λύσουμε με ένα CHIP (Chromatic Immuno-Precipitation)⁹⁷ στο οποίο αρχίζουμε με έναν γνωστό παράγοντα μεταγραφής και προσπαθούμε να προσδιορίσουμε τις περιοχές στο γονιδίωμα που ο παράγοντας μεταγραφής δεσμεύεται. Αντ' αυτού αρχίζουμε με μια *λειτουργικά* χαρακτηρισμένη περιοχή και προσπαθούμε να συμπεράνουμε τον καταστολέα (repressor). Σχηματικά το σύστημα έχει ως εξής:

	<u>Κατάσταση</u>				
	Ξ_1	Ξ_2	Ξ_3	Ξ_4	Ξ_5
Παράγοντας X	+	-	-	+	-
Υποκινητής με μοτίβο Y	+	+	+	-	+
Έκφραση γονιδίου υποκινούμενων από Y	-	+	+	+	+
Έκφραση γονιδίου MH υποκινούμενων από Y	+/-	+/-	+/-	+/-	+/-

Ο άμεσος τρόπος εντοπισμού του παράγοντα X με καθαρά βιοχημικές τεχνικές θα ξεκινούσε από πυρηνικά εκχυλίσματα από την κατάσταση Ξ_1 τα οποία θα διηθούνταν από στήλες χρωματογραφίας συγγένειας. Οι στήλες θα εκλούονταν και τα πρωτεϊνικά εκχυλίσματα (κλάσματα) θα ελέγχονταν για δέσμευση με το στοιχείο Y με πειράματα EMSA για να εντοπισθεί το κλάσμα που περιέχει τον παράγοντα X. Εφόσον το σωστό κλάσμα εντοπισθεί, η υποθετική πρωτεΐνη X θα πρέπει να απομονωθεί από μια πληθώρα άλλων πρωτεϊνών/πεπτιδικών θραυσμάτων με αντίστοιχες ιδιότητες, μια διαδικασία που περνά από τεχνικές φασματογραφίας μάζας, αλληλούχησης θραυσμάτων κατά EDMAN. Το τελικό αποτέλεσμα είναι βραχείες πεπτιδικές αλληλουχίες που χρησιμοποιούνται για αναζητήσεις στις βάσεις αλληλουχίας πρωτεϊνικών δεδομένων για τον εντοπισμό των αντίστοιχων πρωτεϊνών. Αφού εντοπισθούν οι πιθανές πηγές των πεπτιδικών σημάτων (δηλαδή οι μεταγραφικοί παράγοντες), θα πρέπει να επιβεβαιωθούν ότι πραγματικά παίζουν το ρόλο του παράγοντα X με μια σειρά πειραμάτων δέσμευσης (στο στοιχείο Y) και λειτουργικότητας (διαμολύνσεις – reporter assays). Επιπλέον, η έκφραση των υποψηφίων στις διαφορετικές πειραματικές καταστάσεις $\Xi_1, \Xi_2, \dots, \Xi_n$ θα έπρεπε να μελετηθεί για να διαπιστωθεί αν όντως φυσιολογικά ρυθμίζονται με τρόπο που αναμένεται για τον καταστολέα X. Η όλη διαδικασία είναι αρκετά σύνθετη και ειδικά το αρχικό πειραματικό τμήμα περιέχει βήματα εξαιρετικά χρονοβόρα, πολύπλοκα και

⁷ Αν και η έμφαση από εδώ και στο εξής θα είναι στο φαινόμενο της καταστολής, είναι σημαντικό να τονισθεί πως η όλη συμπερασματολογία είναι συμμετρική και ως προς τη ανίχνευση ενεργοποιητών.

με μικρά ποσοστά επιτυχίας λόγω των χαμηλών επιπέδων των μεταγραφικών παραγόντων σε κυτταρικό επίπεδο και των φυσικοχημικών ιδιοτήτων των που τους καθιστά ιδιαίτερα ευπαθείς στους διάφορους χειρισμούς. Σχηματικά ή όλη διαδικασία μπορεί να αναπαρασταθεί ως:

Εικόνα 11 Κλασσική προσέγγιση στο πρόβλημα εντοπισμού μεταγραφικών παραγόντων

Το πρώτο βήμα (δηλαδή ο εντοπισμός ενός ή περισσότερων πιθανών υποψηφίων με τεχνικές πρωτεϊνικής χημείας) που είναι και το πιο επισφαλές δεν εξασφαλίζει ότι οι απομονωθέντες πρωτεΐνες είναι ο παράγοντας που ψάχνουμε. Για να επιβεβαιωθεί ο πιθανός υποψήφιος θα πρέπει να πληρεί τα κριτήρια που υπονοεί ο παραπάνω πίνακας δηλαδή:

- ο να εκφράζεται διαφορεικά στις καταστάσεις $\Xi_1, \Xi_2, \dots, \Xi_n$ (π.χ. παρθενικά λεμφοκύτταρα, ενεργοποιημένα λεμφοκύτταρα, T, B)
- ο να δεσμεύεται στο μοτίβο Y
- ο να ελέγχει (καταστέλλει) την έκφραση των γονιδίων στους υποκινητές των οποίων βρίσκεται το στοιχείο Y

Εφόσον κάποιο από αυτά τα κριτήρια δεν ικανοποιείται, τότε δεν μπορεί να τεκμηριωθεί η υποψηφιότητα ενός παράγοντα. Π.χ. ένας παράγοντας του οποίου το γονίδιο δεν εκφράζεται στην υπό μελέτη κατάσταση, δεν μπορεί να καταστέλλει γιατί απλά δεν υπάρχει στην κατάσταση. Αν ένας παράγοντας καταστέλλει ενά γονίδιο στον υποκινητή του οποίου υπάρχει το Y αλλά δεν δεσμεύεται στο μοτίβο, η δράση του αυτή δεν μπορεί να αποδοθεί στην αλληλεπίδραση του με το στοιχείο αυτό αλλά οφείλεται σε άλλους έμμεσους μηχανισμούς (π.χ. ο παράγοντας ελέγχει την έκφραση μιας άλλης πρωτεΐνης που αλληλεπιδρά με τον υποκινητή του γονιδίου).

Οποιαδήποτε πειραματική προσέγγιση στο πρόβλημα, θα πρέπει να απαντήσει στα αντίστοιχα ερωτήματα: *υπάρχει, δεσμεύεται, καταστέλλει* τα οποία επιβάλλει η υπόθεση εργασίας. Η κλασσική προσέγγιση στο πρόβλημα είναι μια σύνθετη διαδικασία που χρησιμοποιεί τις διαδικασίες πρωτεϊνικής χημείας για να τεκμηριώσει τις προϋποθέσεις (1) και (2) και λειτουργικές μελέτες (π.χ. προβλήματα διαμόλυνσης) για το (3). Επιπλέον πειράματα συνήθως απαιτούνται για να αποδειχθούν τα (1) και (2) (π.χ. τεχνικές PCR και EMSA χρησιμοποιώντας εκχυλίσματα από τις υπό μελέτη πειραματικές καταστάσεις).

Θεωρητικά θα μπορούσαμε να εντοπίζαμε τον παράγοντα, αν για ΟΛΟΥΣ τους μεταγραφικούς παράγοντες στο γονιδίωμα είχαμε:

- ο ποσοτικές πληροφορίες για τα πρωτεϊνικά επίπεδα ή για το επίπεδο του mRNA στις διάφορες καταστάσεις που μελετούμε

- ο στοιχεία όσον αφορά την περιοχή δέσμευσης των υποψηφίων παραγόντων.
- ο μπορούσαμε να σχεδιάσουμε ένα πείραμα όπου με ελεγχόμενο τρόπο μελετούσαμε την αλληλεπίδραση μεταξύ της πρωτεΐνης και ενός γονιδίου που ελέγχεται από το στοιχείο Y

Λόγω του μεγάλου αριθμού των πειραμάτων που θα απαιτούνται για τη συλλογή των δεδομένων για όλους τους μεταγραφικούς παράγοντες μια ευρετική προσέγγιση στο πρόβλημα που θα μπορούσε να μειώσει δραστικά τον αριθμό των πειραμάτων θα ήταν η ακόλουθη:

1. συλλογή δεδομένων γονιδιακής έκφρασης (ένα μέσω πειραμάτων μαζικής αποτύπωσης είτε πειραμάτων που περιορίζονται σε παράγοντες μεταγραφής) ώστε να αποκλεισθούν εξ αρχής οι μεταγραφικοί παράγοντες που δεν είναι συμβατοί με το πρότυπο έκφρασης που αναμένουμε
2. μια συνάρτησης *ομοιότητας* που σε κάθε μεταγραφικό παράγοντα αντιστοιχούν ένα μετρικό ανάλογο με το αν η αλληλουχία του Y μοιάζει με τη θέση δέσμευσης του παράγοντα όπως αυτή έχει καταγραφεί από άλλα πειράματα στις βάσεις δεδομένων.
3. Ένθεση των παραπάνω σε ένα τυπικό πλαίσιο πειραματικού σχεδιασμού των επιβεβαιωτικών πειραμάτων. Τα τελευταία τα οποία ελέγχουν έναν παράγοντα τη φορά πρέπει να είναι εξαιρετικά απλά σε τεχνικό επίπεδο και να ιδανικά να επιδέχονται διχοτομική ερμηνεία

Η ευρετική προσέγγιση μπορεί να αποδοθεί σχηματικά σαν μια διαδικασία 3 βημάτων· σε κάθε βήμα φιλτράρουμε τους υποψηφίους που δεν ταιριάζουν με τις απαιτήσεις του προβλήματος. Σχηματικά η προσέγγιση μπορεί να αποδοθεί με το διάγραμμα ροής της εικόνας 3.

Σημειώστε ότι αυτή η μεθοδολογία μπορεί μόνο να αποκαλύψει παράγοντες που ρυθμίζουν έναν σταθερό φαινότυπο με την παρουσία ή την απουσία τους και όχι μέσω κάποιου άλλου μηχανισμού (δηλ. διαφορικό μάτισμα του γονιδίου του παράγοντα μεταγραφής, μετα-μεταφραστικές τροποποιήσεις της πρωτεΐνης) επειδή τέτοιοι μηχανισμοί θα σχετίζονται με παροδικές, ενδεχομένως, αλλαγές επιπέδων αντιγράφων. Δωθείσας της σημασίας μεταγραφικών καταστολέων σταθεράς κατάστασης σε πολλά συστήματα (ειδικά αυτά που σχετίζονται με την λειτουργία άνοση), αυτό δεν θεωρείται ως σημαντικός περιορισμός. Το κομβικό σημείο της μεθοδολογίας είναι η χρήση των πιθανοτήτων ως μέτρο του βαθμού βεβαιότητας για την ποσοτική συμπεριφορά του mRNA παράγοντα μεταγραφής. Εάν το επίπεδο ενός παράγοντα μεταξύ των δύο καταστάσεων λειτουργεί ως διακόπτης, τότε η ποσοτική συμπεριφορά του mRNA πρέπει να είναι ένας από τους παράγοντες που ενέχονται στο σχεδιασμό του επιβεβαιωτικού πειράματος(-ων) . Στην πραγματικότητα με τον περιορισμό της προσοχής μας στους παράγοντες που δυνητικά λειτουργούν ως ποσοτικοί διακόπτες προκαλούμε μια λογική εξάρτηση μεταξύ των πειραμάτων αποτύπωσης της γονιδιακής έκφρασης και των επιβεβαιωτικών πειραμάτων · η εξάρτηση αρκεί για να καθορίσει μονοσήμαντα τη στατιστική δομή του προβλήματος και των αντίστοιχων πιθανοκρατικών μέτρων.

Εικόνα 12
Σειριακή

πειραματική διαδικασία ανίχνευσης μεταγραφικού καταστολέα που δεσμεύεται σε γνωστό μοτίβο Y

3.5 Δεντροειδής δομή μιας σειριακής πειραματικής διαδικασίας για ανίχνευση μεταγραφικών παραγόντων

Η αναπαράσταση του υποθετικού πειράματος του σχήματος με τη μορφή ενός δέντρου λήψης αποφάσης προϋποθέτει τον εντοπισμό των κόμβων αβεβαιότητας και των κόμβων απόφασης, τον ορισμό μιας συνάρτησης χρησιμότητας για όλα τα πιθανά ενδεχόμενα και την αντιστοίχιση των πιθανοτήτων στα αβέβαια ενδεχόμενα.

Τα αβέβαια ενδεχόμενα στην περίπτωση μας είναι τα ακόλουθα:

1. το επίπεδο του μεταγραφικού παράγοντα μειώνεται στις καταστάσεις Ξ_1 , Ξ_2 ;
2. μεταγραφικός παράγοντας δεσμεύεται στη θέση Y;
3. ο μεταγραφικός παράγοντας καταστέλλει την γονιδιακή έκφραση μέσω του Y;

Οι κόμβοι απόφασης σχετίζονται με τη διενέργεια των πειραμάτων των οποίων η έκβαση εξαρτάται από τα αβέβαια ενδεχόμενα:

1. γονιδιακή αποτύπωση για την μελέτη των επιπέδων των πρωτεϊνών ή του mRNA των μεταγραφικών παραγόντων

2. EMSA χρησιμοποιώντας εκχυλίσματα που να περιέχουν τον παράγοντα (π.χ. in vitro translation συστήματα μετά από κλωνοποίηση)
3. πειράματα διαμόλυνσης χρησιμοποιώντας το (κλωνοποιημένο) γονίδιο του μεταγραφικού παράγοντα και υβριδικό υποκινητή με τη θέση δέσμευσης Y

Για όλα τα πιθανά ενδεχόμενα ορίζουμε μια συνάρτηση χρησιμότητας διακριτών τιμών που παίρνει την τιμή 1 στην περίπτωση που και τα 3 ενδεχόμενα είναι συμβατά με την υπόθεση εργασίας και μηδέν σε όλες τις υπόλοιπες. Η σημασιολογία της συνάρτησης αυτής είναι αρκετά απλή: προτιμούμε να επιλέξουμε έναν παράγοντα που να εκπληρεί ΟΛΕΣ τις προϋποθέσεις, ενώ αντίθετα δεν επιλέγουμε έναν παράγοντα που δεν πληρεί έστω και ένα από τα κριτήρια. Με βάση τα παραπάνω κατασκευάζουμε το ακόλουθο δέντρο απόφασης:

Εικόνα 13 Δεντροειδές διάγραμμα απόφασης ενός σύνθετου πειράματος που συνδιάζει δεδομένα ποσοτικής βιολογίας, EMSA και διαμολύνσεις. Η αφηρητή είναι το πείραμα μαζικής αποτύπωσης από το οποίο επιλέγουμε τον παράγοντα X για μελέτη· ο παράγοντας αυτός μελετάται με EMSA και με πειράματα διαμόλυνσης (δράση α_X). Η βέλτιστη δράση είναι αυτή που μεγιστοποιεί την αναμενόμενη ωφελιμότητα για το όλο δέντρο, δηλαδή αυτή που μεγιστοποιεί το γινόμενο των πιθανοτήτων των τριών αβέβαιων ενδεχομένων (υποέκφραση, δέσμευση στο στοιχείο και καταστολή). Οι τρεις αποφάσεις έχουν να κάνουν με τη διεξαγωγή των πειραμάτων E: EMSA, C: Co-transfection, SS: Supershift, NSS: Not SuperShift, DN: Downregulation, NDN: Notdownregulation. Σε κάθε κόμβο απόφασης έχει παραληφθεί από το σχήμα η μηδενική δράση (α_0), δηλαδή η μη διενέργεια περαιτέρω πειραμάτων η οποία έχει χρησιμότητα μηδέν (επειδή δεν πρόκειται να προσφέρει όλες τις απαραίτητες πληροφορίες για να απαντηθούν τα τρία ερωτήματα : *υπάρχει, δεσμεύεται, καταστέλλει*

Λόγω της διακριτής φύσης της συνάρτησης χρησιμότητας, η αναμενόμενη μέγιστη χρησιμότητα δίνεται από τη δράση που μεγιστοποιεί το τριπλό γινόμενο των πιθανοτήτων των αβέβαιων ενδεχομένων. Για να το αποδείξουμε αυτό εξετάζουμε τον τύπο που μας δίνει την τιμή της αναμενόμενης χρησιμότητας στο πρώτο βήμα (έχοντας δεδομένα D_i για το προφίλ έκφρασης του παράγοντα):

$$\bar{u}(a_x, D_i) = \sum_{E_j \in \{DN, NDN\}} \bar{u}(a_x, D_i, E_j) P(E_j | D_j)$$

Για την περίπτωση που απεικονίζεται στην Εικόνα 13, αυτή παίρνει τη μορφή⁸:

⁸ Οι πιθανότητες που εμφανίζονται στη σχέση αυτή είναι δεσμευμένες ως προς τα πειραματικά δεδομένα, αλλά για λόγους συντομίας γραφής η εξάρτηση αυτή παραλείπεται

$$P(DN) \times [P(SS) \times (P(R|SS) \times 1 + P(NR|SS) \times 0) + P(NSS) \times (P(R|NSS) \times 0 + P(NR|NSS) \times 0)] + \\ P(NDN) \times [P(SS) \times (P(R|SS) \times 0 + P(NR|SS) \times 0) + P(NSS) \times (P(R|NSS) \times 0 + P(NR|NSS) \times 0)] =$$

$$P(DN) \times P(SS) \times P(R|SS)$$

Λόγω του τρόπου διεξαγωγής του σύνθετου πειράματος, η πιθανότητα του ενδεχόμενου SS είναι ανεξάρτητη της πιθανότητας του ενδεχομένου DN (γιατί στο ιδεατό αυτό πείραμα που χρησιμοποιείται in vitro μετάφραση για την παραγωγή της πρωτεΐνης του παράγοντα το ενδεχόμενο του SS δεν περιορίζεται από την σχετική έλλειψη της αντίστοιχης πρωτεΐνης). Η πιθανότητα καταστολής όμως δεν είναι ανεξάρτητη από την πιθανότητα του SS γιατί και τα δύο ενδεχόμενα είναι λογικά συζευγμένα λόγω της ύπαρξης του στοιχείου Y τόσο στο ιχνηθέτη (probe) όσο και στον υποκινητή του πλασμιδίου που χρησιμοποιείται για το πείραμα γονιδιακής ανταπόκρισης. Αν ο παράγοντας δεν δεσμεύεται στο στοιχείο Y, τότε η πιθανότητα καταστολής είναι διαφορετική απ'ότι αν δεσμεύεται (γιατί η πρωτεΐνη δεν αλληλεπιδρά μέσω του cis-στοιχείου, αλληλεπίδραση που χρησιμεύει για το να κατασταλλεί ο μεταγραφικός μηχανισμός). Η απλοϊκή αυτή ανάλυση δεν σημαίνει ότι τα αβέβαια ενδεχόμενα του SS και R είναι απόλυτα εξαρτημένα μεταξύ τους, αλλά εκφράζει την πεποίθηση ότι αν ένας παράγοντας δεσμεύεται σε ένα στοιχείο, τότε η πιθανότητα να επηρεάζει, και άρα να καταστέλλει, ένα γονίδιο που φέρει στον υποκινητή του το Y είναι μεγαλύτερη απ'ότι αν δεν το φέρει. Χρησιμοποιώντας το πολλαπλασιαστικό νόμο των πιθανοτήτων η παραπάνω σχέση μετατρέπεται σε:

$$P(DN) \times P(R, SS)$$

Δεδομένου ότι το ενιαίο ενδεχόμενο (R,SS) εξαρτάται από το αν ο παράγοντας X δεσμεύεται στο Y με τέτοιο τρόπο ώστε όσο μεγαλύτερη η πιθανότητα ο X να δεσμεύεται στο Y τόσο μεγαλύτερη η πιθανότητα του ενδεχόμενου (R,SS) μπορούμε να μετατρέψουμε τη σχέση σε :

$$P(DN) \times f(P(X \text{ δεσμεύεται στο } Y))$$

Επεκτείνοντας την άγνωστη συνάρτηση f κατά Taylor στο σημείο $P(X \text{ δεσμεύεται στο } Y)=0$, έχουμε την ακόλουθη σχέση:

$$f(0) + f'(0)P(X \text{ δεσμεύεται στο } Y) + \frac{1}{2} f''(0)P(X \text{ δεσμεύεται στο } Y)^2 + O(P(X \text{ δεσμεύεται στο } Y)^3)$$

Κρατώντας την προσέγγιση πρώτης τάξης, μπορούμε να πούμε ότι η αναμενόμενη χρησιμότητα της δράσης a_X είναι ανάλογη της

$$P(DN) \times P(X \text{ δεσμεύεται στο } Y)$$

και ότι η βέλτιστη δράση είναι αυτή που μεγιστοποιεί τη το γινόμενο των δύο αυτών πιθανοτήτων. Η πιθανότητα δέσμευσης του παράγοντα X στο στοιχείο Y είναι τόσο μεγαλύτερη όσο πιο «όμοια» είναι η ακολουθία του Y με ήδη γνωστές πειραματικά επιβεβαιωμένες αλληλουχίες στις οποίες δεσμεύεται ο X, μια πιθανότητα που μπορεί να εκτιμηθεί είτε υποκειμενικά, είτε με πιθανοκρατικά μοντέλα ανίχνευσης προτύπων σε αλληλουχίες⁹⁸⁻¹⁰⁰. Και στις δύο περιπτώσεις, το πλαίσιο παρέχεται από τη Bayesian θεωρία πιθανοτήτων. Σε παρόμοια συμπεράσματα για τον καλύτερο

υποψήφιο προς επιβεβαίωση μπορούμε να καταλήξουμε και με μια διαφορετική πειραματική προσέγγιση που χρησιμοποιεί πειράματα EMSA με εκχυλίσματα από τις διαφορετικές καταστάσεις που εξετάζονται.

3.6 Το “διαφορικό” πείραμα EMSA

Μια διαφορετική πειραματική προσέγγιση θα χρησιμοποιούσε εκχυλίσματα από δύο καταστάσεις, στην μία εκ των οποίων εκφράζεται ο παράγοντας (Ξ_1) ενώ στην άλλη δεν εκφράζεται (Ξ_2) και το στοιχείο μέσω του οποίου διαμεσολαβείται η καταστολή σε ένα πείραμα EMSA. Η επιλογή του αντισώματος εναντίον ενός μεταγραφικού παράγοντα μοντελοποιείται σαν ένα πρόβλημα λήψης απόφασης με την ακόλουθη δομή:

1. ένα σύνολο διαθέσιμων δράσεων, μία από τις οποίες πρέπει να επιλεγεί (στην περίπτωση μας η δράση είναι η απόφαση να χρησιμοποιηθεί το αντίσωμα ενός συγκεκριμένου μεταγραφικού παράγοντα στο πείραμα EMSA), με σκοπό να ανιχνευτούν φαινόμενα υπερμετάθεσης (supershift) μεταξύ των καταστάσεων Ξ_1, Ξ_2
2. για κάθε δράση ένα σύνολο αβέβαιων γεγονότων που περιγράφουν τα πιθανά ενδεχόμενα που σχετίζονται με κάθε μια από τις δράσεις. Στην περίπτωση μας αυτά τα ενδεχόμενα είναι τα ίδια για όλες τις δράσεις, δηλαδή $SS_{\Xi_1}, NSS_{\Xi_2}, NSS_{\Xi_1}, SS_{\Xi_2}, SS_{\Xi_1}, SS_{\Xi_2}, NSS_{\Xi_1}, NSS_{\Xi_2}$ που αντιστοιχούν στα φαινόμενα υπερμετάθεσης (SS) και μη υπερμετάθεσης (NSS)
3. ένα σύνολο συνεπειών που αντιστοιχεί σε καθένα από τα αβέβαια γεγονότα. Υιοθετούμε μια απλή συνάρτηση χρησιμότητας που δίνεται από τις ακόλουθες σχέσεις: $u(SS_{\Xi_1}, NSS_{\Xi_2})=1, u(NSS_{\Xi_1}, SS_{\Xi_2})=0, u(SS_{\Xi_1}, SS_{\Xi_2})=0, u(NSS_{\Xi_1}, NSS_{\Xi_2})=0$. Η ερμηνεία της συνάρτησης αυτής είναι προφανής: μεταξύ όλων των πιθανών εκβάσεων του πειράματος, προτιμούμε απόλυτα εκείνη στην οποία παίρνουμε σήμα υπερμετάθεσης στην κατάσταση εκείνη που αναμένουμε την ύπαρξη του καταστολέα, ενώ δεν παίρνουμε σήμα εκεί που δεν αναμένουμε να υπάρχει ο καταστολέας.

Η δεντροειδής αναπαράσταση του προβλήματος δίνεται στο παρακάτω σχήμα:

Εικόνα 14 Δεντροειδής δομή δέντρου απόφασης ενός διαφορικού πειράματος EMSA. Οι διαθέσιμες δράσεις σχετίζονται με την επιλογή του αντισώματος για την διενέργεια του πειράματος EMSA: α_X επιλογή του παράγοντα X, α_Θ επιλογή του παράγοντα Θ , SS_{Ξ_1} : Supershift στην κατάσταση Ξ_1 , NSS_{Ξ_1} : Non Supershift στην κατάσταση Ξ_1 , SS_{Ξ_2} : Supershift στην κατάσταση Ξ_2 , NSS_{Ξ_2} : Non Supershift στην κατάσταση Ξ_2

Με βάση την παραπάνω συνάρτηση χρησιμότητας, η βέλτιστη δράση (α_X) έχει να κάνει με την μεγιστοποίηση της πιθανότητας του ενδεχόμενου SS_{Ξ_1}, NSS_{Ξ_2} . Η πιθανότητα του ενδεχόμενου αυτού εξαρτάται από δύο ανεξάρτητα ενδεχόμενα: την πιθανότητα ο παράγοντας X να υπάρχει σε διαφορετικές συγκεντρώσεις στις καταστάσεις Ξ_1, Ξ_2 και δεδομένου ότι υπάρχει να δεσμεύεται στο στοιχείο Y που χρησιμοποιείται στο πείραμα EMSA. Συμβολικά:

$$P(SS_{\Xi_1}, NSS_{\Xi_2}) = f(P(\delta_{\Xi_1-\Xi_2} > 0, X \text{ δεσμεύεται στο } Y))$$

Αναλύοντας τη συνάρτηση f κατά Taylor και κρατώντας τον πρώτο όρο, έχουμε ότι η πιθανότητα του ενδεχόμενου SS_{Ξ_1}, NSS_{Ξ_2} είναι ανάλογη του

$$P(\delta_{\Xi_1-\Xi_2} > 0, X \text{ δεσμεύεται στο } Y) = P(\delta_{\Xi_1-\Xi_2} > 0) \times P(X \text{ δεσμεύεται στο } Y \mid \delta_{\Xi_1-\Xi_2} > 0)$$

Μια μηχανιστική ανάλυση του πειράματος, υποδεικνύει ότι τουλάχιστον σε πρώτη προσέγγιση η πιθανότητα δέσμευσης είναι ανεξάρτητη από τη διαφορά των επιπέδων του μεταγραφικού παράγοντα στις δύο καταστάσεις και συνεπώς η δεσμευμένη πιθανότητα στον παραπάνω τύπο μπορεί να αντικατασταθεί από την πιθανότητα του ενδεχομένου « X δεσμεύεται στο Y ». Συνεπώς η βέλτιστη δράση είναι αυτή που μεγιστοποιεί το γινόμενο :

$$P(\delta_{\Xi_1-\Xi_2} > 0) \times P(X \text{ δεσμεύεται στο } Y)$$

Το ενδεχόμενο όμως $\delta_{\Xi_1-\Xi_2} > 0$ είναι ισοδύναμο με το ενδεχόμενο DN του ιδεατού πειράματος. Συνεπώς η βέλτιστη δράση στο πρόβλημα του πειραματικού σχεδιασμού ενός «διαφορικού» πειράματος EMSA είναι ίδια με αυτή ενός ιδεατού πειράματος που θα περιελάμβανε ένα πείραμα ποσοτικοποίησης, ακολουθόμενο από ένα EMSA που θα χρησιμοποιούσε *in vitro* μεταφρασμένο παράγοντα και ένα πείραμα συνδιαμόλυνσης. Οι παραπάνω εξισώσεις εκφράζουν με μαθηματικό τρόπο τη στρατηγική που υπαγορεύει η κοινή λογική, ότι δηλαδή όταν ψάχνει κανείς για μεταγραφικούς καταστολείς η βέλτιστη στρατηγική θα επικεντρωνόταν στους παράγοντες αυτούς που η λειτουργική τους συμπεριφορά είναι συμβατή με το φαινόμενο της υποέκφρασης και των οποίων οι λειτουργικά επιβεβαιωμένες θέσεις δέσμευσης είναι συμβατές με το *cis-acting* στοιχείο που εξετάζεται.

3.7 Ανίχνευση μεταγραφικών καταστολέων μεταξύ περισσοτέρων των δύο καταστάσεων.

Το πρόβλημα της ανίχνευσης των μεταγραφικών καταστολέων όπως εισήχθει σε προηγούμενη παράγραφο είχε να κάνει με την ανίχνευση ενός μεταγραφικού καταστολέα μεταξύ δύο καταστάσεων (π.χ. παρθενικά vs ενεργοποιημένα T λεμφοκύτταρα), στη μία εκ των οποίων ο παράγοντας υπάρχει, ενώ στην άλλη όχι (έστω Ξ_1, Ξ_2). Αν είχαμε δεδομένα που να υποδηλώναν ότι ο άγνωστος καταστολέας ΔΕΝ εκφράζεται σε καταστάσεις $\Xi_3, \Xi_4, \dots, \Xi_v$ πως θα μπορούσαμε να εκμεταλλευτούμε την επιπλέον πληροφορία για την καλύτερη στόχευση του πειράματος;

Ακολουθώντας συλλογισμούς παρόμοιους με αυτούς των προηγούμενων παραγράφων καταλήγουμε στο συμπέρασμα ότι βέλτιστη δράση (δηλαδή η επιλογή ενός παράγοντα για την διεξαγωγή ενός επιβεβαιωτικού πειράματος συνδιαμόλυνσης) είναι αυτή που μεγιστοποιεί την πιθανότητα :

$$P(\delta_{\Xi_1-\Xi_2} > 0, \delta_{\Xi_1-\Xi_3} > 0, \dots, \delta_{\Xi_1-\Xi_v} > 0) \times P(X \text{ δεσμεύεται στο } Y)$$

Διαδοχικές εφαρμογές του θεωρήματος του Bayes μετασχηματίζουν την παραπάνω εξίσωση σε:

$$P(\delta_{\Xi_1-\Xi_2} > 0) \times P(\delta_{\Xi_1-\Xi_3} > 0, \dots, \delta_{\Xi_1-\Xi_v} > 0 \mid \delta_{\Xi_1-\Xi_2} > 0) \times \dots \times P(\delta_{\Xi_1-\Xi_v} > 0 \mid \delta_{\Xi_1-\Xi_2} > 0, \delta_{\Xi_1-\Xi_3} > 0, \dots, \delta_{\Xi_1-\Xi_{v-1}} > 0) \times P(X \text{ δεσμεύεται στο } Y)$$

Μία προσεγγιστική τοποθέτηση γνωστής σαν αφελής Bayes (naïve Bayes) προτείνει την αντικατάσταση των δεσμευμένων πιθανοτήτων στην παραπάνω εξίσωση με μη δεσμευόμενες πιθανότητες οδηγώντας τελικά στην εξίσωση:

$$P(\delta_{\Xi 1-\Xi 2} > 0) \times P(\delta_{\Xi 1-\Xi 3} > 0) \times \dots \times P(\delta_{\Xi 1-\Xi v} > 0) \times P(X \text{ δεσμεύεται στο } Y)$$

Το πρόβλημα λοιπόν απλοποιείται στον υπολογισμό του γινομένου των πιθανοτήτων για κάθε μια από τις συγκρίσεις μεταξύ δύο καταστάσεων.

3.8 Συναρτήσεις ωφελιμότητας και κατώφλια υποεκφρασιμότητας.

Η υιοθέτηση της συνάρτησης χρησιμότητας των προηγούμενων παραγράφων, υπαγορεύει τη βέλτιστη πειραματική σχεδίαση σαν εκείνη που μεγιστοποιεί το γινόμενο της πιθανότητας υποέκφρασης επί της πιθανότητας δέσμευσης του παράγοντα στο στοιχείο Y . Αν και σε εννοιολογικό επίπεδο η συνάρτηση αυτή είναι εξαιρετικά χρήσιμη για την ψηλάφηση των συστατικών του προβλήματος, η εφαρμογή της σε πρακτικό επίπεδο δημιουργεί ορισμένα προβλήματα. Παραδείγματος χάρη έστω οι ακόλουθες (πολύ πιθανές) καταστάσεις:

	Παράγοντας Α	Παράγοντας Β
$P(\delta_{\Xi 1-\Xi 2} > 0)$	0.100	0.900
$P(X \text{ δεσμεύεται στο } Y)$	0.920	Άγνωστη θέση δέσμευσης
Συνάρτηση χρησιμότητας	0.092	???

και

	Παράγοντας Α	Παράγοντας Β
$P(\delta_{\Xi 1-\Xi 2} > 0)$	0.100	0.900
$P(X \text{ δεσμεύεται στο } Y)$	0.920	0.100
Συνάρτηση χρησιμότητας	0.092	0.090

Στην πρώτη περίπτωση το πλαίσιο δεν μπορεί να εφαρμοσθεί, ενώ στη δεύτερη περίπτωση θα οδηγούμαστε να ελέγξουμε τον παράγοντα Α (ο οποίος λόγω της ποσοτικής συμπεριφοράς του, είναι περισσότερο πιθανό να είναι ενεργοποιητής παρά καταστολέας) πριν από τον παράγοντα Β. Με άλλα λόγια η υιοθέτηση της συνάρτησης χρησιμότητας που προαναφέρθηκε οδηγεί στα ακόλουθα πρακτικά προβλήματα:

1. δεν μπορεί να εφαρμοσθεί στην περίπτωση μεταγραφικών παράγοντων για τους οποίους δεν υπάρχουν δεδομένα δέσμευσης. Εξαιτίας του προγράμματος του Ανθρώπινου Γονιδιώματος και άλλων προγραμμάτων αλληλούχησης, οι διαθέσιμες βιολογικές βάσεις δεδομένων περιέχουν πληθώρα μεταγραφικών παραγόντων για τα οποία, ενώ γνωρίζουμε την αλληλουχία (και άρα μπορούμε να χρησιμοποιήσουμε σαν ιχνηθέτες σε πειράματα), δεν ξέρουμε τη θέση δέσμευσης τους. Λόγω της απουσίας τέτοιων στοιχείων ο υπολογισμός της πιθανότητας δέσμευσης σε ένα δεδομένο στοιχείο είναι αδύνατος.
2. μπορεί να οδηγήσει σε λογικά παράδοξα, καθώς μια μεγάλη πιθανότητα δέσμευσης μπορεί να αντισταθμίσει μια μικρή πιθανότητα υποέκφρασης, οδηγώντας στη διεξαγωγή πειραμάτων με παράγοντες με μικρή πιθανότητα υποέκφρασης. Ακραία τέτοια περίπτωση μπορεί να θεωρηθεί εκείνη στην

οποία ο παράγοντας **A** είναι γνωστός ενεργοποιητής του στοιχείου **Y**, ενώ ο πραγματικός καταστολέας **B** ανήκει σε οικογένεια με πολύ ευρεία ειδικότητα των θέσεων δέσμευσης της. Στην τελευταία περίπτωση αν δεν γνωρίζουμε ένα πολύ μεγάλο αριθμό θέσεων δέσμευσης, ο υπολογισμός της πιθανότητας δέσμευσης του **B** στο στοιχείο **Y** θα είναι ψεύτικα χαμηλός και η τυφλή εφαρμογή του πλαισίου θα οδηγήσει στο να προτείνουμε ένα γνωστό ενεργοποιητή σαν καταστολέα που δρα μέσω του στοιχείου **Y**.

Τα προβλήματα αυτά ανακύπτουν λόγω του συμμετρικού τρόπου με τον οποίο η συνάρτηση χρησιμότητας αντιμετωπίζει τις απαντήσεις στα ερωτήματα «υπάρχει;» και «δεσμεύεται;», συμμετρία η οποία δεν δικαιολογείται λόγω:

1. ασύμμετρης διαθεσιμότητας όγκου δεδομένων έκφρασης και δεδομένων δέσμευσης (δηλαδή λειτουργίας), τέτοια που η πιθανότητα υποέκφρασης είναι υπολογίσιμη για πολύ περισσότερους μεταγραφικούς παράγοντες απ'ότι η πιθανότητα δέσμευσης
2. διαφορετικής φύσης/αξιοπιστίας των εργαλείων / μεθόδων που παρέχουν τα δεδομένα για τον υπολογισμό των αντίστοιχων πιθανοτήτων. Η πιθανότητα υποέκφρασης θα βασισθεί σε μετρήσεις με πολύ καλά επικυρωμένες τεχνικές μοριακής βιολογίας, ενώ ο υπολογισμός της πιθανότητας δέσμευσης θα γίνει με *υπολογιστικές τεχνικές* που δεν έχουν φθάσει στο ίδιο επίπεδο ωριμότητας.
3. ακόμα και στην περίπτωση που είχαμε εξίσου αξιόπιστο τρόπο υπολογισμού των δύο πιθανοτήτων, τα δύο ενδεχόμενα δηλαδή της διαφορικής έκφρασης και δέσμευσης, ΔΕΝ μπορεί να είναι λογικά «ανταλλάξιμα» αν θέλουμε να αποφύγουμε λογικά παράδοξα όπως εκείνο της προηγούμενης παραγράφου. Με άλλα λόγια δωθέντος δύο παραγόντων **A**, **B** με αντίστροφη σχέση πιθανοτήτων υποέκφρασης – δέσμευσης μεγαλύτερο βάρος πρέπει να δοθεί στον παράγοντα με την μεγαλύτερη πιθανότητα υποέκφρασης.

Η διόρθωση των αδυναμιών του πλαισίου είναι εφικτή εφόσον καταφέρουμε να λάβουμε υπόψη την ασυμμετρία μεταξύ των ενδεχομένων διαφορικής έκφρασης και δέσμευσης, και να την αναπαραστήσουμε στην ανάλυση με την υιοθέτηση της κατάλληλης συνάρτησης χρησιμότητας. Τα γενικά χαρακτηριστικά αυτής της συνάρτησης είναι τα ακόλουθα δύο:

- ο λαμβάνει υπόψη ότι θα υπάρχουν μεταγραφικοί παράγοντες για τους οποίους ο υπολογισμός της πιθανότητας δέσμευσης είναι αδύνατος και δεν τους «τιμωρεί» γι' αυτό. Από την άλλη πριμοδοτεί παράγοντες με μεγάλη πιθανότητα δέσμευσης στο υπό εξέταση μοτίβο
- ο δίνει μεγαλύτερη έμφαση στην πιθανότητα υποέκφρασης από την πιθανότητα δέσμευσης · μεγάλες πιθανότητες υποέκφρασης πριμοδοτούνται σε σχέση με μεγάλες πιθανότητες δέσμευσης, ενώ πολύ μικρές πιθανότητες υποέκφρασης δεν μπορούν να αντισταθμισθούν από μεγάλες πιθανότητες δέσμευσης

Συνεπώς η αναπαράσταση της ασυμμετρίας των δύο ενδεχομένων στο πεδίο που πρέπει να λυθεί το πρόβλημα ανίχνευσης μεταγραφικών καταστολέων οδηγεί στην αναζήτηση μορφών συναρτήσεων χρησιμότητας που είτε να εξαρτώνται από τις πιθανότητες υποέκφρασης και δέσμευσης (εφόσον η τελευταία μπορεί να υπολογισθεί), ή να συναρτώνται μόνο από την πιθανότητα υποέκφρασης, εφόσον η πιθανότητα δέσμευσης δεν μπορεί να υπολογισθεί. Η πιο απλή συνάρτηση

χρησιμότητας που καλύπτει και τις δύο περιπτώσεις είναι αυτή που παίρνει τη μορφή γινομένου δύο συναρτήσεων δηλαδή:

$$u(\mathbf{P}(\delta_{\Xi 1-\Xi 2} > 0), \mathbf{P}(\mathbf{X} \text{ δεσμεύεται στο } \mathbf{Y})) = f(\mathbf{P}(\delta_{\Xi 1-\Xi 2} > 0)) \times g(c), c \in \{0,1\} \text{ με}$$

1. $g(0) = h(\mathbf{P}(\mathbf{X} \text{ δεσμεύεται στο } \mathbf{Y})) \in [0,1]$ εφόσον η πιθανότητα δέσμευσης μπορεί να υπολογισθεί και
2. $g(1) = c_1, c_1 \in [0,1]$ αν η πιθανότητα δέσμευσης δεν υπολογίζεται (π.χ λόγω ελλειπών δεδομένων).

Για να πληρωθεί η δεύτερη «προδιαγραφή» για τη συνάρτηση χρησιμότητας πρέπει οι συναρτήσεις f, g να ικανοποιούν τις σχέσεις:

3. $f(x) > c_1$ και $f(x) > g(x) \wedge f(x) > g(1-x)$ για $x \rightarrow 1$
4. $f(x) < c_1$ και $f(x) < g(x) \wedge f(x) < g(1-x)$ για $x \rightarrow 0$

Η συνθήκη 3 εκφράζει μαθηματικά την απαίτηση το πλαίσιο ανίχνευσης να ευνοεί μεταγραφικούς παράγοντες για τους οποίους η πιθανότητα υποέκφρασης (δηλαδή το x) είναι μεγάλη. Τότε η συνάρτηση χρησιμότητας επικυριαρχείται από τη συνιστώσα της που εξαρτάται από την πιθανότητα υποέκφρασης. Η συνάρτηση 4 εκφράζει μαθηματικά την αντίστροφη απαίτηση: μικρές πιθανότητες υποέκφρασης ($x \rightarrow 0$) να μην μπορούν να αντισταθμισθούν από μεγάλες πιθανότητες δέσμευσης ($1-x \rightarrow 1$). Ο συνδιασμός των παραπάνων σχέσεων οδηγεί τελικά στην αναζήτηση μορφών που ικανοποιούν τη σύνθετη συνθήκη:

$$f(0) < h(0) \leq h(1) < f(1) \text{ ή } f(0) < c_1 < f(1)$$

επιπλέον των συνθηκών 3 και 4.

Η πιο απλή περίπτωση που ικανοποιεί τις ανωτέρω σχέσεις προκύπτει όταν η συνάρτηση f παίρνει τη μορφή ενός διχοτομικού νόμου δηλαδή :

$$f(x) = 1 \text{ για } x > C_p \text{ και } f(x) = 0 \text{ } x \leq C$$

χωρίς να επιβάλλονται περιορισμοί στη μορφή της συνάρτησης h .

Αναφερόμενοι στην κατάσταση που απεικονίζεται στην Εικόνα 13 για όλα τα πιθανά ενδεχόμενα ορίζουμε μια συνάρτηση χρησιμότητας που παίρνει την τιμή $f(\mathbf{P}(\delta_{\Xi 1-\Xi 2} > 0)) \times g(c)$ στην περίπτωση που και τα 3 ενδεχόμενα είναι συμβατά με την υπόθεση εργασίας και μηδέν σε όλες τις υπόλοιπες. Σε τεχνικό επίπεδο ανάλυσης η συνάρτηση χρησιμότητας που ορίζεται σαν το γινόμενο ενός διχοτομικού νόμου και μιας φραγμένης συνάρτησης στο διάστημα $[0,1]$, επιβάλλει ένα υψηλοπερατό φίλτρο (high – pass filter) : παράγοντες των οποίων η πιθανότητα υποέκφρασης είναι μικρότερη από το κατώφλι C_p δεν λαμβάνονται υπόψη στο σχεδιασμό πειραμάτων υποέκφρασης.

Για όλους τους υπόλοιπους παράγοντες η βέλτιστη πειραματική σχεδίαση θα είναι αυτή που θα μεγιστοποιεί:

1. το γινόμενο της πιθανότητας υποέκφρασης επί την πιθανότητα δέσμευσης επί μια αυθαίρετη συνάρτηση της πιθανότητας υποέκφρασης σε περίπτωση που αυτή υπολογίζεται
2. το γινόμενο της πιθανότητας υποέκφρασης επί μια σταθερά στο διάστημα $[0,1]$ εφόσον η πιθανότητα υποέκφρασης δεν μπορεί να υπολογισθεί.

Αν η συνάρτηση h παίρνει τη μορφή σταθεράς ίση με 1, και ερμηνεύσουμε την τιμή της c_1 σαν πιθανότητα δέσμευσης, τότε τα αποτελέσματα που λαμβάνονται από την παραπάνω ανάλυση θα είναι ταυτόσημα με το πλαίσιο που παρουσιάστηκε στην Εικόνα 13 με τις εξής διαφορές:

- ο παράγοντες για τους οποίους δεν υπάρχουν δεδομένα δέσμευσης αντιστοιχούνται σε μια αυθαίρετη πιθανότητα δέσμευσης
- η μεγιστοποίηση του γινομένου των πιθανοτήτων γίνεται στο χώρο $[C_p, 1] \times [0, 1]$ αντί για το χώρο $[0, 1]^2$

Έχοντας τροποποιήσει το πλαίσιο με την εισαγωγή δύο επιπλέον σταθερών, προκύπτει ερώτημα: Με ποια κριτήρια πρέπει να γίνει η επιλογή της σταθεράς c_1 , και του κατωφλίου C_p ; Όσον αφορά την απάντηση στο πρώτο ερώτημα μπορούμε να προδιαγράψουμε δύο διαφορετικές προσεγγίσεις: 1) την περίπτωση που η τιμή αυτή επιλέγεται να είναι ίση με 1 και αντιστοιχεί σε μια μη συντηρητική τοποθέτηση κατά την οποία ο χρήστης του πλαισίου αυτού προτίθεται να «δοκιμάσει» όλους τους παράγοντες για τους οποίους δεν έχει λειτουργικά δεδομένα στοχεύοντας έτσι σε μια «νέα» ανακάλυψη 2) στην περίπτωση που η σταθερά αυτή είναι πολύ μικρή ή ίση με μηδέν ο ερευνητής ακολουθεί μια συντηρητική γραμμή που στοχεύει στον εμπλουτισμό των πληροφοριών για ήδη μελετημένους μεταγραφικούς παράγοντες.

Η επιλογή της σταθεράς C_p μπορεί να γίνει με περισσότερο «αντικειμενικό» τρόπο αξιοποιώντας τα διάφορα επίπεδα πολλαπλότητας των μετρήσεων μικροσυστοιχειών για τη διεξαγωγή πειραμάτων δειγματοληψίας (re-sampling), μεθοδολογία που αναλύεται στο κεφάλαιο επεξεργασίας των μετρήσεων μικροσυστοιχειών που χρησιμοποιήθηκαν στην παρούσα εργασία.

3.9 Ο ρόλος των μικροσυστοιχειών (microarrays) στο σχεδιασμό πειραμάτων ανίχνευσης μεταγραφικών καταστολέων

Και στις δύο παραπάνω περιπτώσεις εναλλακτικών πειραματικών προσεγγίσεων είδαμε ότι η βέλτιστη δράση, δηλαδή η επιλογή ενός μεταγραφικού παράγοντα σαν το δυνητικό καταστολέα που δρα μέσω ενός στοιχείου Y , εξαρτάται από την πιθανότητα υποέκφρασης της αντίστοιχης πρωτεΐνης μεταξύ των καταστάσεων Ξ_1, Ξ_2 . Εφ'όσον κάποιος είχε τη δυνατότητα να υπολογίσει την πιθανότητα αυτή ξεχωριστά για όλους τους μεταγραφικούς παράγοντες ενός οργανισμού, θα μπορούσε να πραγματοποιήσει τη μεγιστοποίηση της αναμενόμενης ωφελιμότητας χρησιμοποιώντας επιπλέον δεδομένα για τις θέσεις δέσμευσης των παραγόντων αυτών. Ένα πείραμα μαζικής αποτύπωσης (high throughput profiling) είτε σε επίπεδο πρωτεΐνης είτε σε επίπεδο mRNA (θεωρώντας ότι μεταβολές στα επίπεδα του mRNA σε συνθήκες σταθεράς κατάστασης αντανακλούν μεταβολές σε επίπεδο πρωτεΐνης), παρέχει αυτή τη δυνατότητα. Στην περίπτωση που δεδομένα δέσμευσης δεν είναι διαθέσιμα, ο ρόλος των πειραμάτων μαζικής αποτύπωσης καθίσταται ακόμα πιο σημαντικός · στην περίπτωση αυτή, το μόνο διαθέσιμο δεδομένο για το αν ένας παράγοντας παίζει το ρόλο του καταστολέα είναι η λειτουργική του συμπεριφορά, όπως αυτή αποτυπώνεται στην πιθανότητα υποέκφρασης. Η παράλληλη φύση των πειραμάτων αυτών καθιστά εφικτή την σειριακή προσέγγιση στο πρόβλημα του σχεδιασμού

πειραμάτων ανίχνευσης μεταγραφικών καταστολέων με το μικρότερο αριθμό πειραμάτων, αποτέλεσμα το οποίο είναι εφικτό με το να ελέγχει κανείς τους μεταγραφικούς παράγοντες με σειρά φθίνουσας πιθανότητας υποέκφρασης . Με την εισαγωγή της τεχνολογίας μικροσυστοιχειών, το πρόβλημα μετασχηματίζεται στο πρόβλημα του υπολογισμού της πιθανότητας υποέκφρασης των επιπέδων των mRNA βασισμένη σε πειράματα γονιδιακής αποτύπωσης στις καταστάσεις $\Xi_1, \dots \Xi_n$. Το πρόβλημα του υπολογισμού της πιθανότητας της διαφοράς δύο δειγμάτων παρατηρήσεων είναι γνωστό σαν πρόβλημα των δύο μέσων όρων ή πρόβλημα Behrens-Fisher. Μια καινοτόμα τεχνική προσέγγιση στον υπολογισμό των αντιστοιχών πιθανοτήτων θα εξεταστεί στο επόμενο κεφάλαιο.

4 Το πρόβλημα Behrens Fisher

Σε προηγούμενο κεφάλαιο είδαμε ότι η βέλτιστη πειραματική σχεδίαση στο πρόβλημα του εντοπισμού μεταγραφικών καταστολέων, σχετίζεται με την μεγιστοποίηση της συνάρτησης που ορίζεται σαν γινόμενο της πιθανότητας υποέκφρασης (down-regulation) των επιπέδων ενός μεταγραφικού παράγοντα και της πιθανότητας να δεσμεύεται ο παράγοντας στο υπό εξέταση μοτίβο. Το κεφάλαιο αυτό έχει σαν θέμα τον υπολογισμό πιθανοτήτων που σχετίζονται με το φαινόμενο της υποέκφρασης.

Δωθίσων δύο καταστάσεων Ξ_1, Ξ_2 θεωρούμε ότι το συμβάν της υποέκφρασης για τον μεταγραφικό παράγοντα X είναι ταυτόσημο με το συμβάν $\delta = L_{\Xi_2} - L_{\Xi_1} < C$ όπου C ένα κατώφλι «σημαντικότητας» που χρησιμοποιείται για να ελέγξει την αυστηρότητα της αναζήτησης διαφοράς μεταξύ των επιπέδων L_{Ξ_2}, L_{Ξ_1} . Το κατώφλι είναι εντελώς αυθαίρετο: κατα αναλογία με τα πειράματα Northern blotting όπου ένας λόγος μεταβολής συγκεντρώσεων μεταξύ 2-3 χρησιμοποιείται για να διαχωρίσει τα διαφορετικώς εκφρασμένα από τα μη γονίδια θα μπορούσαμε να ορίσουμε την τιμή του C να είναι 2, 3, 5 ή οποιαδήποτε άλλη τιμή θέλουμε. Ανεξάρτητα από τη συγκεκριμένη τιμή που λαμβάνει το C , ο ορισμός του επιτρέπει την κατάτμηση της πεποίθησης για τη διαφορά των συγκεντρώσεων δ σε δύο περιοχές που αντιστοιχούμε στα ενδεχόμενα της υποέκφρασης και μη υποέκφρασης αντίστοιχα (DN και NDN για συντομία) όπως δείχνει το ακόλουθο σχήμα:

Εικόνα 15 Κατάτμηση της πεποίθησης για την τιμή της διαφοράς μεταξύ των επιπέδων ενός γονιδίου σε δύο καταστάσεις. Το κατώφλι C διαχωρίζει την συνάρτηση πυκνότητας πιθανότητας σε δύο περιοχές που αντιστοιχούν στα ενδεχόμενα της υπο-έκφρασης και μη υπο-έκφρασης. Η πιθανότητα υποέκφρασης $P(DN)$ δίνεται από το εμβαδό της σκιασμένης περιοχής κάτω από την καμπύλη, ενώ η πιθανότητα μη υποέκφρασης από το εμβαδό της λευκής περιοχής.

Η πιθανότητα του ενδεχομένου υποέκφρασης βρίσκεται με τον υπολογισμό του ολοκληρώματος :

$$I = \int_{-\infty}^C pdf(\delta) d\delta$$

όπου $pdf(\delta)$ η συνάρτηση πυκνότητας πιθανότητας που εκφράζει την πεποίθησή μας για την τιμή του δ .

Η ουσιαστική παραδοχή έχει να κάνει με την παραμετρική μορφή που παίρνει η πεποίθησή μας για τα επίπεδα του mRNA σε κάθε μία από τις διαφορετικές καταστάσεις $\Xi_1, \dots, \Xi_v$. Συγκεκριμένα δεχόμαστε ότι:

- ο κλίμακα που εκφράζουμε τις πεποιθήσεις μας είναι συμμετρική ως προς το μηδέν
- απουσία μεγάλου αριθμού μετρήσεων, θεωρούμε ότι η συνάρτηση πυκνότητας πιθανότητας έχει πεπερασμένη ορμή πρώτης και δεύτερης τάξης.

Κάτω από αυτές τις γενικές προϋποθέσεις η συνάρτηση που εκφράζει τις πεποιθήσεις μας για τα επίπεδα του mRNA στις διάφορες καταστάσεις είναι το κανονικό μοντέλο πιθανότητας (ΚΜΠ, normal probability model). Η χρήση του κανονικού μοντέλου πιθανότητας παρέχει τη δυνατότητα χρήσης ισχυρών αναλυτικών εργαλείων για τον υπολογισμό των αντίστοιχων πιθανοτήτων και επιπλέον αποτελεί την αφετηρία για λιγότερο παραμετρικά μοντέλα (μοντέλα μεικτών κατανομών, mixture probability models). Σε επιστημονολογικό επίπεδο επιτρέπει τη διάκριση μεταξύ της Bayesian και συχνοτικής προσέγγισεων στην πιθανότητα, διάκριση με σαφείς χρηστικές προεκτάσεις.

4.1 Ανάλυση του κανονικού μοντέλου πιθανότητας

Η ανάλυση πεποιθήσεων με βάση το ΚΜΠ αναλύεται σε πλήθος κειμένων αναφοράς και μόνο περιληπτικά θα αναφερθεί εδώ. Το βασικό εργαλείο για την έκφραση των πεποιθήσεων μας από πειραματικά μετρήσεις εντός του ΚΜΠ είναι η a posteriori συνάρτηση πυκνότητας πιθανότητας. Η ακριβής μορφή της πιθανότητας εξαρτάται από τις a priori συναρτήσεις πιθανότητας (δηλαδή για τις πεποιθήσεις μας πριν πραγματοποιηθεί το πείραμα) για τις παραμέτρους μ, σ όπως δείχνει ο παρακάτω πίνακας:

Prior	$p(\mu, \sigma^2)$	Παράμετρος	Posterior
Uniform	1	μ, σ^2	$T(n-3, m, S), \chi^2(n-3, S)$
Jeffreys	$1/\sigma^4$	μ, σ^2	$T(n+1, m, S), \chi^2(n+1, S)$
MDIP ⁹ Reference	$1/\sigma^2$	μ, σ^2	$T(n-1, m, S), \chi^2(n-1, S)$
Conjugate	$N(\mu_0, \sigma/\kappa_0^{1/2}) \chi^2(\nu_0, \sigma_0)$	μ, σ^2	$T(n+\nu_0, (\kappa_0\mu_0+nm)/(n+\kappa_0), \sigma_n/(n+\kappa_0)^{1/2}), \chi^2(n+\nu_0, \sigma_n)$

Πίνακας 1 Παραμετρική μορφή κατανομής πυκνότητας πιθανότητας για διαφορετικές επιλογές priors

$$\text{Όπου } m = \frac{1}{n} \sum_{i=1}^n d_i, \quad S^2 = \frac{1}{n} \sum_{i=1}^n (d_i - m)^2, \quad (\nu_0 + n)\sigma_n^2 = \nu_0\sigma_0^2 + (n-1)S^2 + \frac{k_0 n}{k_0 + n} (m - \mu_0)^2$$

Και $T(n, \mu, s)$ η κατανομή Student t με παράμετρο θέσης μ , παράμετρο κλίμακας (scale parameter) s και βαθμούς ελευθερίας n , κατανομή που ορίζεται από τη σχέση :

$$\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)s\sqrt{n\pi}} \left[1 + \frac{1}{n} \left(\frac{x-\mu}{s} \right)^2 \right]^{-(n+1)/2}$$

⁹ MDIP : Maximal Data Information Prior

Με $N(\mu, \sigma)$ συμβολίζουμε τη κανονική κατανομή με παράμετρο θέσης μ , παράμετρο κλίμακας σ οριζόμενη από $\left(\frac{1}{2\pi\sigma^2}\right)^{1/2} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$, ενώ

$\chi^2(n, S)$ είναι η αντίστροφη χ^2 κατανομή με παράμετρο μη-κεντρικότητας S οριζόμενη

$$\text{από } \frac{2^{-n/2} e^{-nS^2/2x} n^{n/2} S^n x^{-(1+n/2)}}{\Gamma\left(\frac{n}{2}\right)}$$

και $\Gamma(x)$ είναι η συνάρτηση Γάμμα, οριζόμενη από $\Gamma(x) = x!$ για ακέραιες τιμές του x μέσω αναλυτικής συνέχισης στο υπόλοιπο μιγαδικό επίπεδο.

Σε όλες τις περιπτώσεις βλέπουμε ότι η *a posteriori* πεποίθησή μας για την παράμετρο μ (που αντιστοιχεί στην «αναμενόμενη» τιμή του επιπέδου του mRNA σε μια ορισμένη κατάσταση) εκφράζεται μέσω της κατανομής t · διαφορετικές *a priori* πεποιθήσεις οδηγούν στην ίδια γενική παραμετρική λύση αλλά με διαφορετικές τιμές των παραμέτρων που εμφανίζονται στην ειδική μορφή της κατανομής. Η συχνοτική θεωρία προτείνει τη λύση που ενώ τυπολογικά είναι ίδια με αυτή της MDIP δεν επιδέχεται της ίδιας ερμηνείας.

4.2 Το πρόβλημα των «δύο μέσων όρων»

Το πρόβλημα των «δύο μέσων όρων» ή πρόβλημα Behrens – Fisher σχετίζεται με την εξαγωγή συμπερασμάτων σχετικά με την διαφορά είτε των δύο μέσων τιμών πληθυσμιακών κατανομών (συχνοτική προσέγγιση) είτε των παραμέτρων θέσης *posterior* κατανομών. Το κλασσικό πρόβλημα Behrens – Fisher μελετά την περίπτωση που οι δύο μέσοι όροι (ή πεποιθήσεις) ακολουθούν το ΚΜΠ, ενώ το γενικευμένο πρόβλημα αναφέρεται στην περίπτωση που άρεται ο περιορισμός αυτός. Και στις δύο περιπτώσεις, καμία παραδοχή δεν γίνεται για την τιμή ή τη σχέση μεταξύ των πληθυσμιακών διακυμάνσεων (παραμέτρων κλίμακας). Μαθηματικά το πρόβλημα έχει ένα εξαιρετικά απλοϊκό ορισμό: *την εύρεση της συνάρτησης πυκνότητας πιθανότητας, της διαφοράς μεταξύ των παραμέτρων θέσης δύο τυχαίων μεταβλητών (ή κατανομών) t χωρίς κανένα περιορισμό όσον αφορά τους βαθμούς ελευθερίας ή τις παραμέτρους κλίμακας*¹⁰.

Για τη λύση του φαινομενικά απλού αυτού προβλήματος έχουν προταθεί διάφορες προσεγγίσεις στη διάρκεια των τελευταίων 75 χρόνων (^{101;102} για ιστορική αναδρομή). Το χαρακτηριστικό αυτών των λύσεων, είναι ότι η προτεινόμενη από τη Bayesian θεωρία πιθανότητας είναι μια *ακριβής* λύση, ενώ οι προτεινόμενες από τη συχνοτική θεωρία είναι προσεγγιστικές λύσεις που ανακύπτουν από την αδυναμία να αντιμετωπισθεί με μαθηματικούς όρους η «προβληματική» παράμετρος του λόγου των παραμέτρων κλίμακας των προς σύγκριση κατανομών. Οι προτεινόμενες λύσεις εκφράζονται με συναρτήσεις πυκνότητας πιθανότητας που πρέπει στη συνέχεια να ολοκληρωθούν για τον υπολογισμό της πιθανότητας υποέκφρασης · για τη Bayesian λύση ο αριθμητικός υπολογισμός ολοκληρωμάτων της κατανομής Behrens-Fisher δεν είναι επαρκώς μελετημένος.

Με τυπικούς όρους τα στοιχεία του προβλήματος των δύο μέσων όρων είναι τα ακόλουθα:

- Δύο καταστάσεις Ξ_1 , και Ξ_2 στις οποίες πραγματοποιούμε τις πειραματικές μέτρησεις

¹⁰ Στην περίπτωση που δεχόμαστε *a priori* ότι οι παράμετροι κλίμακας (ή ισοδύναμα οι πληθυσμιακές διακυμάνσεις) είναι ίσες, τότε η λύση του προβλήματος δίνεται από την κατανομή t δύο δειγμάτων.

- Κάθε μέτρηση αποτελείται από n_i πειραματικά σημεία
- Κάθε μέτρηση οδηγεί σε πρωτογενή δεδομένα των οποίων ο μέσος όρος είναι $\bar{x}_i$ και η δειγματική τυπική απόκλιση s_i $i=1,2$
- Το κανονικό μοντέλο πιθανότητας χρησιμοποιείται για την συμπερασματολογία επί των μετρήσεων
- Το υπό εξέταση μέγεθος είναι η διαφορά παραμέτρων θέσης των δύο κατανομών Student t

Συχνотική λύση στο πρόβλημα Behrens-Fisher

Η συχνотική λύση στο πρόβλημα Behrens-Fisher ξεκινά από τον υπολογισμό του μεγέθους :

$$T = \frac{\delta - (\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{v_1} + \frac{s_2^2}{v_2}}}$$

με βάση τις δειγματικές μέσες τιμές και τυπικές αποκλίσεις. Οι βαθμοί ελευθερίας v_1, v_2 ισούνται με το μέγεθος των αντίστοιχων δειγμάτων μειωμένων κατά ένα. Ο υπολογισμός της κατανομής αυτού του μεγέθους ΔΕΝ είναι εφικτός εντός της συχνотικής προσέγγισης. Προσεγγίσεις στη Bayesian λύση έχουν παρουσιαστεί στη βιβλιογραφία¹⁰³⁻¹¹⁶ με πιο γνωστή τη μέθοδο του Welch¹¹⁷⁻¹²¹ που χρησιμοποιώντας τη μέθοδο των ροπών (method of moments) προσέγγισαν την κατανομή του T με την κατανομή Student με βαθμούς ελευθερίας που δίνονται από τη σχέση:

$$\frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

Η σύγκριση της λύσης αυτής με την Bayesian λύση έχει γίνει αντικείμενο πολλών ερευνών στην στατιστική βιβλιογραφία¹²²⁻¹²⁴ με την κοινή παραδοχή ότι για μεγάλα δείγματα ($>15-25$) οι λύσεις τείνουν να συγκλίνουν. Σημαντικές αποκλίσεις εμφανίζονται για πολύ μικρά δείγματα, και σε τέτοιες καταστάσεις η πιθανότητα που δίνει η Bayesian λύση φαίνεται να υπερτερεί^{101,125}

Bayesian λύση στο πρόβλημα Behrens-Fisher

Η κλασσική Bayesian λύση¹²⁶⁻¹²⁹ στο πρόβλημα Behrens – Fisher που χρησιμοποιεί τις prior του Πίνακα 1 παίρνει τη μορφή της κατανομής Fisher-Behrens με βαθμούς ελευθερίας v_1, v_2 και γωνίας φ : $BF(X, v_1, v_2, \varphi)$. Η κατανομή αυτή είναι η ίδια με την κατανομή της μεταβλητής $T_1 \sin \varphi - T_2 \cos \varphi$ όπου T_1 και T_2 είναι ανεξάρτητες Student μεταβλητές με v_1 και v_2 βαθμούς ελευθερίας αντίστοιχα. Η γωνία φ υπολογίζεται από το λόγο των παραμέτρων θέσεων των δύο μεταβλητών και τους βαθμούς ελευθερίας των, σύμφωνα με τη σχέση:

$$\tan \varphi = \frac{s_1/\sqrt{v_1}}{s_2/\sqrt{v_2}}$$

Οι βαθμοί ελευθερίας των δύο μεταβλητών T_1 και T_2 σχετίζεται με το μέγεθος των δειγμάτων με τις σχέσεις:

- $n_1 = v_1 + 3$ $n_2 = v_2 + 3$ (Uniform prior)
- $n_1 = v_1 - 1$ $n_2 = v_2 - 1$ (Jeffrey's prior)
- $n_1 = v_1 + 1$ $n_2 = v_2 + 1$ (MDIP prior)
- $n_1 = v_1 - v_0 = v_2 - v_0$ (Conjugate prior)

Οι δύο μεταβλητές T_1 και T_2 εκφράζουν την πεποίθησή μας για τα επίπεδα mRNA σε κάθε μία από τις καταστάσεις Ξ_1 και Ξ_2 και σχετίζονται με τα δειγματικά (πειραματικά) δεδομένα με τις ακόλουθες σχέσεις:

$$T_1 = \frac{L_{\Xi_1} - \bar{x}_1}{s_1/\sqrt{\nu_1}}, \quad T_2 = \frac{L_{\Xi_2} - \bar{x}_2}{s_2/\sqrt{\nu_2}} \quad (\text{στην περίπτωση των Uniform, Jeffreys, MDIP priors})$$

$$T_1 = \frac{L_{\Xi_1} - (\kappa_0 \mu_0 + n_1 \bar{x}_1)/(\kappa_0 + n_1)}{s_1/\sqrt{\nu_1}}, \quad T_2 = \frac{L_{\Xi_2} - (\kappa_0 \mu_0 + n_2 \bar{x}_2)/(\kappa_0 + n_2)}{s_2/\sqrt{\nu_2}} \quad (\text{συζυγής prior})$$

Η αναλυτική μορφή της συνάρτησης πυκνότητας πιθανότητας αποδεικνύεται ότι δίνεται από τη σχέση¹¹:

$$BF(X, \nu_1, \nu_2, \varphi) = \frac{\int_{-\infty}^{+\infty} \left(1 + \frac{(X \sin \varphi + z \cos \varphi)^2}{\nu_1}\right)^{-\frac{\nu_1+1}{2}} \left(1 + \frac{(z \sin \varphi - X \cos \varphi)^2}{\nu_2}\right)^{-\frac{\nu_2+1}{2}} dz}{B\left(\frac{\nu_1}{2}, \frac{1}{2}\right) B\left(\frac{\nu_2}{2}, \frac{1}{2}\right) \sqrt{\nu_1 \nu_2}}$$

Κατα αναλογία με την συχνοτική προσέγγιση, το μέγεθος:

$$T = \frac{\delta - (\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{\nu_1} + \frac{s_2^2}{\nu_2}}} \quad \text{ή} \quad T = \frac{\delta - ((\kappa_0 \mu_0 + n_1 \bar{x}_1)/(\kappa_0 + n_1) - (\kappa_0 \mu_0 + n_2 \bar{x}_2)/(\kappa_0 + n_2))}{\sqrt{\frac{s_1^2}{\nu_1} + \frac{s_2^2}{\nu_2}}} \quad \text{ανάλογα με το}$$

είδος της prior χρησιμοποιείται στο άνω(κάτω) όριο των ολοκληρωμάτων που αντιστοιχούν στην υπόθεση της υπό(μη υπό-)έκφρασης.

Λόγω της σύνθετης μορφής αυτής της κατανομής, ο υπολογισμός ολοκληρωμάτων με βάση την παραπάνω κατανομή έχει γίνει μόνο σε ειδικές περιπτώσεις (π.χ. ίσοι βαθμοί ελευθερίας, γωνία ίση με 45 μοίρες, άπειροι βαθμοί ελευθερίας στη μία εκ των περιπτώσεων^{131;132}). Μοντέρνα βιβλία Bayesian στατιστικής χρησιμοποιούν την προσέγγιση της Patil¹³³ που έχει τα ακόλουθα μειονεκτήματα:

- ο είναι επαρκής μόνο για μεγάλα δείγματα (με βαθμούς ελευθερίας >6-7 και ουσιαστικά δίνει σχετικά ακριβή αποτελέσματα για σημαντικά μεγαλύτερους δειγματικούς βαθμούς ελευθερίας)
- ο είναι χαμηλής ακρίβειας (γύρω στα 3 δεκαδικά ψηφία)
- ο δεν καλύπτει όλες τις περιπτώσεις βαθμών ελευθερίας, και γωνιών
- ο μόνο αρκετά περιορισμένοι πίνακες διατίθενται στη βιβλιογραφία
- ο το λογισμικό που χρησιμοποιήθηκε για την παραγωγή των πινάκων είναι διαθέσιμο σε μια pre-Fortran «νεκρή» πλέον γλώσσα προγραμματισμού

Για όλους τους παραπάνω λόγους η προσέγγιση της Patil δεν προσφέρεται παρά για σκοπούς επίδειξης και μόνο. Ειδικά για την εφαρμογή των Bayesian μεθόδων σε προβλήματα μικροσυστοιχείων που απαιτούν τον αυτόματο υπολογισμό χιλιάδων τέτοιων ολοκληρωμάτων (ένα για κάθε γονίδιο) με μεγάλη ακρίβεια και μικρούς βαθμούς ελευθερίας είναι εντελώς ανεπαρκής. Το κύριο τεχνικό επίτευγμα της εργασίας αυτής είναι η απόδειξη μιας κλειστής ολοκληρωτικής μορφής για ολοκληρώματα αυτής της κατανομής και ενός βέλτιστου αριθμητικού τρόπου υπολογισμού της που ξεπερνούν τους περιορισμούς των προηγούμενων λύσεων. Η προτεινόμενη λύση είναι ακριβής (με την έννοια ότι η μορφή που προτείνεται έχει

¹¹ Η απόδειξη της σχέσης αυτής μπορεί να βρεθεί σε βιβλία αναφοράς στατιστικής¹³⁰ και γι' αυτό παραλείπεται

καθολική ισχύ) και η ακρίβεια της περιορίζεται από την ακρίβεια των αριθμητικών υπολογισμών των υπάρχοντων αλγορίθμων.

4.3 Υπολογισμός ολοκληρωμάτων της κατανομής Behrens Fisher

Σχέση της κατανομής Behrens Fisher με τη θεωρία των υπεργεωμετρικών συναρτήσεων

Το έναυσμα για την τεχνική καινοτομία της παρούσας εργασίας ήταν η συστηματική ανασκόπηση της μαθηματικής στατιστικής βιβλιογραφίας στη βάση JSTOR με σκοπό τον εντοπισμό δημοσιεύσεων που πραγματεύονταν μαθηματικά ισοδύναμες μορφές της συνάρτησης πυκνότητας πιθανότητας της κατανομής Behrens – Fisher. Η πρώτη σχετική δημοσίευση εμφανίζεται το 1960 από τον Ruben¹³⁴ που πρώτος εξέφρασε την συνάρτηση μέσω μία υπεργεωμετρικής συνάρτησης στις ακόλουθες οριακές περιπτώσεις : 1) ίσοι βαθμοί ελευθερίας και ίσες δειγματικές διακυμάνσεις 2) άπειρους βαθμούς ελευθερίας για ένα από τα δύο δείγματα. Στη συνέχεια ο Dickey^{135;136} εξέφρασε την BF μέσω μιας συνάρτησης Appell (υπεργεωμετρική συνάρτηση δύο μεταβλητών) εξετάζοντας την κατανομή της σταθμισμένης συνέλιξης δύο τυχαίων Student μεταβλητών¹².

Το αποτέλεσμα του Dickey αποτελεί την αφετηρία της παρούσας προσέγγισης και δίνεται από την εξίσωση:

$$\frac{\Gamma\left(\frac{\nu+1}{2}\right)(\nu_1\beta_1^2)^{\frac{\nu_1}{2}}(\nu_2\beta_2^2)^{\frac{\nu_2}{2}}}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)\pi^{\frac{1}{2}}}\mathcal{B}\left(\frac{\nu_1+1}{2},\frac{\nu_2+1}{2}\right)\mathcal{F}_1\left(\frac{\nu_1+1}{2};\frac{\nu+1}{2},\frac{\nu+1}{2};\frac{\nu}{2}+1;\rho_1,\rho_2\right),$$

$$\nu = \nu_1 + \nu_2,$$

όπου $\Gamma(\alpha)$ είναι συνάρτηση Γ , $\mathcal{B}(\alpha,\beta)$ είναι η συνάρτηση B και $\mathcal{F}_1(\alpha;\beta_1,\beta_2;\gamma;x,y)$ είναι η συνάρτηση Appell¹³⁸⁻¹⁴² οριζόμενη από:

$$\mathcal{F}_1(\alpha;\beta_1,\beta_2;\gamma;x,y) = \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} \frac{(\alpha)_{m+n}(\beta_1)_m(\beta_2)_n}{m!n!(\gamma)_{m+n}} x^m y^n, \quad |x| < 1, |y| < 1$$

και μέσω αναλυτικής συνέχισης (analytical continuation) στο υπόλοιπο μιγαδικό επίπεδο. Το σύμβολο $(\omega)_\mu$ αναφέρεται στην συνάρτηση Pochhammer που ορίζεται από : $(\omega)_\mu = \Gamma(\omega+\mu)/\Gamma(\omega)$.

Τα μεγέθη ρ_1, ρ_2 που εμφανίζονται στην παραπάνω έκφραση είναι οι πραγματικές λύσεις του πολυωνύμου:

$$\rho^2 + \left(\delta^2/\nu_2\beta_2^2 + \nu_1\beta_1^2/\nu_2\beta_2^2 - 1\right)\rho - \delta^2/\nu_2\beta_2^2 = 0$$

Ειδικά για την κατανομή Behrens Fisher οι συντελεστές στάθμισης δίνονται από τις σχέσεις $\beta_1 = \sin \theta$, $\beta_2 = \cos \theta$,

¹² Σε παρόμοια μορφή κατέληξαν και οι Rahman και Saleh¹³⁷ χρησιμοποιώντας την μέθοδο των χαρακτηριστικών συναρτήσεων.

Έκφραση της κατανομής Behrens Fisher με ολοκληρώματα τύπου Euler

Στην υποενότητα αυτή θα καταληξουμε σε μια κλειστή ολοκληρωτική μορφή για την πιθανότητα υποέκφρασης βασιζόμενοι στην κατανομή Behrens Fisher. Πρώτα αποδεικνύουμε το ακόλουθο θεώρημα:

Θεώρημα 1: Το ολοκλήρωμα $\int_0^\delta (b + ax^2)^{-m} dx$ $a > 0, b > 0$ συγκλείνει για όλες τις τιμές

του m στην συνάρτηση ${}_2F^1(m, 1/2; 3/2; -a\delta^2/b)b^{-m}\delta$

όπου ${}_2F^1(a, b; c; z)$ είναι η υπεργεωμετρική συνάρτηση του Gauss που ορίζεται σαν

$$\sum_{i=0}^{\infty} \frac{(a)_i (b)_i}{(c)_i i!} z^i \quad \forall |z| < 1 \text{ και μέσω αναλυτικής συνέχισης στα άλλα σημεία του επιπέδου}$$

Απόδειξη : Το θεώρημα 1 προκύπτει με την αντικατάσταση $y = x^2/\delta^2$ στο αρχικό ολοκλήρωμα και με σύγκριση της συνεπακόλουθης έκφρασης με την ολοκληρωτική αναπαράσταση κατά Picard για την υπεργεωμετρική συνάρτηση :

$${}_2F^1(a, b; c; z) = \frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)} \int_0^1 t^{b-1} (1-t)^{c-b-1} (1-zt)^{-a} dt$$

που ισχύει για όλες τις τιμές $\text{Re}(c) > 0 \wedge \text{Re}(c-b) > 0$ ¹⁴³.

Το κύριο αποτέλεσμα του τμήματος αυτού προκύπτει αρχίζοντας από την κατά Picard ολοκληρωτική αναπαράσταση της συνάρτησης Appell ¹⁴⁴ :

$$F_1(\alpha; \beta_1, \beta_2; \gamma; x, y) = \frac{\Gamma(\gamma)}{\Gamma(\alpha)\Gamma(\gamma-\alpha)} \int_0^1 u^{\alpha-1} (1-u)^{\gamma-\alpha-1} (1-xu)^{-\beta_1} (1-yu)^{-\beta_2} du$$

εφόσον $\Gamma(\alpha) > 0, \Gamma(\gamma-\alpha) > 0$.

Οι παραδοχές αυτές ισχύουν στην περίπτωση της συνάρτησης Appell που εμφανίζεται στην συνάρτηση πυκνότητας πιθανότητας Behrens Fisher με αποτέλεσμα το ολοκλήρωμα της συνάρτησης αυτής από το $-\infty$ ως το δ (το ολοκλήρωμα αυτό αντιστοιχεί στην σκιασμένη επιφάνεια της Εικόνας 1) να μπορεί να εκφραστεί σαν:

$$\frac{\Gamma\left(\frac{\nu+1}{2}\right) \left(\nu_1 \beta_1^2\right)^{\frac{\nu_1}{2}} \left(\nu_2 \beta_2^2\right)^{\frac{\nu_2}{2}}}{\Gamma\left(\frac{\nu_1}{2}\right) \Gamma\left(\frac{\nu_2}{2}\right) \pi^{\frac{1}{2}}} \int_{-\infty}^{\delta} \int_0^1 u^{(\nu_1-1)/2} (1-u)^{(\nu_2-1)/2} (1-\rho_1 u)^{-(\nu+1)/2} (1-\rho_2 u)^{-(\nu+1)/2} du dx$$

Δεδομένης της συμμετρίας της κατανομής Behrens Fisher το παραπάνω διπλο ολοκλήρωμα από $-\infty$ ως το μηδέν είναι ίσο με $1/2$, και έτσι η παραπάνω έκφραση απλοποιείται σε:

$$\frac{1}{2} + \frac{\Gamma\left(\frac{\nu+1}{2}\right) \left(\nu_1 \beta_1^2\right)^{\frac{\nu_1}{2}} \left(\nu_2 \beta_2^2\right)^{\frac{\nu_2}{2}}}{\Gamma\left(\frac{\nu_1}{2}\right) \Gamma\left(\frac{\nu_2}{2}\right) \pi^{\frac{1}{2}}} \int_0^{\delta} \int_0^1 \lambda(u, \delta) du dx,$$

$$\lambda(u, \delta) = u^{(\nu_1-1)/2} (1-u)^{(\nu_2-1)/2} (1-\rho_1 u)^{-(\nu+1)/2} (1-\rho_2 u)^{-(\nu+1)/2}$$

Η σειρά ολοκλήρωσης μπορεί να αλλάξει στο παραπάνω διπλό ολοκλήρωμα επικαλούμενοι το θεώρημα Fubini μια και:

1. $u^{(\nu_1-1)/2}(1-u)^{(\nu_2-1)/2}(1-\rho_1 u)^{-(\nu+1)/2}(1-\rho_2 u)^{-(\nu+1)/2} > 0 \quad \forall u \in [0,1]$, καθώς
 $\text{sgn}[u^{(\nu_1-1)/2}(1-u)^{(\nu_2-1)/2}(1-\rho_1 u)^{-(\nu+1)/2}(1-\rho_2 u)^{-(\nu+1)/2}] = \text{sgn}[(1-\rho_1 u)(1-\rho_2 u)] =$
 $\text{sgn}\left[(1-u) + u \frac{\nu_1 \beta_1^2}{\nu_2 \beta_2^2} + u(1-u) \frac{\delta^2}{\nu_2 \beta_2^2}\right] > 0$
2. Η $\lambda(u, \delta)$ είναι συνεχής $\forall u \in [0,1] \wedge \delta \in [0, \infty)$ και
3. $(1-u) + u \frac{\nu_1 \beta_1^2}{\nu_2 \beta_2^2} + u(1-u) \frac{\delta^2}{\nu_2 \beta_2^2} \geq 1 + u \frac{\nu_1 \beta_1^2}{\nu_2 \beta_2^2} + u(1-u) \frac{\delta^2}{\nu_2 \beta_2^2} \geq 1 + \frac{\nu_1 \beta_1^2}{\nu_2 \beta_2^2} + (1-u) \frac{\delta^2}{\nu_2 \beta_2^2} \geq$
 $1 + \frac{\nu_1 \beta_1^2}{\nu_2 \beta_2^2} + \frac{\delta^2}{\nu_2 \beta_2^2} \Rightarrow (1-\rho_1 u)^{-(\nu+1)/2}(1-\rho_2 u)^{-(\nu+1)/2} \leq \left(1 + \frac{\nu_1 \beta_1^2}{\nu_2 \beta_2^2} + \frac{\delta^2}{\nu_2 \beta_2^2}\right)^{-(\nu+1)/2}.$

Το ολοκλήρωμα:

$$\int_0^\delta \left(1 + \frac{\nu_1 \beta_1^2}{\nu_2 \beta_2^2} + \frac{x^2}{\nu_2 \beta_2^2}\right)^{-(\nu+1)/2} dx$$

συγκλείνει $\forall \delta \in R$, σύμφωνα με το θεώρημα 1 ικανοποιώντας έτσι την τελευταία συνθήκη του θεωρήματος Fubini.

Ανταλλάσσοντας τη σειρά ολοκλήρωσης και εφαρμόζοντας το θεώρημα 1 για μια ακόμα φορά καταλήγουμε στο ακόλουθο αποτέλεσμα για την κατανομή Behrens Fisher (**Εκφραση 1**):

$$\frac{1}{2} + C \int_0^1 x^{(\nu_1-1)/2}(1-x)^{(\nu_2-1)/2} [\nu_1 \beta_1^2 x + \nu_2 \beta_2^2 (1-x)]^{-(\nu+1)/2} {}_2F_1\left(\frac{\nu+1}{2}, \frac{1}{2}; \frac{3}{2}; \frac{-\delta^2 x(1-x)}{\nu_1 \beta_1^2 x + \nu_2 \beta_2^2 (1-x)}\right) dx,$$

$$\text{όπου } C = \frac{\Gamma\left(\frac{\nu+1}{2}\right) (\nu_1 \beta_1^2)^{\frac{\nu_1}{2}} (\nu_2 \beta_2^2)^{\frac{\nu_2}{2}} \delta}{\Gamma\left(\frac{\nu_1}{2}\right) \Gamma\left(\frac{\nu_2}{2}\right) \pi^{\frac{1}{2}}}$$

Το παραπάνω ολοκλήρωμα τύπου Euler είναι αριθμητικά ίσο με την πιθανότητα υποέκφρασης όπως αυτή ορίστηκε στο αρχικό τμήμα του κεφαλαίου αυτού. Σε σχέση με την εξίσωση με την οποία ορίστηκε η πιθανότητα υποέκφρασης παρατηρούμε τις ακόλουθες σημαντικές διαφορές:

1. το ολοκλήρωμα ορίζεται στο κλειστό διάστημα $[0,1]$ αντί για το διάστημα $(-\infty, \delta]$
2. το δ εμφανίζεται σαν μια παράμετρος του ολοκληρώματος, γεγονός που θα μπορούσε να σημαίνει ότι υπάρχει βέλτιστη αριθμητική μέθοδος υπολογισμού του ολοκληρώματος ανεξάρτητη από την τιμή της δ .

Προτού διερευνήσουμε την αριθμητική συμπεριφορά του παραπάνω ολοκληρώματος, θα προσπαθήσουμε να το εκφράσουμε σε ισοδύναμες αναλυτικές μορφές χρησιμοποιώντας βασικές σχέσεις μεταξύ των υπεργεωμετρικών συναρτήσεων. Οι ισοδύναμες αυτές μορφές μπορούν να χρησιμοποιηθούν για την κατασκευή αριθμητικών αλγορίθμων για τον υπολογισμό του ολοκληρώματος αυτού και άρα της πιθανότητας υποέκφρασης.

Ισοδύναμες εκφράσεις της κατανομής Behrens Fisher

Για να εκφράσουμε την κατανομή Behrens Fisher με ισοδύναμες μορφές ξεκινούμε από την ακόλουθη ταυτολογία¹³ για την υπεργεωμετρική συνάρτηση:

$${}_2F_1(a, b; c; z) = {}_2F_1\left(a, b; c; \frac{z}{z-1}\right)(1-z)^{-a}$$

Αντικαθιστώντας τη σχέση αυτή στην έκφραση 1, το παραπάνω ολοκλήρωμα μετασχηματίζεται σε (**Εκφραση 2**):

$$\int_0^1 x^{(\nu_1-1)/2} (1-x)^{(\nu_2-1)/2} \left[\nu_1 \beta_1^2 x + \nu_2 \beta_2^2 (1-x) + \delta^2 x(1-x) \right]^{-(\nu+1)/2} {}_2F_1\left(\frac{\nu+1}{2}, 1; \frac{3}{2}; \frac{\delta^2 x(1-x)}{\nu_1 \beta_1^2 x + \nu_2 \beta_2^2 (1-x) + \delta^2 x(1-x)}\right) dx$$

Η επόμενη μορφή της κατανομής Behrens Fisher εκφράζεται μέσω σταθμισμένων αθροισμάτων υπεργεωμετρικών συναρτήσεων 3 μεταβλητών (**Εκφραση 3**). Το πρώτο βήμα προς την έκφραση της κατανομής μέσω αυτών των συναρτήσεων είναι η ανάλυση της υπεργεωμετρικής συνάρτησης ${}_2F_1(a, 1; c; z)$ με τη βοήθεια της επόμενης σχέσης :

$$\frac{{}_2F_1(1, b-m; c; z) (c-b)_m (1-z)^{-m}}{(1-b)_m} + \frac{(c-1) \sum_{i=0}^{m-1} \frac{(c-b)_i (1-z)^{-i-1}}{(2-b)_i}}{b-1}$$

Διακρίνουμε δύο περιπτώσεις:

Περίπτωση Α:

$$\nu_1 + \nu_2 = 2\kappa, \kappa \in \mathbb{N}$$

οπότε η παραπάνω εξίσωση παίρνει τη μορφή:

$${}_2F_1\left(k + \frac{1}{2}, 1; \frac{3}{2}; z\right) = {}_2F_1\left(k - m + \frac{1}{2}, 1; \frac{3}{2}; z\right) \frac{(1-k)_m (1-z)^{-m}}{\left(\frac{1}{2}-k\right)_m} + \frac{\frac{1}{2} \sum_{i=0}^{m-1} \frac{(1-k)_i (1-z)^{-i-1}}{\left(\frac{3}{2}-k\right)_i}}{k - \frac{1}{2}}$$

ή για $m = k$

$${}_2F_1\left(k + \frac{1}{2}, 1; \frac{3}{2}; z\right) = \frac{\sum_{i=0}^{k-1} \frac{(1-k)_i (1-z)^{-i-1}}{\left(\frac{3}{2}-k\right)_i}}{2\left(k - \frac{1}{2}\right)} + \frac{((1-k)_k (1-z)^{-k}) {}_2F_1\left(\frac{1}{2}, 1; \frac{3}{2}; z\right)}{\left(\frac{1}{2}-k\right)_k}$$

Αλλά $(1-k)_k = 0$, οπότε η παραπάνω έκφραση απλοποιείται σε :

$${}_2F_1\left(k + \frac{1}{2}, 1; \frac{3}{2}; z\right) = \frac{\frac{1}{2} \sum_{i=0}^{m-1} \frac{(1-k)_i (1-z)^{-i-1}}{\left(\frac{3}{2}-k\right)_i}}{k - \frac{1}{2}}$$

Αγνοώντας σταθερούς όρους το ολοκλήρωμα που εμφανίζεται στην σχέση για την κατανομή μπορεί να εκφραστεί σαν:

$$\int_0^1 (1-u)^{\frac{1}{2}(-1+\nu_2)} u^{\frac{1}{2}(-1+\nu_1)} (1-u\rho_1^{-1})^{-k-\frac{1}{2}} (1-u\rho_2^{-1})^{-k-\frac{1}{2}} {}_2F_1\left(k + \frac{1}{2}, 1; \frac{3}{2}; \frac{t^2(1-u)u}{-t^2\rho_1\rho_2(1-u\rho_1^{-1})(1-u\rho_2^{-1})}\right) du$$

¹³ Οι παρακάτω ταυτολογίες για τις υπεργεωμετρικές συναρτήσεις βρίσκονται στο βιβλίο των Andrews et al¹⁴³ και στην ιστοσελίδα της Wolfram Research <http://functions.wolfram.com/>

όπου ρ_1, ρ_2 οι λύσεις του διωνύμου: $\nu_1 \beta_1^2 x + \nu_2 \beta_2^2 (1-x) + \delta^2 x(1-x)$ που δίνονται από τη σχέση:

$$\frac{-t^2 - \sin^2(\theta) \nu_1 + \cos^2(\theta) \nu_2 + \sqrt{4 t^2 \nu_2 \cos^2(\theta) + (t^2 - \nu_2 \cos^2(\theta) + \sin^2(\theta) \nu_1)^2}}{2 t^2}$$

$$\frac{-t^2 - \sin^2(\theta) \nu_1 + \cos^2(\theta) \nu_2 - \sqrt{4 t^2 \nu_2 \cos^2(\theta) + (t^2 - \nu_2 \cos^2(\theta) + \sin^2(\theta) \nu_1)^2}}{2 t^2}$$

Κάνοντας τις σχετικές αντικαταστάσεις στην έκφραση 2 παίρνουμε:

$$\int_0^1 (1-u)^{\frac{1}{2}(-1+\nu_2)} u^{\frac{1}{2}(-1+\nu_1)} (1-u\rho_1^{-1})^{-k-\frac{1}{2}} (1-u\rho_2^{-1})^{-k-\frac{1}{2}} \left(\frac{1}{2k-1} \sum_{i=0}^{k-1} \frac{(1-k)_i (1-z)^{-i-1}}{(\frac{3}{2}-k)_i} \right) du$$

$$1-z = (1 - (t^2 (1-u)u) / (-t^2 \rho_1 \rho_2 (1-u\rho_1^{-1})(1-u\rho_2^{-1})))$$

Αλλάζοντας τη σειρά της πεπερασμένης άθροισης με την ολοκλήρωση παίρνουμε τελικά τη μορφή :

$$\frac{1}{2k-1} \sum_{i=0}^{k-1} \frac{(1-k)_i (-t^2 \rho_1 \rho_2)^{i-k+\frac{1}{2}} (\nu_2 \beta_2^2)^{-i-1}}{(\frac{3}{2}-k)_i} \int_0^1 u^{\frac{1}{2}(-1+\nu_1)} (1-u)^{\frac{1}{2}(-1+\nu_2)} (1-\lambda_1 u)^{i-k+\frac{1}{2}} (1-\lambda_2 u)^{i-k+\frac{1}{2}} (1-\lambda_3 u)^{-i-1} du$$

$$\text{όπου } \lambda_1 = 1/\rho_1, \lambda_2 = 1/\rho_2, \text{ και } \lambda_3 = 1 - \frac{\nu_1 \beta_1^2}{\nu_2 \beta_2^2}$$

Το ολοκλήρωμα εντός του αθροίσματος ισούται με το γινόμενο μιας συνάρτησης Lauricella (υπεργεωμετρική συνάρτηση 3 μεταβλητών)¹⁴⁵ επί την αντίστοιχη σταθερά κανονικοποίησης δηλαδή:

$$B\left(\frac{1}{2}(\nu_1+1), \frac{1}{2}(\nu_2+1)\right) F_D^3\left[\frac{1}{2}(\nu_1+1), k-i, k-i, i+1; \frac{1}{2}(\nu_1+\nu_2+2); \lambda_1, \lambda_2, \lambda_3\right]$$

Κάνοντας όλες τις σχετικές αντικαταστάσεις καταλήγουμε στην τελική μορφή:

$$\frac{\left(t \Gamma\left(\frac{1}{2}(\nu_1+\nu_2+1)\right) (\sin^2(\theta) \nu_1)^{\frac{\nu_1}{2}} (\cos^2(\theta) \nu_2)^{\frac{\nu_2}{2}}\right)}{\sqrt{\pi} \Gamma\left(\frac{1}{2}(\nu_1+\nu_2+2)\right) (2k-1)} X$$

$$\sum_{i=0}^{\frac{1}{2}(\nu_1+\nu_2-2)} \frac{(1-\frac{1}{2}(\nu_1+\nu_2))_i (-t^2 \rho_1 \rho_2)^{i-k+\frac{1}{2}} (\nu_2 \beta_2^2)^{-i-1} F_D^3\left[\frac{1}{2}(\nu_1+1), \frac{1}{2}(\nu_1+\nu_2-1)-i, \frac{1}{2}(\nu_1+\nu_2-1)-i, i+1; \frac{1}{2}(\nu_1+\nu_2+2), \lambda_1, \lambda_2, \lambda_3\right]}{(\frac{3}{2}-\frac{1}{2}(\nu_1+\nu_2))_i}$$

Περίπτωση B:

$$\nu_1+\nu_2 = 2\kappa+1, \kappa \in \mathbb{N}$$

Στην περίπτωση αυτή το ολοκλήρωμα δίνεται από τη σχέση :

$$\int_0^1 (1-u)^{\frac{1}{2}(-1+\nu_2)} u^{\frac{1}{2}(-1+\nu_1)} (1-u\rho_1^{-1})^{-k-1} (1-u\rho_2^{-1})^{-k-1} {}_2F_1\left(k+1, 1; \frac{3}{2}; \frac{t^2(1-u)u}{-t^2\rho_1\rho_2(1-u\rho_1^{-1})(1-u\rho_2^{-1})}\right) du$$

Η υπεργεωμετρική συνάρτηση εντός του ολοκληρώματος μπορεί να γραφτεί σαν:

$${}_2F_1\left(k+1, 1; \frac{3}{2}; z\right) = {}_2F_1\left(k-m+1, 1; \frac{3}{2}; z\right) \frac{\left(\frac{1}{2}-k\right)_m (1-z)^{-m}}{(-k)_m} + \frac{\frac{1}{2} \sum_{i=0}^{m-1} \frac{\left(\frac{1}{2}-k\right)_i (1-z)^{-i-1}}{(1-k)_i}}{k}$$

Επεκτείνοντας την παραπάνω σχέση για $k = m$, και κάνοντας χρήση της σχέσης :

$${}_2F_1\left(1, 1; \frac{3}{2}; z\right) = \frac{\sin^{-1}(\sqrt{z})}{\sqrt{1-z}\sqrt{z}}$$

καταλήγουμε ότι το παραπάνω ολοκλήρωμα μπορεί να γραφτεί σαν άθροισμα δύο όρων:

$$\int_0^1 (1-u)^{\frac{1}{2}(-1+\nu_2)} u^{\frac{1}{2}(-1+\nu_1)} (1-u\rho_1^{-1})^{-k-1} (1-u\rho_2^{-1})^{-k-1} \sum_{i=0}^{k-1} \frac{\left(\frac{1}{2}-k\right)_i \left(\frac{(1-u)\nu_2 \cos^2(\theta) + u \sin^2(\theta)\nu_1}{-t^2\rho_1\rho_2(1-u\rho_1^{-1})(1-u\rho_2^{-1})}\right)^{-i-1}}{(1-k)_i} du$$

και :

$$\frac{\left(\frac{1}{2}-k\right)_k}{(-k)_k |t|} (-t^2\rho_1\rho_2)^{k+1} (\nu_2 \cos^2(\theta))^{-k-\frac{1}{2}} \int_0^1 (1-u)^{\frac{-2+\nu_2}{2}} u^{\frac{-2+\nu_1}{2}} (1-u\lambda_3)^{-k-\frac{1}{2}} \sin^{-1}\left(\sqrt{\frac{(1-u)u}{-\rho_1\rho_2(1-u\rho_1^{-1})(1-u\rho_2^{-1})}}\right) du$$

$$\text{όπου } \lambda_1 = 1/\rho_1, \lambda_2 = 1/\rho_2, \text{ και } \lambda_3 = 1 - \frac{\nu_1\beta_1^2}{\nu_2\beta_2^2}$$

Ο πρώτος όρος ισούται με το άθροισμα k συναρτήσεων Lauricella, γεγονός που αποδεικνύεται εύκολα με την αλλαγή της σειρά ολοκλήρωσης και άθροισης και συγκρίνοντας το ολοκλήρωμα εντός των αθροισμάτων με τον κατά Picard ολοκληρωτικό ορισμό της συνάρτησης Lauricella F_D ¹⁴⁶:

$$\sum_{i=0}^{k-1} \left(\frac{-t^2\rho_1\rho_2}{C}\right)^{i+1} \frac{\left(\frac{1}{2}-k\right)_i}{(1-k)_i} \int_0^1 (1-u)^{\frac{1}{2}(-1+\nu_2)} u^{\frac{1}{2}(-1+\nu_1)} (1-u\lambda_1)^{-k+i} (1-u\lambda_2)^{-k+i} (1-u\lambda_3)^{-i-1} du$$

και:

$$\int_0^1 (1-u)^{\frac{1}{2}(-1+\nu_2)} u^{\frac{1}{2}(-1+\nu_1)} (1-u\lambda_1)^{-k+i} (1-u\lambda_2)^{-k+i} (1-u\lambda_3)^{-i-1} du =$$

$$B\left(\frac{\nu_1+1}{2}, \frac{\nu_2+1}{2}\right) F_D^{(3)}\left(\frac{\nu_1+1}{2}, k-i, k-i, i+1; \frac{\nu_1+\nu_2+2}{2}; \lambda_1, \lambda_2, \lambda_3\right)$$

Για τις μεταβλητές της παραπάνω συνάρτησης Lauricella εύκολα αποδεικνύεται ότι ισχύουν οι σχέσεις: $\lambda_1 \in (0,1)$, $\lambda_2 \in (-\infty,0)$, $\lambda_3 \in (-\infty,1)$.

4.4 Ασυμπτωτική ανάλυση των ολοκληρωμάτων της κατανομής Behrens Fisher

Έχοντας καταλήξει σε μια σειρά ισοδύναμων μορφών για την κατανομή Behrens Fisher, θα εξετάσουμε μεθόδους υπολογισμού της, μέσω τεχνικών αριθμητικής ολοκλήρωσης. Μέσω της ανάλυσης της συμπεριφοράς των προς ολοκλήρωση συναρτήσεων που παρουσιάστηκαν στις προηγούμενες παραγράφους θα καταλήξουμε στο να προτείνουμε μια μέθοδο αριθμητικής ανάλυσης (μέθοδο διπλού εκθετικού μετασχηματισμού)¹⁴⁷ που έχει αποδειχτεί ότι είναι βέλτιστη για την ολοκλήρωση συναρτήσεων¹⁴⁸ που έχουν γενικά χαρακτηριστικά παρόμοια με τις δικιές μας.

Ανάλυση της έκφρασης 1

Η προς ολοκλήρωση συνάρτηση που εμφανίζεται στην πρώτη μορφή της κατανομής Behrens Fisher έχει τα ακόλουθα χαρακτηριστικά:

- ο είναι αναλυτική στο κλειστό διάστημα $[0, I]$.
- ο μπορούμε να βρούμε μια αντικατάσταση μεταβλητής $x = y(t)$, $0 = y(-\infty)$, $1 = y(\infty)$ τέτοια ώστε η πτώση της τιμής της συνάρτησης προς ολοκλήρωση να ακολουθεί εκθετικό $|f(y(t))y'(t)| \approx ce^{-\gamma|t|}$, $|t| \rightarrow \infty$ ή διπλό εκθετικό ρυθμό¹⁴: $|f(y(t))y'(t)| \approx \exp(-ce^{\gamma|t|})$, $|t| \rightarrow \infty$.

Ένας τέτοιος μετασχηματισμός δίνεται από τη σχέση $x \rightarrow \tanh(\exp(u))$, που μετασχηματίζει την συνάρτηση προς ολοκλήρωση σε :

$$e^{2u} \tanh(e^u)^{(v_1-1)/2} (1 - \tanh(e^u))^{(v_2-1)/2} [v_2 \beta_2^2 + (v_1 \beta_1^2 - v_2 \beta_2^2) \tanh(e^u)]^{-(v+1)/2} \times {}_2F_1\left(\frac{v+1}{2}, \frac{1}{2}, \frac{3}{2}; \frac{-\delta^2 \tanh(e^u)[1 - \tanh(e^u)]}{v_2 \beta_2^2 + (v_1 \beta_1^2 - v_2 \beta_2^2) \tanh(e^u)}\right) \sec(e^u)^2$$

Αντικαθιστώντας τις τριγωνομετρικές συναρτήσεις με τις ισοδύναμες εκθετικές και μετά από αρκετές αλγεβρικές τροποποιήσεις παίρνουμε το εξής αποτέλεσμα :

$$\frac{2^{(v_2+3)/2} [2v_2 \beta_2^2 - v_1 \beta_1^2 + v_1 \beta_1^2 e^{2e^u}]^{-(v+1)/2} e^{2e^u+2u} (e^{2e^u} - 1)^{(v_1-1)/2} \sec(e^u)^2} {\sqrt{1+e^{2e^u}}} {}_2F_1\left(\frac{v+1}{2}, \frac{1}{2}, \frac{3}{2}; \frac{-2\delta^2 (e^{2e^u} - 1)/(e^{2e^u} + 1)}{(2v_2 \beta_2^2 - v_1 \beta_1^2 + v_1 \beta_1^2 e^{2e^u})}\right)$$

Για να αναλύσουμε την ασυμπτωτική συμπεριφορά της παραπάνω συνάρτησης της γράφουμε στην ακόλουθη μορφή:

$$2^{(v_2+3)/2} [2v_2 \beta_2^2 - v_1 \beta_1^2 + v_1 \beta_1^2 e^{2e^u}]^{-(v+1)/2} \sec(e^u)^2 e^{2u} \left(\frac{e^{2e^u} - 1}{2v_2 \beta_2^2 - v_1 \beta_1^2 + v_1 \beta_1^2 e^{2e^u}} \right)^{v_1/2} \left(\frac{e^{e^u}}{\sqrt{1+e^{2e^u}}} \right) \left(\frac{e^{e^u}}{\sqrt{e^{2e^u} - 1}} \right) \times {}_2F_1\left(\frac{v+1}{2}, \frac{1}{2}, \frac{3}{2}; \frac{-2\delta^2 (e^{2e^u} - 1)/(e^{2e^u} + 1)}{(2v_2 \beta_2^2 - v_1 \beta_1^2 + v_1 \beta_1^2 e^{2e^u})}\right)$$

¹⁴ Στην περίπτωση που είναι γνωστό ότι η $|y'(t)|$ χαρακτηρίζεται από εκθετική ή διπλή εκθετική πτώση, τότε αρκεί να δειχθεί ότι η $|f(y(t))|$ χαρακτηρίζεται από διπλή εκθετική πτώση, λόγω των ιδιοτήτων των σχέσεων τάξεως

Καθώς το $u \rightarrow +\infty$, η ασυμπτωτική συμπεριφορά των όρων του γινομένου (πλην της υπεργεωμετρικής συνάρτησης) δίνεται από τις ακόλουθες σχέσεις :

$$\left(\frac{e^{2e^u} - 1}{2\nu_2\beta_2^2 - \nu_1\beta_1^2 + \nu_1\beta_1^2 e^{2e^u}} \right)^{\nu_1/2} \approx (\nu_1\beta_1^2)^{-\nu_1/2}, \quad \left(\frac{e^{e^u}}{\sqrt{1+e^{2e^u}}} \right) \approx 1, \quad \left(\frac{e^{e^u}}{\sqrt{e^{2e^u} - 1}} \right) \approx 1$$

$$\left[2\nu_2\beta_2^2 - \nu_1\beta_1^2 + \nu_1\beta_1^2 e^{2e^u} \right]^{-(\nu_2+1)/2} \approx (\nu_1\beta_1^2)^{-(\nu_2+1)/2} e^{-(\nu_2+1)e^u}$$

$$e^{2u} \sec(e^u)^2 = \frac{4e^{2u}}{(e^{-e^u} + e^{e^u})^2} < \frac{4e^{2u}}{e^{2e^u}} = O(\exp(-(2-\varepsilon_+)e^u))$$

Χρησιμοποιώντας της ολοκληρωτική αναπαράσταση κατά Picard's μπορούμε να προσεγγίσουμε την ασυμπτωτική συμπεριφορά της υπεργεωμετρικής συνάρτησης ως:

$${}_2F_1\left(\frac{\nu+1}{2}, \frac{1}{2}; \frac{3}{2}; \frac{-2\delta^2(e^{2e^u}-1)/(e^{2e^u}+1)}{(2\nu_2\beta_2^2 - \nu_1\beta_1^2 + \nu_1\beta_1^2 e^{2e^u})}\right) \approx \frac{\Gamma(3/2)}{\Gamma(1/2)} \int_0^1 t^{-1/2} (1 + \frac{2\delta^2}{\nu_1\beta_1^2} e^{-2e^u} t)^{-(\nu+1)/2} dt$$

όπου χρησιμοποιήσαμε το γεγονός :

$$\frac{-2\delta^2(e^{2e^u}-1)/(e^{2e^u}+1)}{(2\nu_2\beta_2^2 - \nu_1\beta_1^2 + \nu_1\beta_1^2 e^{2e^u})} \approx \frac{-2\delta^2}{\nu_1\beta_1^2} e^{-2e^u}$$

Συνεπώς, καθώς το u τείνει στο $+\infty$ η υπεργεωμετρική συνάρτηση συμπεριφέρεται σαν¹⁵:

$$\frac{\Gamma(3/2)}{\Gamma(1/2)} \int_0^1 t^{-1/2} (1 + \frac{2\delta^2}{\nu_1\beta_1^2} e^{-2e^u} t)^{-(\nu+1)/2} dt \approx \frac{\Gamma(3/2)}{\Gamma(1/2)} \int_0^1 t^{-1/2} dt \sim 1$$

Συνεπώς η συνάρτηση προς ολοκλήρωσης φθίνει σαν $O(\exp(-(\nu_2 + 3 - \varepsilon_+)e^u))$, $0 < \varepsilon_+ < 1$ δηλαδή με διπλό εκθετικό ρυθμό καθώς η μεταβλητής προς ολοκλήρωση τείνει στο $+\infty$.

Στην περίπτωση $u \rightarrow -\infty$ έχουμε τα εξής αποτελέσματα:

$$\left(\frac{e^{e^u}}{\sqrt{1+e^{2e^u}}} \right) \approx 1/\sqrt{2}, \quad \sec(e^u)^2 = \frac{4}{(e^{-e^u} + e^{e^u})^2} \approx 1$$

$$\left[2\nu_2\beta_2^2 - \nu_1\beta_1^2 + \nu_1\beta_1^2 e^{2e^u} \right]^{-(\nu+1)/2} \approx (2\nu_2\beta_2^2)^{-(\nu+1)/2}, \quad \text{ενώ το ολοκλήρωμα που ορίζει την υπεργεωμετρική συνάρτηση τείνει προς σταθερά τιμή διαφορετική από το μηδέν, καθώς:}$$

$${}_2F_1\left(\frac{\nu+1}{2}, \frac{1}{2}; \frac{3}{2}; \frac{-2\delta^2(e^{2e^u}-1)/(e^{2e^u}+1)}{(2\nu_2\beta_2^2 - \nu_1\beta_1^2 + \nu_1\beta_1^2 e^{2e^u})}\right) \approx \frac{\Gamma(3/2)}{\Gamma(1/2)} \int_0^1 t^{-1/2} (1 + \frac{\delta^2}{2\nu_2\beta_2^2} (e^{e^u} - 1)t)^{-(\nu+1)/2} dt = 2 \frac{\Gamma(3/2)}{\Gamma(1/2)}$$

Συνεπώς η πτώση της συνάρτησης είναι τάξης που καθορίζεται από τον όρο:

$$e^{2u} (e^{2e^u} - 1)^{(\nu_1-1)/2}$$

Για τον όρο αυτό ισχύει ότι:

¹⁵ Στο σημείο αυτό χρησιμοποιούμε το γεγονός ότι μπορούμε να μεταφέρουμε τις σχέσεις τάξεως εντός της πράξης του ολοκληρώματος

$$\left(e^{2u}e^{2e^u}-e^u\right)^{(v_1-1)/2} < e^{2u}\left(e^{2e^u}-1\right)^{(v_1-1)/2} < e^{2u}\left(e^{2e^u}-1\right) < e^{2u} \text{ για } u \rightarrow -\infty.$$

Κάνοντας την αντικατάσταση της μεταβλητής $y \rightarrow e^{-u}$ στο διάστημα $(-\infty, 0]$ στο γινόμενο $e^{2u}\left(e^{2e^u}-1\right)$ παίρνουμε τη σχέση: $y^{-2}\left(e^{2/y}-1\right)$ η οποία έχει ανάπτυγμα κατά Taylor $\sum_{i=1}^{\infty} \frac{2^i y^{-i}}{i!}$ (η ακτίνα σύγκλισης αυτής της δυναμοσειράς είναι το διάστημα $(0, \infty)$). Μια και οι όροι της δυναμοσειράς είναι μεγαλύτεροι του μηδέν, η τιμή της τελευταίας είναι αυστηρά μεγαλύτερη του πρώτου όρου της, δηλαδή του $2y^{-1}$. Συνεπώς :

$$2y^{-1} < \sum_{i=1}^{\infty} \frac{2^i y^{-i}}{i!} \Rightarrow 2e^u < e^{2u}\left(e^{2e^u}-1\right) \Rightarrow \left(2e^u\right)^{(v_1-1)/2} < \left(e^{2u}e^{2e^u}-e^u\right)^{(v_1-1)/2}$$

Αντικαθιστώντας την σχέση αυτή στην αρχική ανισότητα έχουμε το εξής αποτέλεσμα:

$$2^{(v_1-1)/2} e^{u(v_1-1)/2} < e^u \left(e^{2e^u}-1\right)^{(v_1-1)/2} < e^u$$

δηλαδή η πτώση της προς ολοκλήρωση συνάρτησης είναι εκθετικής τάξης στο διάστημα $(-\infty, 0]$ μια και οι τιμές της είναι φραγμένες (bounded) από συναρτήσεις με εκθετική πτώση.

Ανάλυση της έκφρασης 2

Η ανάλυση της έκφρασης 2 μπορεί να γίνει με εντελώς ανάλογο τρόπο με την προηγούμενη περίπτωση. Παρά το γεγονός ότι οι δύο εκφράσεις είναι ισοδύναμες, η χρήση της έκφρασης 2 παρέχει ορισμένα πλεονεκτήματα όσον αφορά τον υπολογισμό των τιμών της υπεργεωμετρικής συνάρτησης κατά τη διάρκεια της αριθμητικής ολοκλήρωσης της πυκνότητας πιθανότητας της κατανομής Behrens-Fisher. Το γεγονός αυτό σχολιάζεται περαιτέρω στην ανάλυση των αλγορίθμων υπολογισμού βασισμένων στην μέθοδο ολοκλήρωσης διπλού εκθετικού μετασχηματισμού.

Για την ανάλυση της ασυμπτωτικής συμπεριφοράς της έκφρασης 2, εφαρμόζουμε διαδοχικά δύο αλλαγές μεταβλητής στο ολοκλήρωμα, που την ορίζει:

1. $x \rightarrow \frac{x+1}{2}$ που έχει ως αποτέλεσμα την αλλαγή των ορίων ολοκλήρωσης από $[0,1]$ σε $[-1,1]$ και μετά:
2. $\frac{x+1}{2} \rightarrow \tanh\left(\frac{\pi}{2}\sinh u\right)$ (μετασχηματισμός η παράγωγος του οποίου εμφανίζει διπλή εκθετική μείωση) που οδηγεί στην αλλαγή των ορίων ολοκλήρωσης από $[-1,1]$ στο $[-\infty, \infty]$

Η αλλαγής των ορίων ολοκλήρωσης σε δύο βήματα επιτρέπει την ανάλυση της συμπεριφοράς της συνάρτησης όπως αυτή ολοκληρώνεται από προυπάρχουσες αριθμητικές βιβλιοθήκες βασισμένες στη μέθοδο του διπλού εκθετικού μετασχηματισμού (που πρώτα μετασχηματίζουν το προς ολοκλήρωσης διάστημα στο $[-1,1]$ και μετά εφαρμόζουν το μετασχηματισμό 2).

Σαν αποτέλεσμα της εφαρμογής του μετασχηματισμού 1, η προς ολοκλήρωση συνάρτηση παίρνει τη μορφή:

$$2^{\frac{\nu_1}{2} + \frac{\nu_2}{2} + 1} (1-x)^{\frac{\nu_2}{2} - \frac{1}{2}} (x+1)^{\frac{\nu_1}{2} - \frac{1}{2}} {}_2F_1\left(\frac{1}{2}(\nu_1 + \nu_2 + 1), 1; \frac{3}{2}; \frac{(x-1)(x+1)\delta^2}{x^2\delta^2 - \delta^2 - 2x\beta_1^2\nu_1 - 2\beta_1^2\nu_1 + 2x\beta_2^2\nu_2 - 2\beta_2^2\nu_2}\right) \cdot$$

$$(-x^2\delta^2 + \delta^2 + 2x\beta_1^2\nu_1 + 2\beta_1^2\nu_1 - 2x\beta_2^2\nu_2 + 2\beta_2^2\nu_2)^{-\frac{\nu_1}{2} - \frac{\nu_2}{2} - \frac{1}{2}}$$

Η δεύτερη αντικατάσταση οδηγεί στην ακόλουθη μορφή:

$$2^{\frac{\nu_1}{2} + \frac{\nu_2}{2} + 1} \pi \cosh(u) {}_2F_1\left(\frac{1}{2}(\nu_1 + \nu_2 + 1), 1; \frac{3}{2}; \frac{\delta^2(\tanh(\frac{1}{2}\pi \sinh(u)) - 1)(\tanh(\frac{1}{2}\pi \sinh(u)) + 1)}{\tanh^2(\frac{1}{2}\pi \sinh(u))\delta^2 - \delta^2 - 2\beta_1^2\nu_1 - 2\beta_2^2\nu_2 - 2\beta_1^2\nu_1 \tanh(\frac{1}{2}\pi \sinh(u)) + 2\beta_2^2\nu_2 \tanh(\frac{1}{2}\pi \sinh(u))}\right) \cdot$$

$$\operatorname{sech}^2\left(\frac{1}{2}\pi \sinh(u)\right) \left(1 - \tanh\left(\frac{1}{2}\pi \sinh(u)\right)\right)^{\frac{\nu_2}{2} - \frac{1}{2}} \left(\tanh\left(\frac{1}{2}\pi \sinh(u)\right) + 1\right)^{\frac{\nu_1}{2} - \frac{1}{2}} \cdot$$

$$\left(-\tanh^2\left(\frac{1}{2}\pi \sinh(u)\right)\delta^2 + \delta^2 + 2\beta_1^2\nu_1 + 2\beta_2^2\nu_2 + 2\beta_1^2\nu_1 \tanh\left(\frac{1}{2}\pi \sinh(u)\right) - 2\beta_2^2\nu_2 \tanh\left(\frac{1}{2}\pi \sinh(u)\right)\right)^{-\frac{\nu_1}{2} - \frac{\nu_2}{2} - \frac{1}{2}}$$

Αντικαθιστώντας τις τριγωνομετρικές συναρτήσεις με τις ισοδύναμες εκθετικές και μετά από αρκετές αλγεβρικές τροποποιήσεις παίρνουμε το εξής αποτέλεσμα :

$$e^{-u - \frac{e^{-u}\pi}{2} + \frac{e^u\pi}{2}} \left(1 + e^{-\frac{e^{-u}\pi}{2} - \frac{e^u\pi}{2}}\right)^{\frac{1}{2} - \frac{\nu_1}{2}} \left(1 + e^{-\frac{1}{2}e^{-u}\pi + \frac{e^u\pi}{2}}\right)^{\nu_1 + \frac{\nu_2}{2} - \frac{1}{2}} (1 + e^{2u})\pi$$

$${}_2F_1\left(\frac{\nu_1}{2} + \frac{\nu_2}{2} + \frac{1}{2}, 1; \frac{3}{2}; \frac{\delta^2}{\delta^2 + e^{-\frac{1}{2}e^{-u}\pi + \frac{e^u\pi}{2}}\beta_1^2\nu_1 + \beta_1^2\nu_1 + e^{-\frac{e^{-u}\pi}{2} - \frac{e^u\pi}{2}}\beta_2^2\nu_2 + \beta_2^2\nu_2}\right)$$

$$\left(e^{-\frac{1}{2}e^{-u}\pi + \frac{e^u\pi}{2}}\delta^2 + e^{-\frac{1}{2}e^{-u}\pi + \frac{e^u\pi}{2}}\beta_1^2\nu_1 + e^{-e^{-u}\pi + e^u\pi}\beta_1^2\nu_1 + e^{-\frac{1}{2}e^{-u}\pi + \frac{e^u\pi}{2}}\beta_2^2\nu_2 + \beta_2^2\nu_2\right)^{-\frac{\nu_1}{2} - \frac{\nu_2}{2} - \frac{1}{2}}$$

Η μεταβλητή της υπεργεωμετρικής συνάρτησης, είναι αυστηρά μικρότερη του ένα και για μεγάλες (κατά απόλυτη τιμή) του u τείνει στο μηδέν. Συνεπώς η προς ολοκλήρωση συνάρτηση είναι αυστηρά μικρότερη του :

$${}_2F_1\left(\frac{\nu_1}{2} + \frac{\nu_2}{2} + \frac{1}{2}, 1; \frac{3}{2}; 1\right) e^{-u - \frac{e^{-u}\pi}{2} + \frac{e^u\pi}{2}} \left(1 + e^{-\frac{e^{-u}\pi}{2} - \frac{e^u\pi}{2}}\right)^{\frac{1}{2} - \frac{\nu_1}{2}} \left(1 + e^{-\frac{1}{2}e^{-u}\pi + \frac{e^u\pi}{2}}\right)^{\nu_1 + \frac{\nu_2}{2} - \frac{1}{2}} (1 + e^{2u})\pi$$

$$\left(e^{-\frac{1}{2}e^{-u}\pi + \frac{e^u\pi}{2}}\delta^2 + e^{-\frac{1}{2}e^{-u}\pi + \frac{e^u\pi}{2}}\beta_1^2\nu_1 + e^{-e^{-u}\pi + e^u\pi}\beta_1^2\nu_1 + e^{-\frac{1}{2}e^{-u}\pi + \frac{e^u\pi}{2}}\beta_2^2\nu_2 + \beta_2^2\nu_2\right)^{-\frac{\nu_1}{2} - \frac{\nu_2}{2} - \frac{1}{2}}$$

η οποία (εκτός από μια σταθερά πολλαπλασιασμού) προκύπτει με εφαρμογή των προαναφερθέντων μετασχηματισμών στο ολοκλήρωμα

$$\int_0^1 (1-u)^{\frac{1}{2}(\nu_2-1)} u^{\frac{1}{2}(\nu_1-1)} \delta(u\beta_1^2\nu_1 - (u-1)(u\delta^2 + \beta_2^2\nu_2))^{\frac{1}{2}(-\nu_1-\nu_2)} du$$

το οποίο ορίζει μια συνάρτηση Lauricella δύο μεταβλητών. Συνεπώς η ασυμπτωτική συμπεριφορά της έκφρασης 2 είναι φραγμένη από την ασυμπτωτική συμπεριφορά μιας συνάρτησης Lauricella. Αν η τελευταία φθίνει με διπλό εκθετικό ή εκθετικό ρυθμό, τότε και η έκφραση 2 θα πρέπει να φθίνει με ρυθμό τουλάχιστον αυτής της τάξης και αναμένεται ότι μέθοδοι ολοκλήρωσης που μπορούν να εφαρμοσθούν στην περίπτωση της συνάρτησης Lauricella θα βρίσκουν εφαρμογή στη συνάρτηση που ορίζει την έκφραση 2.

Ανάλυση ασυμπτωτικής συμπεριφοράς συναρτήσεων Lauricella

Η ασυμπτωτική ανάλυση της έκφρασης 3 παρουσιάζει ενδιαφέρον λόγω της χρήσης συναρτήσεων Lauricella. Οι συναρτήσεις αυτές εμφανίζονται σε ενδιαφέροντα προβλήματα Bayesian στατιστικής και συγκεκριμένα στον υπολογισμό συναρτήσεων

κατανομών βασισμένων σε διαδικασίες Dirichlet (Dirichlet processes)^{149;150}. Οι τελευταίες εμφανίζονται σε πληθώρα προβλημάτων της βιοπληροφορικής όπως μοντελοποίηση δεδομένων γονιδιακής έκφρασης¹⁵¹, ανάλυση απλοτύπων SNPs (Single Nucleotide Polymorphisms)¹⁵², text mining^{153;154}. Λόγω έλλειψης λογισμικού για τον υπολογισμό αυτών των συναρτήσεων, η προσέγγιση αυτών των προβλημάτων γίνεται με μεθόδους Monte Carlo και δειγματοληψίας (sampling)¹⁵⁵.

Η μέθοδος του διπλού εκθετικού μετασχηματισμού είναι ιδανική για τον υπολογισμό του ολοκληρώματος της κατανομής Behrens – Fisher λόγω των ιδιοτήτων της τελευταίας. Μια και το ολοκλήρωμα αυτό μπορεί επίσης να ορισθεί σαν σταθμισμένο πεπερασμένο άθροισμα συναρτήσεων Lauricella, τίθεται το ερώτημα μήπως η μέθοδος του διπλού εκθετικού μετασχηματισμού μπορεί (και υπό ποιους περιορισμούς) να χρησιμοποιηθεί για τον υπολογισμό τιμών τέτοιων συναρτήσεων.

Οι συναρτήσεις αυτές ορίζονται από τον μετασχηματισμό:

$$F_D^n(\alpha; \beta_1, \beta_2, \dots, \beta_n; c; x_1, x_2, \dots, x_n) = \frac{\Gamma(c)}{\Gamma(\alpha)\Gamma(c-\alpha)} \int_0^1 s^{\alpha-1} (1-s)^{c-\alpha-1} \prod_{i=1}^n (1-sx_i)^{-\beta_i} ds, x_i \in [1, \infty) \wedge \beta_i > 1$$

Θα περιορίσουμε στη μελέτη των περιπτώσεων όπου x_i παίρνει τιμές στο $(-\infty, 1]$ και $\alpha > 0, c-\alpha > 0$ οι οποίες παρουσιάζουν ενδιαφέρον στην περίπτωση της κατανομής Behrens Fisher¹⁶. Η περίπτωση όπου $n=1$ αντιστοιχεί στην υπεργεωμετρική συνάρτηση του Gauss, η περίπτωση $n=2$, αντιστοιχεί στη συνάρτηση Appell. Η απόδειξη της ασυμπτωτικής συμπεριφοράς για συναρτήσεις περισσότερων μεταβλητών αποδεικνύεται ότι είναι ανεξάρτητη από τον αριθμό των μεταβλητών n .

Η πρώτη περίπτωση $n=1$, αντιστοιχεί στην υπεργεωμετρική συνάρτηση του Gauss. Εφαρμόζουμε τους δύο διαδοχικούς μετασχηματισμούς του προηγούμενου τμήματος και καταλήγουμε στη μορφή¹⁷:

$$\left(\frac{e^{\frac{s}{2}}}{e^{\frac{s}{2}} + e^{\frac{s}{2}}} \right)^{\beta-1} \left(\frac{1}{1 + e^{\frac{1}{2}(-e^{-s} + e^s)\pi}} \right)^{c-\beta-1} \left(\frac{1 - e^{\frac{1}{2}(-e^{-s} + e^s)\pi}(x-1)}{1 + e^{\frac{1}{2}(-e^{-s} + e^s)\pi}} \right)^{-\alpha}$$

Διακρίνουμε δύο περιπτώσεις:

A) $u \rightarrow +\infty$

Ο πρώτος όρος τείνει ασυμπτωτικά στη σταθερά 1 για $\cdot$ για το δεύτερο όρο του γινομένου ισχύει ότι:

$$\left(\frac{1}{1 + e^{\frac{1}{2}(-e^{-s} + e^s)\pi}} \right)^{c-\beta-1} < \left(\frac{1}{e^{\frac{1}{2}(-e^{-s} + e^s)\pi}} \right)^{c-\beta-1} < e^{\frac{1}{2}(e^{-s} - e^s)\pi(c-\beta-1)} < (1-x)^{-\alpha} (1+\varepsilon_+) e^{\frac{1}{2}(-e^s)\pi(c-\beta-1)}$$

με $0 < \varepsilon_+ < 1$ δηλαδή μια διπλή εκθετική πτώση.

Η συμπεριφορά του τρίτου όρου για μεγάλες θετικές τιμές του u μπορεί να αποδειχθεί με τον κανόνα *L'Hospital* ότι τείνει στην τιμή $(1-x)^{-\alpha}$. Συνδυάζοντας τη συμπεριφορά των τριών όρων βλέπουμε ότι η προς ολοκλήρωση συνάρτηση έχει ασυμπτωτική

¹⁶ Με τη χρήση σχέσεων αναλυτικής συνέχισης είναι εφικτός ο μετασχηματισμός των συναρτήσεων Lauricella ώστε οι μεταβλητές τους να παίρνουν τιμές στο διάστημα $(-\infty, 1]$.

¹⁷ Εδώ παραλείπεται ο όρος που αντιστοιχεί στο μετασχηματισμό του διαφορικού γιατί λόγω του είδους του δεύτερου μετασχηματισμού, το διαφορικό της πρώτης ολοκλήρωσης μετατρέπεται στο γινόμενο μιας συνάρτησης με διπλή εκθετική πτώση και του νέου διαφορικού.

συμπεριφορά που καθορίζεται από το δεύτερο όρο, ο οποίος φθίνει διπλά εκθετικά, με ρυθμό φραγμένο προς τα πάνω από: $(1-x)^{-\alpha} (1+\varepsilon_+) e^{\frac{1}{2}(-e^s)(\pi(c-\beta-1))}$

B) $u \rightarrow -\infty$

Ο πρώτος όρος τείνει ασυμπτωτικά στο μηδέν · ο ρυθμός πτώσης στην περίπτωση αυτή βρίσκεται ότι είναι τουλάχιστον διπλός εκθετικός μια και:

$$\left(\frac{e^{\frac{s}{2}\pi}}{e^{\frac{s}{2}\pi} + e^{\frac{s}{2}\pi}} \right)^{\beta-1} = \left(\frac{1}{e^{e^{-s}\pi} + 1} \right)^{\beta-1} < \left(\frac{1}{e^{e^{-s}\pi}} \right)^{\beta-1} = e^{-e^{-|s|}\pi(\beta-1)}$$

Ο δεύτερος όρος και ο τρίτος όρος στην περίπτωση αυτή τείνει ασυμπτωτικά προς τη σταθερά 1.

Η εισαγωγή περισσότερων της μίας μεταβλητών δεν αλλάζει την βασική ασυμπτωτική συμπεριφορά της συνάρτησης προς ολοκλήρωση, η πτώση του οποίου εξαρτάται από τον πρώτο όρο (για $u \rightarrow -\infty$) και από τον δεύτερο όρο για $u \rightarrow +\infty$. Η εισαγωγή περισσότερων της μίας μεταβλητών επηρεάζει βέβαια την τιμή του s η οποία απαιτείται για να μειωθεί η τιμή της συνάρτησης κάτω από ένα συγκεκριμένο όριο (και άρα τον αριθμό όρων που πρέπει να χρησιμοποιηθούν για τον υπολογισμό της τιμής του ολοκληρώματος). Είναι η πρώτη φορά που ένα τέτοιο αποτέλεσμα αποδεικνύεται για τις συναρτήσεις Lauricella ανοίγοντας έτσι το δρόμο για την αυτοματοποίηση βιοπληροφορικών εργαλείων που τις χρησιμοποιούν. Λόγω του περιορισμένου ενδιαφέροντος της εφαρμογής του αποτελέσματος αυτού για τους σκοπούς της παρούσης εργασίας δεν προχωρούμε στην περαιτέρω διερεύνηση του.

4.5 Αριθμητικός υπολογισμός της κατανομής Behrens Fisher με τη μέθοδο του διπλού εκθετικού μετασχηματισμού

Εφαρμογή του διπλού εκθετικού μετασχηματισμού στη κατανομή Behrens Fisher

Στο προηγούμενο τμήμα είδαμε ότι η προς ολοκλήρωση συνάρτηση που εμφανίζεται στην πρώτη μορφή της κατανομής Behrens Fisher έχει τα ακόλουθα χαρακτηριστικά:

1. είναι αναλυτική στο διάστημα $[0, 1]$, χωρίς ανώμαλα σημεία (singularities) εντός του κλειστού διαστήματος ολοκλήρωσης
2. μπορούμε να βρούμε μια αντικατάσταση μεταβλητής $x = y(t)$, $0 = y(-\infty)$, $1 = y(\infty)$ τέτοια ώστε η πτώση της τιμής της συνάρτησης προς ολοκλήρωση να ακολουθεί εκθετικό $|f(y(t))y'(t)| \approx ce^{-\gamma|t|}$, $|t| \rightarrow \infty$ ή διπλό εκθετικό ρυθμό: $|f(y(t))y'(t)| \approx \exp(-ce^{\gamma|t|})$, $|t| \rightarrow \infty$.

Αποδείχτηκε ότι υπάρχουν τουλάχιστον δύο τέτοιοι μετασχηματισμοί που ικανοποιούν τις συνθήκες 1 και 2. Από τη βιβλιογραφία της αριθμητικής ανάλυσης ^{147;148;156} είναι γνωστό για ότι τέτοια τέτοια ολοκληρώματα ο βέλτιστος τρόπος ολοκλήρωσης (με το μικρότερο απαιτούμενο αριθμό υπολογισμών της προς ολοκλήρωση συνάρτησης) είναι η τραπεζοειδής μέθοδος. Η τελευταία εφαρμόζεται στην συνάρτηση που προκύπτει μετά από την εφαρμογή του διπλού εκθετικού μετασχηματισμού των Takahasi και Mori. Για ολοκληρώματα της γενικής μορφής

$\int_0^1 f(x)dx$ ο μετασχηματισμός αυτός είναι μια διαδικασία δύο βημάτων.

Συγκεκριμένα:

1. Στο πρώτο βήμα κάνουμε την αντικατάσταση $x \rightarrow \frac{x+1}{2}$ που έχει ως αποτέλεσμα την αλλαγή των ορίων ολοκλήρωσης από $[0,1]$ σε $[-1,1]$ και την αλλαγή της μορφής του ολοκληρώματος σε $(1/2) \int_{-1}^1 f((x+1)/2)dx$

2. Στη συνέχεια εφαρμόζουμε το μετασχηματισμό¹⁸ $\frac{x+1}{2} \rightarrow \tanh\left(\frac{\pi}{2} \sinh u\right)$ που οδηγεί στην αλλαγή των ορίων ολοκλήρωσης από $[-1,1]$ στο $[-\infty, \infty]$ και στην αλλαγή της μορφής του ολοκληρώματος $\int_{-\infty}^{\infty} f\left(\tanh\left(\frac{\pi}{2} \sinh u\right)\right) \frac{\pi/2 \cosh u}{\cosh^2(\sinh u \pi/2)} du$

3. Λόγω της διπλής εκθετικής πτώσης, η συνάρτηση εντός του ολοκληρώματος, τείνει γρήγορα προς το μηδέν για μεγάλες απόλυτες τιμές του u . Συνεπώς το σφάλμα που γίνεται κατά τη προσεγγίση του υπό εξέταση ολοκληρώματος με το $\int_{-L_-}^{L_+} f\left(\tanh\left(\frac{\pi}{2} \sinh u\right)\right) \frac{\pi/2 \cosh u}{\cosh^2(\sinh u \pi/2)} du$ γίνεται αρκετά μικρό για μεγάλες τιμές των L_- , L_+ .

Εφ' όσον ικανοποιούνται οι παραπάνω απαιτήσεις, η τιμή του ολοκληρώματος δίνεται από το άθροισμα $h \sum_{k=-N_-}^{N_+} f\left(\tanh\left(\frac{\pi}{2} \sinh kh\right)\right) \frac{\pi/2 \cosh kh}{\cosh^2(\sinh kh \pi/2)}$, όπου h το βήμα της ολοκλήρωσης.

Συγκριτική ανάλυση μεθόδων αριθμητικής ολοκλήρωσης για τον υπολογισμό πιθανοτήτων βασισμένων στην κατανομή Behrens-Fisher

Έχοντας καταλήξει στο ότι ο διπλός εκθετικός μετασχηματισμός, μπορεί να χρησιμοποιηθεί για να υπολογισθεί με βέλτιστο τρόπο η κατανομή Behrens Fisher θα προσπαθήσουμε να ποσοτικοποιήσουμε το κέρδος από την υιοθέτηση της μεθόδου αυτού. Για το σκοπό αυτό κατασκευάζουμε μια σειρά προγραμμάτων χρησιμοποιώντας τους αυτόματους αριθμητικούς ολοκληρωτές του συστήματος υπολογιστικής άλγεβρας *Mathematica* και μετρώντας τον αριθμό των υπολογισμών της συνάρτησης που εμφανίζεται εντός του ολοκληρώματος που ορίζει την κατανομή Behrens Fisher που απαιτείται για να υπολογισθεί το ολοκλήρωμα με δεδομένη ακρίβεια. Τα αποτελέσματα δίνονται γραφικά στα ακόλουθα σχήματα που συγκρίνουν τη μέθοδο του διπλού εκθετικού μετασχηματισμού με τη τραπεζοειδή μέθοδο (εφαρμοζόμενη στη μη μετασχηματισμένη μορφή της κατανομής Behrens

¹⁸ ο μετασχηματισμός εμφανίζεται στην πρώτη δημοσίευση της διπλής εκθετικής μεθόδου και είναι αυτός που χρησιμοποιείται στις υπάρχουσες βιβλιοθήκες αυτόματης αριθμητικής ολοκλήρωσης. Μετασχηματίζοντας το αρχικό ολοκλήρωμα στο διάστημα $[-1,1]$ μπορούμε να κάνουμε χρήση αποτελεσμάτων που έχουν εξαχθεί για αυτή την αρκετά μελετημένη περίπτωση χωρίς να ξανα-ανακαλύπτουμε τον τροχό.

Fisher) και τη μέθοδο Gauss-Kronrod. Το μέτρο της σύγκρισης είναι το πόσες φορές πρέπει να υπολογισθούν σημεία της συνάρτησης προς ολοκλήρωση. Όσες λιγότερες φορές πρέπει να γίνει αυτό, τόσο πιο αποδοτική είναι η μέθοδος ολοκλήρωσης για το προς εξέταση ολοκλήρωμα.

Τα γραφήματα αυτά δείχνουν ότι κατά μέσο όρο και για ένα μεγάλο εύρος τιμών διαφοράς βαθμών ελευθερίας και γωνιών ο υπολογισμός της πιθανότητας υποέκφρασης με βάση την μέθοδο του διπλού εκθετικού μετασχηματισμού απαιτεί περίπου 2-3 φορές λιγότερους υπολογισμούς της συνάρτησης προς ολοκλήρωση σχέση με τη μέθοδο Gauss Kronrod. Σε σχέση με την τραπεζοειδή μέθοδο, η μέθοδος του διπλού εκθετικού μετασχηματισμού παρουσιάζει βελτιωμένη απόδοση κατά 5-6 φορές τουλάχιστον. Τα αποτελέσματα του υπολογιστικού αυτού πειράματος μπορούν να δωθούν με τη μορφή ιστογράμματος, του λόγου των αριθμητικών υπολογισμών που απαιτούνται στην περίπτωση του διπλού εκθετικού μετασχηματισμού (DE) ως προς τη μέθοδο Gauss Kronrod (GK) και την απλή τραπεζοειδή (TR) για κάθε μια από τις περιπτώσεις που εξετάστηκαν. Τα ιστογράμματα δείχνουν ότι για ένα μεγάλο μέρος περιπτώσεων η διπλή εκθετική μέθοδος υπερτερεί (λόγος >1) ως προς τις άλλες δύο. Ένα μικρό μέρος των μετρήσεων χαρακτηρίζονται από λόγους μικρότερους του 1 και για τις δύο συγκρίσεις, γεγονός που οφείλεται στην ταχεία μείωση της συνάρτησης προς ολοκλήρωση για τιμές στη γειτονιά του 1 (αριθμητική near-singularity).

Εικόνα 16 Ιστογράμματα του λόγου του αριθμού υπολογισμών της συνάρτησης εντός του ολοκληρώματος της συνάρτησης Behrens Fisher όταν οι υπολογισμοί γίνονται με μία από τις μεθόδους Gauss Kronrod (GK) ή τραπεζοειδούς (TR) και της διπλής εκθετικής. Για την συντριπτική πλειοψηφία των παραμέτρων της κατανομής Behrens Fisher η διπλή εκθετική μέθοδος δίνει ακριβή αποτελέσματα με μικρότερο αριθμό υπολογισμών.

Εικόνα 17 Γραφική αναπαράσταση του αριθμού σημείων της συνάρτησης που εμφανίζεται στο ολοκλήρωμα της συνάρτησης Behrens Fisher που απαιτούνται (άξονας Z) για τον υπολογισμό της με ακρίβεια 6 δεκαδικών ψηφίων για διάφορες τιμές της σταθμισμένης διαφοράς μεταξύ των δειγματικών μέσων τιμών (δ) , βαθμών ελευθερίας (άξονες X, Y) και γωνιών (ϕ). Το αποτελέσματα αυτά αντιστοιχούν στη μέθοδο Gauss Kronrod.

Εικόνα 18 Γραφική αναπαράσταση του αριθμού σημείων της συνάρτησης που εμφανίζεται στο ολοκλήρωμα της συνάρτησης Behrens Fisher που απαιτούνται (άξονας Z) για τον υπολογισμό της κατανομής με ακρίβεια 6 δεκαδικών ψηφίων για διάφορες τιμές της σταθμισμένης διαφοράς μεταξύ των δειγματικών μέσων τιμών (δ), βαθμών ελευθερίας (άξονες X, Y) και γωνιών (ϕ). Το αποτελέσματα αυτά αντιστοιχούν στο διπλό εκθετικό μετασχηματισμό.

Εικόνα 19 Γραφική αναπαράσταση του αριθμού σημείων της συνάρτησης που εμφανίζεται στο ολοκλήρωμα της συνάρτησης Behrens Fisher που απαιτούνται (άξονας Z) για τον υπολογισμό της με ακρίβεια 6 δεκαδικών ψηφίων για διάφορες τιμές της σταθμισμένης διαφοράς μεταξύ των δειγματικών μέσων τιμών (δ) , βαθμών ελευθερίας (άξονες X, Y) και γωνιών (ϕ). Το αποτελέσματα αυτά αντιστοιχούν στη τραπεζοειδή μέθοδο.

Σχεδιασμός και υλοποίηση λογισμικού για τον υπολογισμό της κατανομής Behrens Fisher

Γενική δομή και προδιαγραφές του λογισμικού υπολογισμού της κατανομής Behrens Fisher

Έχοντας κατασκευάσει αρχικά πρωτότυπα που πιστοποιούν την εγκυρότητα του προτεινόμενου αλγορίθμου εντός μιας υψηλής γλώσσας προγραμματισμού, στρέφουμε την προσοχή μας στην σχεδίαση και υλοποίηση λογισμικού που να υλοποιεί τους παραπάνω υπολογισμούς εντός μιας μεταγλωτιζόμενης (compiled) γλώσσας προγραμματισμού. Η υλοποίηση τέτοιων πρωτοτύπων προσφέρει τα ακόλουθα πλεονεκτήματα σε σχέση με τα αρχικά πρωτότυπα εντός του *Mathematica*:

1. ταχύτητα εκτέλεσης (π.χ. ακόμα και με τη μέθοδο του διπλού εκθετικού μετασχηματισμού, ο υπολογισμός 10000 τέτοιων τιμών πιθανότητας απαιτεί περίπου 3 λεπτά εντός του *Mathematica* χρόνος που καθιστά απαγορευτική την διενέργεια συγκριτικών δοκιμών μέσω «τεχνητών» συνόλων δεδομένων
2. δυνατότητα εκτέλεσης σε συστήματα που δεν έχουν το *Mathematica*
3. δυνατότητα ενσωμάτωσης σε προϋπάρχοντα πακέτα στατιστικής επεξεργασίας (π.χ. R) ή πακέτα επεξεργασίας δεδομένων αποτύπωσης (BioPERL, BioConductor)

Οι δυσκολίες που έπρεπε να υπερνικηθούν για την επιτυχή έκβαση αυτής της προσπάθειας ήταν:

1. περιορισμένη ακρίβεια των βιβλιοθηκών κινητής υποδιαστολής εκτός του *Mathematica* (το οποίο υποστηρίζει πράξεις μεταβλητής ακρίβειας). Απόρροια της διαφορετικής αυτής ακρίβειας, είναι ότι υπολογισμοί δυνητικά εφικτοί εντός του *Mathematica* να μην μπορούν να γίνουν σε συστήματα μικρότερης ακρίβειας
2. ανάγκη χρήσης βιβλιοθηκών ειδικών συναρτήσεων (π.χ. συναρτήσεις B, Γ, υπεργεωμετρικών) που συνήθως είναι γραμμένες σε γλώσσες προγραμματισμού που τυπικά ΔΕΝ χρησιμοποιούνται σε εφαρμογές βιοπληροφορικής
3. ανάγκη ευελιξίας στη διασύνδεση (interfacing) με υπάρχοντα συστήματα ανάλυσης δεδομένων μικροσυστοιχειών που παρέχεται από scripting γλώσσες προγραμματισμού (PERL) που όμως δεν προσφέρονται για αριθμητικούς υπολογισμούς

Η ευρετική προσέγγιση που υιοθετήθηκε ήταν η ακόλουθη:

1. χρήση αναλυτικών συνεχίσεων για τις πράξεις που περιελάμβανουν τις υπεργεωμετρικές συναρτήσεις ώστε να αποφύγουμε να κάνουμε υπολογισμούς σε διαστήματα του άξονα των πραγματικών αριθμών όπου οι τιμές των συναρτήσεων αυτών βρίσκονται εκτός των αντιπροσωπεύσιμων αριθμών από το 32-bit IEEE πρότυπο.
2. υλοποίηση του προγράμματος για τον υπολογισμό της πιθανότητας με βάση της κατανομής Behrens Fisher σε περισσότερες τις μιας γλώσσες προγραμματισμού. Προϋπάρχον κώδικας σε κάθε μια από τις γλώσσες που χρησιμοποιήθηκε παρείχε τμήματα της απαιτούμενης συνολικής λειτουργικότητας · τα μοναδικά τμήματα που χρειάστηκε να γραφτούν από το μηδέν είχαν να κάνουν με τον υπολογισμό των τιμών της προς ολοκλήρωση συνάρτησης σε επίπεδο αφαίρεσης παρόμοιο με αυτό του *Mathematica* καθώς και τα τμήματα που είχαν να κάνουν με τη διασύνδεση του κώδικα στις

διάφορες γλώσσες προγραμματισμού. Έτσι κατέστη δυνατή η χρησιμοποίηση υπάρχουσων βιβλιοθηκών ειδικών συναρτήσεων και αριθμητικής ολοκλήρωσης μειώνοντας κατά πολύ το χρόνο συνολικής ανάπτυξης του ειδικού λογισμικού.

Οι γλώσσες προγραμματισμού που χρησιμοποιήθηκαν και οι αντίστοιχες λειτουργικότητες είναι οι ακόλουθες (βλέπε και επόμενο σχήμα για γραφική αναπαράσταση των σχέσεων μεταξύ τους):

- *Fortran 77*: Ειδικές συναρτήσεις Βήτα, Γάμμα και υπεργεωμετρικές συναρτήσεις όπως αυτές βρίσκονται στο βιβλίο των Shanjie Zhang and Jianming¹⁵⁷. Ο κώδικας σε Fortran διατίθεται στην ιστοσελίδα: <http://jin.ece.uiuc.edu/>
- *C*: Αυτόματη συνάρτηση ολοκλήρωσης με βάση τη μέθοδο διπλού εκθετικού μετασχηματισμού από τον T. Ooura <http://momonga.t.u-tokyo.ac.jp/~ooura/>. Στη C γράφτηκαν επίσης και οι ρουτίνες υπολογισμού της συνάρτησης προς ολοκλήρωση· η δυνατότητα απευθείας κλήσης συναρτήσεων από τη C στη Fortran επέτρεψε την χρήση των συναρτήσεων της Fortran χωρίς περαιτέρω παρεμβάσεις
- *C++*: για την παραγωγή interfaces από τη C προς τη γλώσσα PERL μέσω της οποίας ο κώδικας είναι προσπελάσιμος από εξωτερικά προγράμματα. Η παραγωγή του κώδικα αυτού έγινε αυτόματα με το σύστημα SWIG (<http://www.swig.org>)
- *PERL*: για τη χρήση του κώδικα από προγράμματα ανάλυσης δεδομένων μικροσυστοιχειών.

Εικόνα 20 Σχηματική αναπαράσταση των διαφόρων τμήματων του λογισμικού για τον υπολογισμό της κατανομής Behrens Fisher

Μαθηματικοί μετασχηματισμοί της προς ολοκλήρωση συνάρτησης

Για τους σκοπούς της ολοκλήρωσης της κατανομής Behrens Fisher χρησιμοποιήθηκε η έκφραση 2 λόγω της έλλειψης αριθμητικών βιβλιοθηκών που να μπορούν να υπολογίζουν την τιμή της αντίστοιχης υπεργεωμετρικής για μεγάλες (αρνητικές) τιμές της μεταβλητής της. Η μεταβλητή της υπεργεωμετρικής της έκφρασης 2 είναι αυστηρά μικρότερη της μονάδας, περίπτωση που καλύπτεται από τις υπάρχουσες βιβλιοθήκες για τον υπολογισμό των υπεργεωμετρικών συναρτήσεων. Λόγω όμως του τρόπου υπολογισμού των τιμών της υπεργεωμετρικής από τις υπάρχουσες βιβλιοθήκες¹⁹ οι αλγόριθμοι αυτοί δίνουν ικανοποιητικά αποτελέσματα για τιμές της μεταβλητής μικρότερης του 0.5. Συνεπώς κατά τη διάρκεια της ολοκλήρωσης, ο υπολογισμός της υπεργεωμετρικής συνάρτησης γίνεται από την έκφραση:

$${}_2F_1\left(\frac{\nu_1 + \nu_2 + 1}{2}, 1; \frac{3}{2}; \frac{\delta^2 x(1-x)}{\delta^2 x(1-x) + \nu_1 \beta_1^2 x + \nu_2 \beta_2^2 (1-x)}\right) \text{ αν } \frac{\delta^2 x(1-x)}{\delta^2 x(1-x) + \nu_1 \beta_1^2 x + \nu_2 \beta_2^2 (1-x)} < 0.5$$

Στην περίπτωση που η μεταβλητή της υπεργεωμετρικής είναι μεγαλύτερη από 0.5, εφαρμόζουμε τη σχέση :

$${}_2F_1(a, b, c, z) = \frac{\Gamma(c) \Gamma(-a-b+c) {}_2F_1(a, b, a+b-c+1; 1-z)}{\Gamma(c-a) \Gamma(c-b)} + \frac{(1-z)^{-a-b+c} \Gamma(a+b-c) \Gamma(c) {}_2F_1(c-a, c-b, -a-b+c+1; 1-z)}{\Gamma(a) \Gamma(b)}$$

που έχει ως αποτέλεσμα το γράψιμο της υπεργεωμετρικής με τη μορφή αθροίσματος δύο όρων:

$$\frac{\sqrt{\pi} \Gamma\left(\frac{1}{2}(\nu_1 + \nu_2)\right) (-u \beta_1 \nu_1 + u \beta_2 \nu_2 - \beta_2 \nu_2)^{-\frac{\nu_1}{2} - \frac{\nu_2}{2}} (u^2 t^2 - t^2 u - u \beta_1 \nu_1 + u \beta_2 \nu_2 - \beta_2 \nu_2)^{\frac{\nu_1}{2} + \frac{\nu_2}{2} + \frac{1}{2}}}{2t \sqrt{u-1} \sqrt{u} \Gamma\left(\frac{1}{2}(\nu_1 + \nu_2 + 1)\right)} \\ - \frac{\Gamma\left(\frac{1}{2}(\nu_1 + \nu_2)\right) {}_2F_1\left(1, \frac{1}{2}(\nu_1 + \nu_2 + 1); \frac{1}{2}(\nu_1 + \nu_2 + 2); \frac{-u \beta_1 \nu_1 + u \beta_2 \nu_2 - \beta_2 \nu_2}{u^2 t^2 - t^2 u - u \beta_1 \nu_1 + u \beta_2 \nu_2 - \beta_2 \nu_2}\right)}{2 \Gamma\left(\frac{1}{2}(\nu_1 + \nu_2 + 2)\right)}$$

μετά από διαδοχικές απλοποιήσεις μέσα από το *Mathematica*.

Αριθμητική επικύρωση του κώδικα υπολογισμού της κατανομής Behrens Fisher

Το επόμενο βήμα μετά την υλοποίηση του αριθμητικού κώδικα υπολογισμού της κατανομής Behrens Fisher ήταν η συγκριτική ανάλυση των αποτελεσμάτων που λαμβάνονται από το πρόγραμμα αυτό, σε σχέση με τα αποτελέσματα που

¹⁹ Η βιβλιοθήκη που χρησιμοποιήθηκε, υπολογίζει την υπεργεωμετρική συνάρτηση προσέγγιζοντας το άπειρο άθροισμα με το ένα πεπερασμένο μερικό άθροισμα (partial sum) που τερματίζεται όταν η ζητούμενη σχετική ακρίβεια της προσέγγισης έχει επιτευχθεί.

λαμβάνονται από τους αριθμητικούς ολοκληρωτές του *Mathematica*. Η σύγκριση των αποτελεσμάτων, θα απαντήσει στο ερώτημα της αριθμητικής σταθερότητας των προτεινόμενων αλγορίθμων εκτός του ελεγχόμενου περιβάλλοντος ενός συστήματος συμβολικής άλγεβρας που παρέχει δυνατότητες αριθμητικών πράξεων με μεταβαλλόμενη ακρίβεια. Για τους σκοπούς της αριθμητικής σύγκρισης αυτής, υπολογίστηκαν οι τιμές της κατανομής Behrens Fisher για :

- ο βαθμοί ελευθερίας 1-20 (δείγμα 1)
- ο βαθμοί ελευθερίας 1-20 (δείγμα 2)
- ο σταθμισμένη δειγματική διαφορά από 0.1 ως 5.1 με μέγεθος βήματος 1.0
- ο τιμή γωνίας από 0.1 ως 1.4 (σε ακτίνια) με μέγεθος βήματος 0.1

Η επιλογή των παραπάνω αριθμών έγινε με σκοπό την ανάδειξη της αριθμητικής συμπεριφοράς του αλγορίθμου σε περιπτώσεις μεγέθους δειγμάτων χαρακτηριστικών για πειράματα μικροσυστοιχειών και για ένα μεγάλο εύρος τιμών δειγματικών διακυμάνσεων και μέσων τιμών. Στη συνέχεια υπολογίστηκαν οι τιμές της κατανομής Behrens Fisher με τη βοήθεια των πρωτοτύπων που είχαν προηγούμενα αναπτυχθεί στο *Mathematica*. Δεχόμενοι ότι οι τελευταίες αποτελούν την «αληθινή» τιμή της κατανομής, υπολογίσαμε στη συνέχεια το απόλυτο και το σχετικό αριθμητικό σφάλμα (numerical accuracy). Τα αποτελέσματα δίνονται γραφικά στα ακόλουθα σχήματα με για το σύνολο των μετρήσεων, σαν συνάρτηση της γωνίας, της σταθμισμένης δειγματικής διαφοράς. Παρά τις διαφορές μεταξύ των δύο αριθμητικών συστημάτων, βλέπουμε ότι καταλήγουν σε αποτελέσματα που συμφωνούν μέχρι το 6^ο και το 7^ο δεκαδικό ψηφίο για τις διαφορετικές γωνίες.

Εικόνα 21 Μέσο απόλυτο σφάλμα της υλοποίησης εντός του Mathematica σε σχέση με την υλοποίηση στη γλώσσα προγραμματισμού C για διάφορες τιμές της γωνίας της κατανομής Behrens Fisher. Τα αποτελέσματα συμφωνούν μέχρι το 6^ο και 8^ο δεκαδικό ψηφίο

Εικόνα 22 Μέσο απόλυτο σφάλμα της υλοποίησης εντός του Mathematica σε σχέση με την υλοποίηση στη γλώσσα προγραμματισμού C για διάφορες τιμές της σταθμισμένης δειγματικής διαφοράς που εμφανίζεται στην κατανομή Behrens Fisher. Τα αποτελέσματα συμφωνούν μέχρι το 5^ο και 6^ο δεκαδικό ψηφίο

5 Ανάλυση πειραμάτων μικροσυστοιχειών για την ανακάλυψη μεταγραφικών καταστολέων

Στις επόμενες παραγράφους αναλύεται η διαδικασία επεξεργασίας του σετ δεδομένων μικροσυστοιχειών που χρησιμοποιήθηκε για την ανίχνευση μεταγραφικών καταστολέων στα παρθενικά T κύτταρα. Το τμήμα αυτό της διατριβής διαγράφει τα προβλήματα που έπρεπε να αντιμετωπισθούν και τον τρόπο επίλυσης τους τόσο σε θεωρητικό όσο και σε τεχνικό επίπεδο. Εσκεμμένα παραλείπεται η τεχνική επεξεργασία του μαθηματικού προβλήματος υπολογισμών πιθανοτήτων με βάση την κατανομή Behrens Fisher γιατί έχει καλυφθεί σε προηγούμενο τμήμα. Από την πλευρά της ταξινόμησης του κεφαλαίου αυτού σε μια από τις κατηγορίες που συνηθίζονται να ομαδοποιούνται τα επιστημονικά κείμενα, τα στοιχεία που δίνονται στις επόμενες σελίδες αντιστοιχούν τόσο στο κομμάτι της μεθοδολογίας όσο και στο κομμάτι των αποτελεσμάτων.

5.1 Πηγές δεδομένων

Ένα από τα μεγαλύτερα σετ δεδομένων αποτύπωσης λεμφοκυττάρων δημιουργήθηκε από τον Alizadeh και συνεργάτες το 2000 ^{5:158-160} σε μία προσπάθεια χαρακτηρισμού μοριακών προγνωστικών δεικτών σε Non Hodgkin λεμφώματα. Το σετ δεδομένων περιελάμβανε μετρήσεις 17853 ιχνηθετών (probes) που προήλθαν από μια πληθώρα B και T κυτταρικών σειρών και περιελάμβανε ένα μεγάλο αριθμό μεταγραφικών παραγόντων που εκφράζονται σε κύτταρα του ανοσοποιητικού. Οι αλληλουχίες της συντριπτικής πλειοψηφίας των ιχνηθετών βρίσκονται κατατεθειμένες στην GENBANK επιτρέποντας την αναζήτηση ομολόγων μέσω BLAST/UNIGENE αλλά και το σχεδιασμό πληθώρας επιβεβαιωτικών πειραμάτων όπως για παράδειγμα αποτύπωσης Northern, PCRs. Κατά τη διάρκεια των μετρήσεων οι πρωτογενείς ερευνητές είχαν χρησιμοποιήσει ένα κοινό βιολογικό δείγμα αναφοράς για όλες τις μετρήσεις μικροσυστοιχειών που θα μπορούσε να αποτελέσει τη βάση για τον έλεγχο υποθέσεων σχετικά με διάφορα στάδια της ανάλυσης των δεδομένων.

Πριν την αξιοποίηση του σετ δεδομένων, έπρεπε να επιλυθεί μια σειρά προβλημάτων:

1. Στην αρχική δημοσίευση είχε χρησιμοποιηθεί μια τεχνική ομαδοποίησης μετρήσεων για την ανακάλυψη ομάδων προτύπων έκφρασης. Καμία από αυτές τις ομάδες δεν αναφερόταν σε γονίδια που εκφράζονταν διαφορετικά σε ομάδες λεμφοκυττάρων
2. Η αντιστοίχιση των EST ιχνηθετών σε γονίδια δεν ήταν γνωστή για την πλειοψηφία των γονιδίων. Επιπλέον, για αρκετούς από τους ιχνηθέτες η αναζήτηση ομολογίας κατά BLAST είχε γίνει όχι στο υποσύνολο της GENBANK των δεδομένων του *H. sapiens*, αλλά στο σύνολο της πρώτης. Συνεπώς για ένα υπολογίσιμο αριθμό ιχνηθετών η αντιστοίχιση με το γονίδιο με το μεγαλύτερο E-score και όχι με το ανθρώπινο γονίδιο με το μεγαλύτερο E-score ήταν αυτή που εμφανιζόταν στα αρχεία των πρωτογενών δεδομένων.
3. Για όλους τους ιχνηθέτες το αποτέλεσμα της αναζήτησης κατά BLAST δίνεται με το *όνομα* του αντίστοιχου γονιδίου και όχι με τον μοναδιαίο κωδικό με τον οποίο το γονίδιο αποθηκεύεται στην υποσύνολο αλληλουχιών αναφοράς της GENBANK. Το αποτέλεσμα είναι ότι πληροφορίες διαθέσιμες από τις δευτερογενείς βάσεις δεδομένων (π.χ. σχολιογράφηση με τη βοήθεια των Gene Ontology Terms) δεν ήταν άμεσα προσπελάσιμες.
4. Μόνο ένα μικρό σύνολο δεδομένων ήταν σχετικό με το ερώτημα που θέλαμε να απαντήσουμε. Το μέγεθος των αντίστοιχων δειγμάτων ήταν αρκετά μικρό

και έπρεπε να βρεθεί ένας τρόπος να χρησιμοποιηθεί ο μεγαλύτερος δυνατός αριθμός μετρήσεων χωρίς όμως να αυξηθεί ο θόρυβος. Η λογική συνέπεια ήταν η ανάγκη ανάλυσης των μετρήσεων από το επίπεδο της εικόνας.

5. Υπολογίσιμος αριθμός *διαφορετικών* ιχνηθετών αντιστοιχούσε στο ίδιο γονίδιο και έπρεπε να βρεθεί ένας τρόπος να συνδυασθεί η πληροφορία αυτή ώστε ο υπολογισμός της πεποίθησης για την πιθανότητα υποέκφρασης να μπορεί να χρησιμοποιήσει το σύνολο αυτών των μετρήσεων.
6. Η επιλογή της τιμής του κατωφλίου σημαντικότητας για την υπόθεση υποέκφρασης έπρεπε να γίνει με τρόπο που θα μείωνε την πιθανότητα ψευδώς θετικών αποτελεσμάτων.

Πηγή mRNA	Αριθμός Πειραμάτων
B lymphocytes (peripheral unstimulated)	2
B lymphocytes (tonsilar)	2
B lymphocytes (cord blood unstimulated)	1
Memory B lymphocytes CD27+	2
Memory B lymphocytes CD27-	1
Memory B lymphocytes (tonsilar)	2
B lymphocytes (peripheral stimulated using various protocols employing IgM, CD40L,IL-4)	12
Activated CD4+ T lymphocytes	3
CD4+ Naive T lymphocytes	3
Lymph node cells	1
Tonsilar centroblasts	1
CLL (Chronic Lymphocytic Leukemia)	12
DLCL (Diffuse Large Cell Lymphoma)	61
FL (Follicular Lymphoma)	9
Leukemic Cell lines (Jurkat, U937, OCI,SUDHL,WSU1)	10

Πίνακας 2 Πηγή προέλευσης mRNA στο σετ των μετρήσεων που χρησιμοποιήθηκε για τον υπολογισμό της πιθανότητας υποέκφρασης. Οι μετρήσεις που είναι σχετικές με το υπό εξέταση σύστημα δίνονται με έντονη γραμματοσειρά

Η σκιαγράφηση του τρόπου προσέγγισης των τεχνικών αυτών προβλημάτων αναλύεται στις επόμενες παραγράφους μεθοδολογίας. Για τους σκοπούς της εργασίας αυτής τα πρωτογενή δεδομένα όπως ήταν κατατεθειμένα στη βάση μικροσυστοιχειών του Stanford αποτέλεσαν την αφετηρία για τις αναλύσεις των επόμενων τμημάτων

Αντιστοίχιση ιχνηθετών σε γονίδια

Για την αντιστοίχιση των ιχνηθετών που βρίσκονταν πάνω στις μικροσυστοιχειές σε γονίδια χρησιμοποιήθηκε η ακόλουθη διαδικασία:

1. Η βάση dbEST εγκαταστάθηκε τοπικά αφού πρώτα τα αντίστοιχα συμπιεσμένα αρχεία κατέβηκαν από το NCBI. Συγκεκριμένα τα συμπιεσμένα αρχεία est.0x.tar.gz και τα αρχεία «μάσκας» ανθρώπινων ESTs est_human.tar.gz κατεβάστηκαν από τη διεύθυνση του διαδικτύου <ftp.ncbi.nih.gov/blast/db> . Στη συνέχεια τα αρχεία αποσυμπίεστηκαν και

δεικτοδοτήθηκαν (indexed) από το πρόγραμμα formatdb διαθέσιμο από το NCBI.

2. Η βάση αλληλουχιών αναφοράς των mRNA που αντιστοιχούν σε ανθρώπινα γονίδια εγκαταστάθηκε με παρόμοιο τρόπο όπως η dbEST. Η βάση αυτή βρίσκεται στο NCBI σαν ftp.ncbi.nih.gov/refseq/H_sapiens/mRNA_Prot/human.rna.fna.gz (FASTA) ή human.rna.gbff.gz (Σχολιογραφημένη)
3. Οι αριθμοί πρόσβασης (accession numbers) των ιχνηθετών που ήταν τυπωμένοι στη μικροσυστοιχεία εξάχθηκαν από τα αρχεία των πρωτογενών δεδομένων και χρησιμοποιήθηκαν για την εξαγωγή των αλληλουχιών των από τη βάση dbEST που είχε εγκατασταθεί τοπικά στο βήμα 1.
4. Οι αλληλουχίες που ελήφθησαν στο προηγούμενο βήμα χρησιμοποιήθηκαν για την αναζήτηση κατά BLAST στη βάση του βήματος 2. Τα αποτελέσματα της αναζήτησης επεξεργάστηκαν με φίλτρα διαθέσιμα από την BIoPERL και αποθηκεύτηκαν για το επόμενο βήματα.

Επεξεργασία εικόνων μικροσυστοιχειών

Για την επεξεργασία των εικόνων των μικροσυστοιχειών χρησιμοποιήθηκε το ακόλουθο μοντέλο:

Έστω y η καταγεγραμμένη ένταση της κηλίδας όπως αυτή δίνεται από το λογισμικό τμηματοποίησης των πρωτογενών μετρήσεων. Προτείνεται ότι το y έχει την κατανομή μιας μεταβλητής που δίνεται από τη σχέση:

$y=e^x$ όπου x τυχαία μεταβλητή με κανονική κατανομή μέσης τιμής μ και τυπικής απόκλισης σ . Στην περίπτωση αυτή η συνάρτηση πυκνότητας πιθανότητας του y δίνεται από τη σχέση :

$$PDF(y) = \frac{e^{-\frac{(\log y - \mu)^2}{2\sigma^2}}}{\sqrt{2\pi} y \sigma}$$

Η εκτίμηση των παραμέτρων της κατανομής του τοπικού υποστρώματος κάθε κηλίδας γίνεται με βάση τις μετρήσεις του υποστρώματος των γειτονικών της. Η δειγματική μέση τιμή και η δειγματική τυπική απόκλιση χρησιμοποιούνται σαν εκτιμητές των αντίστοιχων πληθυσμιακών τιμών. Ανάλογα με τη θέση κάθε κηλίδας

αξιοποιείται διαφορετικό είδος δείγματος σύμφωνα με ένα σύστημα γειτνίασης δεύτερης τάξης (διπλανό σχήμα)

Οι κηλίδες εκείνες το σήμα των οποίων βρισκόταν εκτός του 95% διαστήματος εμπιστοσύνης της κατανομής του θορύβου υποστρώματος φιλτράρονταν από τα επόμενα βήματα της ανάλυσης. Ιχνηθέτες οι οποίες δεν

είχαν μετρήσιμο σήμα με αυτή την στατιστική έννοια σε ποσοστό $> 60\%$ όλων των πειραμάτων, αποκλείστηκαν από όλες τις επόμενες αναλύσεις ακόμα και για πειράματα στα οποία είχαν δώσει σήμα $> 95\%$ του τοπικού θορύβου υποστρώματος.

Κανονικοποίηση μετρήσεων

Έχοντας φιλτράρει στατιστικά τις μετρήσεις των κηλίδων, προχωρήσαμε στην κανονικοποίηση των μετρήσεων, χρησιμοποιώντας τη μέθοδο κανονικοποίησης

γνωστή ως *global median normalization*, η οποία «κεντράρει» την κατανομή των μετρήσεων ώστε να έχουν διάμεσο ίσο με το μηδέν. Η μέθοδος αυτή χρησιμοποιεί όλες τις μετρήσεις μιας μικροσυστοιχείας σαν εσωτερικό δείγμα αναφοράς (control). Η έμμεση παραδοχή είναι ότι οι περισσότερα από τις μετρήσεις δεν θα μεταβληθούν ανάμεσα στις δύο καταστάσεις που συγκρίνονται, και ότι το ποσοστό των γονιδίων που υποεκφράζονται και υπερεκφράζονται είναι το ίδιο στις δύο καταστάσεις που συγκρίνονται. Διάφοροι λόγοι βρίσκονταν πίσω από την απόφαση να χρησιμοποιηθεί η μέθοδος αυτή κανονικοποίησης:

1. Πρώτα από όλα ο λόγος των επιπέδων έκφρασης δεν επιδέχεται σαφούς ερμηνείας όταν ένα από τα κανάλια είναι δείγμα αναφοράς (όπως στην περίπτωση μας) .
2. Δεύτερον ο εξαιρετικά μικρός συντελεστής μεταβλητότητας των μη κανονικοποιημένων μετρήσεων του σετ που χρησιμοποιήθηκε κάνει την ανάγκη χρήσης πιο «τεχνητών» μεθόδων κανονικοποίησης λιγότερο επιτακτική.
3. Τέλος, αυτή η μέθοδος κανονικοποίησης μπορεί να εφαρμοσθεί μόνο εφόσον έχουμε έγκυρες μετρήσεις και στα δύο κανάλια για μία δεδομένη κηλίδα. Εφόσον ένα από τα δύο κανάλια δεν έδωσε έγκυρη μέτρηση τότε ο λόγος έκφρασης δεν ορίζεται, οδηγώντας σε μία δραματική αύξηση του αριθμού των ελλειπών μετρήσεων (missing data).

Τόσο η επιλογή μεθόδου επεξεργασίας εικόνας όσο και αυτή μεθόδου κανονικοποίησης είχαν μελετηθεί σε προηγούμενο στάδιο αυτής της εργασίας και είχε δειχθεί ότι ο συνδιασμός των δύο οδηγεί σε εξαιρετικά χαμηλό συντελεστή μεταβλητότητας μετρήσεων με το μικρότερο δυνατό αριθμό φιλτραρισμένων μετρήσεων και παρά τη διατήρηση εξαιρετικά ασθενών σημάτων στο τελικό σετ δεδομένων.

Ομαδοποίηση μετρήσεων σε ανταλλάξιμα δείγματα

Όπως αναφέρθηκε προηγούμενα ένα από τα προβλήματα που έπρεπε να αντιμετωπισθεί ήταν ο εξαιρετικά μικρός αριθμός των δειγμάτων ενδιαφέροντος στο σετ δεδομένων (π.χ. ο αριθμός υβριδοποιημένων δειγμάτων από T κύτταρα ήταν 3 τόσο για τα παρθενικά όσο και για τα ενεργοποιημένα δείγματα). Από την άλλη, κάθε μια από τις μικροσυστοιχείες περιέλαμβανε διαφορετικά επίπεδα πολλαπλότητας: αρκετοί από τους ιχνηθέτες είχαν εκτυπωθεί πάνω από μία φορές σε διαφορετικά σημεία της συστοιχείας, ενώ υπήρχαν και γονίδια των οποίων η έκφραση είχε αποτυπωθεί σε περισσότερους του ενός ιχνηθέτες από διαφορετικά σημεία της αλληλουχίας του mRNA. Η λογική συνέπεια των δύο αυτών παρατηρήσεων είναι ότι για μια σημαντική μειοψηφία των γονιδίων, η συμπερασματολογία αναφορικά με την υπόεκφραση ή υπερέκφραση θα μπορούσε να βασιστεί σε δείγματα μεγέθους μεγαλύτερου του αριθμού των πειραματικών δειγμάτων.

Το ερώτημα που φυσικά ανακύπτει στο σημείο αυτό είναι με ποιο τρόπο θα μπορούσε να γίνει η ταξινόμηση των παρατηρήσεων σε ανταλλάξιμα δείγματα; Η έννοια της ανταλλαξιμότητας έχει κομβικό ρόλο στη Bayesian στατιστική και σχετίζεται με την ισοδύναμη έννοια της μη-προτίμησης. Συγκεκριμένα δύο ή περισσότερες μετρήσεις, θεωρούνται ανταλλάξιμες όταν η μία μπορεί να υποκαταστήσει για την άλλη σε οποιαδήποτε τυπική διαδικασία συμπερασματολογίας⁹⁶. Όμως για μετρήσεις με ισχυρό χωρικό χαρακτήρα (και οι μετρήσεις μικροσυστοιχειών είναι τέτοιες), το να ισχυρισθεί κανείς ότι οι πρωτογενείς μετρήσεις (δηλαδή τα σήματα από τις κηλίδες στην επιφάνεια της συστοιχείας) είναι ανταλλάξιμες δεν ευσταθεί. Λόγω της χωρικής κατανομής του θορύβου, τα σήματα

που προέρχονται από διαφορετικές κηλίδες στις οποίες έχει τυπωθεί ο ίδιος ιχνηθέτης θα διαφέρουν λόγω π.χ. του διαφορετικού μεγέθους της εκτυπωτικής ακίδας που χρησιμοποιήθηκε κατά τη διάρκεια παρασκευής της συστοιχείας, της διαφορετικής θέσης (μετρήσεις στην περιφέρεια της συστοιχείας έχουν μεγαλύτερο θόρυβο) κλπ. Από την άλλη, μετρήσεις που προέρχονται από την ίδια κηλίδα, πάντα μπορούν να θεωρηθούν ανταλλάξιμες και μπορούν να αποτελέσουν τη βάση τέτοιων δειγμάτων, ειδικά αν αξιοποιούνται σε καταστάσεις που αντιστοιχούν στο πρόβλημα Behrens Fisher δηλαδή στη διαφορά του μέσου όρου των μετρήσεων αυτών. Εφόσον η διαδικασία κανονικοποίησης έχει επιτύχει να εξαλείψει τις χωροθετικές πηγές σφάλματος, μετρήσεις που προέρχονται από τον ίδιο ιχνηθέτη θα είναι ανταλλάξιμες ανεξάρτητα από τη θέση του. Μετρήσεις από διαφορετικούς ιχνηθέτες ΔΕΝ είναι ανταλλάξιμες λόγω του διαφορετικού μεγέθους και σύστασης σε επίπεδο βάσεων των ιχνηθετών που επηρεάζουν την αποτελεσματικότητα της υβριδοποίησης και άρα τη διαδικασία δημιουργίας σημάτων. Αντίστοιχα αν είχαμε έναν αλγόριθμο κανονικοποίησης που να μπορούσε να φιλτράρει όλες τις πηγές της μετρητικής ετερογένειας, θα μπορούσαμε να ομαδοποιήσουμε τις μετρήσεις από το ίδιο γονίδιο σε ανταλλάξιμα δείγματα ανεξάρτητα από το είδος ή τη θέση του ιχνηθέτη που δημιούργησε τα σήματα αυτά.

Οι παραπάνω συλλογισμοί αν και υποδεικνύουν ένα τρόπο ομαδοποίησης μετρήσεων, δεν επιλύουν το αρχικό πρόβλημα της αξιοποίησης της δυνητικής πολλαπλότητας των μετρήσεων για την ενίσχυση της σταθερότητας της συμπερασματολογίας που πηγάζει από αυτές. Για το σκοπό αυτό προτείνουμε παρακάτω ένα λογικοφανές μοντέλο (plausible model) που μπορεί να αξιοποιηθεί στην πολλαπλότητα των μετρήσεων αυτών.

Συνδυασμός πεποιθήσεων μετρήσεων μικροσυστοιχειών

Η ανάπτυξη του μοντέλου που υπαινίχθηκε στην προηγούμενη παράγραφο, βασίζεται στη φυσική βάση των μετρήσεων των δεδομένων μικροσυστοιχειών δηλαδή στο φαινόμενο του υβριδισμού. Παρά την φαινομενική πολυπλοκότητα τους, η ουσία των πειραμάτων μικροσυστοιχειών βασίζεται στην υβριδοποίηση μόνιων mRNA δεδομένου γονιδίου πάνω σε μια επιφάνεια η οποία φέρει πολλαπλούς στόχους. Πρώτα θεωρούμε την περίπτωση όπου όλοι αυτοί οι στόχοι αντιστοιχούν στον ίδιο ιχνηθέτη. Κάθε *δυνατή* υβριδοποίηση (δηλαδή πρόσδεση ενός μορίου mRNA σε ένα στόχο) οδηγεί σε ένα πιθανό σήμα για το οποίο μπορούμε να έχουμε μια πεποίθηση (δηλαδή μια πιθανότητα). Επειδή το μόριο mRNA μπορεί να υβριδοποιηθεί σε μία μόνο θέση (δηλαδή σε ένα ιχνηθέτη), η πεποίθησή μας για το σήμα του ιχνηθέτη μπορεί να εκφραστεί σαν το μέσο όρο του άθροισματος των μεμονωμένων πεποιθήσεων για κάθε στόχο χωριστά. Με άλλα λόγια οι πεποιθήσεις για τα σήματα αυτά αποτελούν αμοιβαία αποκλειόμενα ενδεχόμενα. Στην περίπτωση που οι μετρήσεις από τον ίδιο ιχνηθέτη είναι ανταλλάξιμες για όλα τις κηλίδες που τον περιέχουν, η πεποίθηση για το σήμα του ιχνηθέτη θα δίνεται από την *a posteriori* κατανομή του μοντέλου πεποίθησης που χρησιμοποιήθηκε λαμβάνοντας υπόψη όλες τις μετρήσεις. Έτσι για γονίδια που αντιπροσωπεύονται από ένα και μόνο ιχνηθέτη, η πιθανότητα υποέκφρασης θα δίνεται από την τιμή της κατανομής Behrens Fisher.

Η γενική περίπτωση, έχει να κάνει με την κατάσταση εκείνη στην οποία πάνω στη συστοιχεία υπάρχουν στόχοι που αντιστοιχούν σε περισσότερους του ενός ιχνηθέτες (cDNA probes). Στην περίπτωση αυτή η πεποίθηση για το φαινόμενο της υποέκφρασης του *κάθε ιχνηθέτη χωριστά* θα δίνεται από την τιμή της κατανομής Behrens Fisher για τις αντίστοιχες μετρήσεις. Επειδή όμως ένα μόριο mRNA μπορεί να υβριδοποιηθεί σε στόχο που ανήκει σε ένα και μόνο ιχνηθέτη, οι πεποιθήσεις της

διαφοράς των διάφορετικών ιχνηθετών αποτελούν αμοιβαία αποκλειόμενα ενδεχόμενα. Στην περίπτωση αυτή η πεποιθήση για το φαινόμενο της υποέκφρασης εκφράζεται με τη μορφή μείγματος κατανομών Behrens Fisher (mixture model), κάθε μία από τις οποίες αναφέρεται στα δεδομένα ενός και μόνο ιχνηθέτη. Οι συντελεστές μείξης (mixture coefficients) εύκολα υπολογίζονται με την επίκληση του επιχειρήματος της ισοπίθανης υβριδοποίησης σε κάθε στόχο· οι συντελεστές αυτοί είναι ίσοι με το αριθμό των κηλίδων στις οποίες έχει τυπωθεί ένας ιχνηθέτης ως προς τον αριθμό όλων των ιχνηθετών ενός δεδομένου γονιδίου στην επιφάνεια της μικροσυστοιχείας. Οι πληθάρθρωτοι αυτοί είναι διαθέσιμοι από τα αποτελέσματα της αναζήτησης ομολογίας κατά BLAST που διενεργήθηκε στα αρχικά στάδια της ανάλυσης των δεδομένων μικροσυστοιχειών.

Επιλογή κατωφλίου πιθανότητας υποέκφρασης μέσω ανάλυσης καμπυλών ROC

Όπως είδαμε σε προηγούμενο τμήμα, μια στρατηγική πειραματικού σχεδιασμού που δεν δημιουργεί λογικά παράδοξα, προϋποθέτει τον ορισμό ενός **κατωφλίου πιθανότητας υποέκφρασης** C_p και ενός αντίστοιχου διχοτομικού κανόνα. Βάσει του τελευταίου, γονίδια με πιθανότητα υποέκφρασης μικρότερης του C_p **δεν** συμπεριλαμβάνονται στα επόμενα βήματα της ανάλυσης. Αν και η επιλογή του C_p αφήνεται στη διακριτικότητα του κάθε ερευνητή, η προδιαγραφή μιας μεθοδολογίας που θα μπορούσε να χρησιμεύσει για την επιλογή της τιμής C_p είναι εφικτή και αναλύεται σε αυτό το τμήμα.

Για να περιγράψουμε τη μεθοδολογία αυτή θα πρέπει να εξετάσουμε τις πρακτικές συνέπειες που έχει ο καθορισμός της τιμής της C_p στις πεποιθήσεις για τις μετρήσεις μικροσυστοιχειών. Σε κάθε τέτοιο πείραμα θα έχουμε έναν αριθμό γονιδίων που πραγματικά υποεκφράζονται και έναν αριθμό γονιδίων που δεν υποεκφράζονται. Η επιλογή της τιμής C_p έχει τα εξής αποτελέσματα:

1. Ένα ποσοστό αληθώς μη υποεκφρασμένων γονιδίων με τιμή πιθανότητας υποέκφρασης $> C_p$ συμπεριλαμβάνονται στις περαιτέρω αναλύσεις (ψευδώς θετικά)
2. Ένα ποσοστό αληθώς υποεκφρασμένων γονιδίων που στο συγκεκριμένο πείραμα όμως είχαν πιθανότητα υποέκφρασης μικρότερης της C_p εξαιρούνται από την ανάλυση (ψευδώς αρνητικά).

Είναι προφανές ότι βέλτιστη τιμή (ή το σύνολο τιμών) της C_p είναι εκείνη που οδηγεί στον μικρότερο αριθμό ψευδώς θετικών **και** ψευδώς αρνητικών γονιδίων. Με αυτό τον τρόπο το πρόβλημα επιλογής της τιμής C ανάγεται σε ένα πρόβλημα ανάλυσης καμπυλών ROC (Receiver Operating Characteristic curves).

Η μεθοδολογία ανάλυσης των καμπυλών ROC γνωστή από την κλινική επιδημιολογία, αναφέρεται στον προσδιορισμό του βέλτιστου κατωφλίου μια διαγνωστικής δοκιμασίας. Ο προσδιορισμός αυτός γίνεται με την εξέταση της καμπύλης, που ορίζεται από τα σημεία του επιπέδου (x,y) με x = ποσοστό ψευδώς θετικών, y = ποσοστό αληθώς θετικών για διαφορετικές τιμές του κατωφλίου. Μια καμπύλη ROC αναπαράστα γραφικά την αντίστροφη σχέση μεταξύ ευαισθησίας και ειδικότητας και μπορεί επιπλέον να χρησιμοποιηθεί για την επιλογή του βέλτιστου κατωφλίου.

Μεθοδολογία και υλοποίηση ανάλυσης καμπυλών ROC

Όπως αναφέρθηκε προηγούμενα η εφαρμογή της μεθοδολογίας ανάλυσης καμπυλών ROC προϋποθέτει την διαθεσιμότητα δύο ομάδων γονιδίων:

- ο Μη αληθώς υποεκφρασμένα (χρησιμοποιούνται για τον υπολογισμό των ψευδώς θετικών)

- ο Αληθώς υποεκφρασμένων (χρησιμοποιούνται για τον υπολογισμό των ψευδώς αρνητικών).

Στην περίπτωση που αναλύουμε χρησιμοποιήσαμε τις μετρήσεις του κοινού βιολογικού δείγματος αναφοράς για να κατασκευάσουμε το δείγμα των αληθώς μη υποεκφρασμένων γονιδίων. Οι υβριδοποιήσεις του κοινού δείγματος αναφοράς, παρέχουν ένα παράθυρο στη μεταβλητότητα των μετρήσεων μικροσυστοιχειών και επιτρέπουν τη θεώρηση της ποσοτικής σχέσης των αντίστοιχων πεποιθήσεων με τη μεταβλητότητα αυτή. Έτσι στην περίπτωση που συγκρίναμε τις πεποιθήσεις για τις διαφορές των εντάσεων δύο δειγμάτων με v_1 και v_2 πειράματα μικροσυστοιχειών από το κοινό βιολογικό δείγμα αναφοράς, με κάθε δυνατή τιμή του κατωφλίου πιθανότητας υποέκφρασης C_p θα μπορούσαμε να κατασκευάσουμε την εξάρτηση του αριθμού των αληθώς αρνητικών αποφάσεων σε σχέση με το C_p . Η ειδικότητα είναι η αναμενόμενη τιμή του αριθμού αυτού, υπολογιζόμενη σε ένα πεπερασμένο αριθμό τέτοιων συγκρίσεων. Η κατασκευή του δείγματος για τον υπολογισμό της ειδικότητας έγινε με τον ακόλουθο τρόπο:

1. Απ' όλα τα πειράματα μικροσυστοιχειών του βιολογικού δείγματος αναφοράς (κανάλι 1) επιλέχθηκαν (με αντικατάσταση) $v_1 = 3$ δείγματα²⁰.
2. Απ' όλα τα πειράματα μικροσυστοιχειών του βιολογικού δείγματος αναφοράς επιλέχθηκαν (με αντικατάσταση) $v_2 = 3$ δείγματα.
3. Τα δύο δείγματα αναλύθηκαν με τον τρόπο που περιγράφηκε προηγούμενα, από το επίπεδο της επεξεργασίας εικόνας μέχρι την κανονικοποίηση
4. Οι μετρήσεις των ιχνηθετών από τα δύο δείγματα χρησιμοποιήθηκαν για τον υπολογισμό των αντίστοιχων πεποιθήσεων με τη βοήθεια της κατανομής Behrens-Fisher
5. Το ποσοστό των πεποιθήσεων με τιμή μικρότερη του κατωφλίου C_p καταγράφηκε για τιμές του C_p από 0.8 μέχρι 1.0. Η επιλογή των τιμών αυτών έγινε για να κρατηθεί ο απόλυτος αριθμός των ψευδώς θετικών σε ένα λογικό νούμερο (π.χ. μια ανάλυση μικροσυστοιχείας με 10000 ιχνηθέτες, από τους οποίους μόνο 5% είναι αληθώς υποεκφρασμένοι με ειδικότητα 80% περιλαμβάνει γύρω στα 1900 ψευδώς υποεκφρασμένα γονίδια)
6. Η διαδικασία 1-5 επαναλήφθηκε 10000 φορές
7. Μετά το τέλος του πειράματος δειγματοληψίας, η ειδικότητα για τις διάφορες τιμές της C_p εκτιμήθηκε από το δειγματικό μέσο όρο του ποσοστού των αληθώς αρνητικών πεποιθήσεων (ποσοστό που μετρήθηκε στο βήμα 5). Ένας δείκτης της σταθερότητας (robustness) της εκτίμησης αυτής δίνεται από την 5^η και 95^η εκατοστιαία θέση (quantile) των δειγματικών αυτών ποσοστών.

Κατά αντιστοιχία ο υπολογισμός της ευαισθησίας έγινε από τα δεδομένα του δεύτερου καναλιού των πειραμάτων των μικροσυστοιχειών με τον ακόλουθο τρόπο:

1. Απ' όλα τα πειράματα μικροσυστοιχειών του καναλιού 2 επιλέχθηκαν (με αντικατάσταση) $v_1 = 3$ δείγματα.
2. Απ' όλα τα πειράματα μικροσυστοιχειών του του καναλιού 2 επιλέχθηκαν (με αντικατάσταση) $v_2 = 3$ δείγματα.
3. Τα δύο δείγματα αναλύθηκαν με τον τρόπο που περιγράφηκε προηγούμενα, από το επίπεδο της επεξεργασίας εικόνας μέχρι την κανονικοποίηση
4. Οι μετρήσεις των ιχνηθετών από τα δύο δείγματα χρησιμοποιήθηκαν για τον υπολογισμό των αντίστοιχων πεποιθήσεων με τη βοήθεια της κατανομής Behrens-Fisher

²⁰ Η επιλογή του αριθμού 3 έγινε βάσει του αριθμού των πειραμάτων μικροσυστοιχειών που αναφέρονταν στα παρθενικά και ενεργοποιημένα λεμφοκύτταρα (3 πειράματα ανά κατάσταση)

5. Το ποσοστό των πεποιθήσεων με τιμή μικρότερη του κατωφλίου C_p καταγράφηκε για τιμές του C_p από 0.8 μέχρι 1.0 . Η επιλογή των τιμών αυτών έγινε για να κρατηθεί ο *απόλυτος* αριθμός των ψευδώς θετικών σε ένα λογικό νούμερο (π.χ. μια ανάλυση μικροσυστοιχείας με 10000 ιχνηθέτες, από τους οποίους μόνο 5% είναι αληθώς υποεκφρασμένοι με ειδικότητα 80% περιλαμβάνει γύρω στα 1900 ψευδώς υποεκφρασμένα γονίδια)
6. Η διαδικασία 1-5 επαναλήφθηκε 10000 φορές
7. Μετά το τέλος του πειράματος δειγματοληψίας, η ευαισθησία για τις διάφορες τιμές της C_p εκτιμήθηκε από το δειγματικό μέσο όρο του ποσοστού των αληθώς θετικών πεποιθήσεων. Λόγω της φύσης των δειγματικών αυτών συγκρίσεων, το ποσοστό αυτό ΔΕΝ είναι ίσο με το ποσοστό των θετικών συγκρίσεων που μετρήθηκε στο βήμα 5 γιατί αυτό περιλαμβάνει τόσο αληθώς όσο και ψευδώς θετικά αποτελέσματα προερχόμενα από τις δύο ομάδες (μη υποεκφρασμένων και υποεκφρασμένων) γονιδίων. Το δειγματικό ποσοστό των αληθώς θετικών μετρήσεων για κάθε τιμή του κατωφλίου υποέκφρασης C_p μπορεί να *εκτιμηθεί* από την ακόλουθη εξίσωση²¹:

$$\frac{x_i - (1 - Sp_C)(1 - f)}{f}$$

όπου x_i το μετρηθέν ποσοστό θετικών πεποιθήσεων, Sp_C η ειδικότητα για την τιμή κατωφλίου C και f το ποσοστό των αληθώς υποεκφρασμένων μετρήσεων. Επειδή η τιμή f δεν είναι γνωστή, η εξίσωση αυτή λύνεται για ένα εύρος υποθετικών τιμών της (π.χ. από 1-10% , ποσοστά που αναμένονται για μαζικά πειράματα μικροσυστοιχειών) και τα αποτελέσματα της ευαισθησίας δίνονται σαν συνάρτηση της τιμής της f που χρησιμοποιήθηκε.

Με βάση τις τιμές της ευαισθησίας και της ειδικότητας όπως αυτές μετρήθηκαν/εκτιμήθηκαν παραπάνω κατασκευάστηκαν οι αντίστοιχες καμπύλες ROC (σαν συνάρτηση της f) με δύο διαφορετικές μη παραμετρικές μεθόδους: α) προσαρμογή με πολυνύμα Hermite και β) Bezier Splines. Συγκεκριμένα οι μετρήσεις αυτές τροφοδοτήθηκαν στις συνάρτησεις **PolynomialFit** και **SplineFit** του *Mathematica* που υλοποιούν τις αντίστοιχες αριθμητικές μεθόδους. Οι καμπύλες ROC που κατασκευάστηκαν με αυτό τον τρόπο παραγωγίστηκαν συμβολικά εντός του *Mathematica* · η τιμή της παραγώγου της καμπύλης ROC εκτιμήθηκε τόσο γραφικά, όσο και με τις αριθμητικές μεθόδους βελτιστοποίησης του *Mathematica* για το εύρος τιμών της C_p που χρησιμοποιήθηκε.

Σχεδιασμός και υλοποίηση του λογισμικού ανάλυσης των δεδομένων μικροσυστοιχειών

Για τους σκοπούς της εργασίας αυτής, απαιτήθηκε η συγγραφή λογισμικού που υλοποιούσε τις προαναφερθέντες αναζητήσεις στις βάσεις δεδομένων και υπολογισμό πιθανοτήτων. Η γλώσσα προγραμματισμού που χρησιμοποιήθηκε ήταν η PERL λόγω της διαθεσιμότητας βιβλιοθηκών για την εκτέλεση και ανάλυση αποτελεσμάτων αναζητήσεων (BIOPERL) σε βάσεις δεδομένων βασισμένων στην GENBANK. Επιπλέον πλεονεκτήματα της PERL είναι η ευκολία συγγραφής λογισμικού ανάλυσης και επεξεργασίας (parsing) μεγάλων αρχείων κειμένου καθώς και η διασυνδεσιμότητα της με τις αριθμητικές βιβλιοθήκες που υπολογίζουν την κατανομή

²¹ Η εξίσωση αυτή προκύπτει από τη θεώρηση του μοντέλου μείζεως τυχαίων μεταβλητών $x_i = (1 - Sp_C^i) \cdot (1 - f) + f \cdot Se_C^i$ με την αντικατάσταση της ειδικότητας από την αναμενόμενη τιμή της, και την επίλυση της ως προς την άγνωστη ποσότητα Se_C^i

Behrens Fisher όπως περιγράφηκε στο αντίστοιχο τμήμα. Τα αποτελέσματα των επιμέρους αναλύσεων αποθηκεύτηκαν τοπικά και εισήχθησαν στο *Mathematica* για τη δημιουργία των αντίστοιχων γραφημάτων.

5.2 Αποτελέσματα αναλύσεων

Επιλογή κατωφλίου υποεκφρασιμότητας

Με βάση τα δεδομένα του πειράματος δειγματοληψίας του βιολογικού δείγματος αναφοράς, εκτιμήσαμε την ειδικότητα για διαφορετικές τιμές του κατωφλίου υποεκφρασιμότητας. Τα αποτελέσματα δίνονται στην Εικόνα 23.

Εικόνα 23 Η ειδικότητα σαν συνάρτηση του κατωφλίου πιθανότητας υποέκφρασης. Δεδομένα από 10000 συγκρίσεις δειγμάτων πειραμάτων μικροσυστοιχείων με 3 πειράματα σε κάθε δείγμα. Στον άξονα X δίνεται το κατώφλι πιθανότητας υποέκφρασης, ενώ στον άξονα Y η αντίστοιχη ειδικότητα για την κατανομή Behrens Fisher. Η κόκκινη γραμμή δίνει την ειδικότητα υπολογιζόμενη σαν το δειγματικό μέσο όρο του % των πεποιθήσεων για τις διαφορές εντάσεων των ιχνηθετών με τιμή μικρότερη από το κατώφλι υποεκφρασιμότητας. Οι γραμμές σφάλματος δίνουν το εμπειρικό διάστημα εμπιστοσύνης (5%,95%) για το ποσοστό των πεποιθήσεων μικρότερων του αντίστοιχου κατωφλίου. Έτσι για το κατώφλι υποεκφρασιμότητας 0.8, το 90% των δειγμάτων εμφάνισαν ποσοστά πεποιθήσεων μικρότερων του 0.8 στο διάστημα (0.51,0.82).

Παρατηρούμε ότι καθώς το κατώφλι υποεκφρασιμότητας γίνεται περισσότερο αυστηρό η ειδικότητα του κατωφλίου αυξάνεται και η μεταβλητότητα της ειδικότητας γίνεται όλο και πιο μικρή. Για τιμές κατωφλίου μικρότερες του 0.88, η ειδικότητα είναι μικρότερη του κατωφλίου υποεκφρασιμότητας, ενώ στο διάστημα 0.88 – 0.91 οι δύο αυτό ποσότητες είναι αριθμητικά ισοδύναμες. Για τιμές του κατωφλίου μεγαλύτερες του 0.91, η ειδικότητα είναι μεγαλύτερη του κατωφλίου υποεκφρασιμότητας. Τα αποτελέσματα αυτά υποδεικνύουν ότι με κριτήριο τον «πραγματικό κόσμο», η κατανομή Behrens Fisher είναι μη συντηρητική στις πεποιθήσεις που εκφράζονται μέσω αυτής όταν οι τιμές των τελευταίων είναι μικρότερες του 0.88, ενώ είναι «συντηρητική» όταν οι αντίστοιχες πεποιθήσεις είναι

μεγαλύτερες του 0.91. Στο διάστημα 0.88-0.91 η εφαρμογή της κατανομής Behrens Fisher οδηγεί σε αποτελέσματα που τείνουν να συμφωνούν με τον «πραγματικό κόσμο». Η σημαντικότητα της παρατήρησης αυτής έγκειται στην πολύ ισχυρή παραμετρική υπόθεση, που προϋποθέτει η κατανομή Behrens Fisher, αυτήν της κανονικότητας των μετρήσεων. Παρά το γεγονός ότι οι μηχανισμοί δημιουργίας των σημάτων μικροσυστοιχιών δεν μπορεί να είναι κανονικοί (π.χ. προέρχονται από μετρητικά όργανα με πεπερασμένη κλίμακα), η παραμετρική υπόθεση της κανονικότητας δεν φαίνεται να δημιουργεί προβλήματα όταν το αντίστοιχο κατώφλι υποεκφρασιμότητας είναι αρκούντως υψηλό. Η γραφική ανάλυση του παραπάνω σχήματος υποδεικνύει ότι η βέλτιστη τιμή του κατωφλίου εντοπίζεται στο διάστημα 0.88-0.91, αλλά για τη εκτίμηση της τιμής αυτής απαιτείται η ανάλυση καμπυλών ROC για διάφορες τιμές του ποσοστού των υποεκφρασμένων γονιδίων.

Λόγω της μαζικής φύσης του πειράματος, κάναμε την υπόθεση ότι ένα μικρό ποσοστό των γονιδίων θα επιδείξει διαφορετική υποέκφραση και εστίασαμε την προσοχή μας στο διάστημα 1-10%. Οι διαφορετικές καμπύλες ROC που παράγονται αυτό τον τρόπο αναπαράστώνται γραφικά στα ακόλουθα 3 σχήματα (μη παραμετρικός εκτιμητής με καμπύλες Bezier, μη παραμετρικός εκτιμητής με πολώνυμα 5^{ου} βαθμού) με χρωματικό κώδικα ανάλογα με την τιμή του f που χρησιμοποιήθηκε για την παραγωγή τους.

Εικόνα 24 Καμπύλη ROC για διαφορετικές τιμές του ποσοστού υποεκφρασμένων γονιδίων στο διάστημα 1-10%. Οι καμπύλες ROC που παράγονται για διαφορετικές τιμές του ποσοστού των υποεκφρασμένων γονιδίων δίνονται με το ίδιο χρώμα με την τιμή που χρησιμοποιήθηκε για την κατασκευή τους. Οι ROC καμπύλες κατασκευάστηκαν με τη βοήθεια καμπυλών Bezier. Οι κουκίδες αντιστοιχούν στα ζεύγη τιμών (1-Ειδικότητα, Ευαισθησία) που χρησιμοποιήθηκαν από το μη παραμετρικό εκτιμητή

Πέρα από τη διαφορετική ομαλότητα των εκτιμητών οι καμπύλες που παράγονται δεν έχουν ουσιαστικές διαφορές μεταξύ τους. Ο ρυθμός μεταβολής των καμπυλών ROC δίνει μια εκτίμηση της αντίστροφης σχέσης μεταξύ ευαισθησίας και ειδικότητας, και μπορεί να ερμηνευτεί σαν το κέρδος σε μονάδες ευαισθησίας για μεταβολή της ειδικότητας κατά 1%. Παρατηρούμε ότι για χαμηλές τιμές ειδικότητας, οι καμπύλες φαίνεται να μεταβάλλονται με χαμηλότερο ρυθμό · έτσι ενώ αρχικά το κέρδος σε

μονάδες ευαισθησίας είναι μεγαλύτερο της μονάδας πέρα από κάποιο σημείο μειώνεται και ασυμπτωτικά φαίνεται να τείνει στο μηδέν.

Εικόνα 25 Καμπύλη ROC για διαφορετικές τιμές του ποσοστού υποεκφρασμένων γονιδίων στο διάστημα 1-10%. Οι καμπύλες ROC που παράγονται για διαφορετικές τιμές του ποσοστού των υποεκφρασμένων γονιδίων δίνονται με το ίδιο χρώμα με την τιμή που χρησιμοποιήθηκε για την κατασκευή τους. Οι ROC καμπύλες κατασκευάστηκαν με τη βοήθεια πολωνύμων 5^{ου} βαθμού. Οι κουκίδες αντιστοιχούν στα ζεύγη τιμών (1-Ειδικότητα, Ευαισθησίας) που χρησιμοποιήθηκαν από το μη παραμετρικό εκτιμητή

Η ανάλυση του ρυθμού μεταβολής της καμπύλης ROC δίνεται γραφικά στα επόμενα δύο σχήματα για τους δύο εκτιμητές (Bezier και πολυωνυμικό). Όπως φαίνεται από τα σχήματα αυτά το κέρδος σε ευαισθησία για κάθε μονάδα μεταβολής ειδικότητας αρχίζει από μια υψηλή τιμή μεγαλύτερη του ενός (π.χ. γύρω στο 4.8 για $f = 0.01$) και πέφτει με σχεδόν εκθετικό ρυθμό στην τιμή 1 για τιμές του κατωφλίου υποέκφρασης κοντά στο 0.9 (π.χ. για $f = 0.01$ η τιμή του κατωφλίου έκφρασης όπου αυτό συμβαίνει είναι 0.895 για τον εκτιμητή Bezier και 0.908 για τον πολυωνυμικό εκτιμητή). Οι καμπύλες του ρυθμού μεταβολής των καμπυλών ROC μπορούν να βοηθήσουν στην επιλογή της τιμής του κατωφλίου μια και μας δείχνουν απο ποιο σημείο και μετά η υιοθέτηση λιγότερο αυστηρών κατωφλίων είναι μη αποδοτική, με το κέρδος σε ευαισθησία να είναι σαφώς μικρότερο από τις απώλειες σε ειδικότητα. Επειδή ο ρυθμός μεταβολής της καμπύλης ROC είναι ίσος με το λόγο πιθανοφάνειας (likelihood ratio) της διαγνωστικής δοκιμασίας, ένας εξαιρετικά απλός τρόπος για τον καθορισμό του κατωφλίου υποεκφρασιμότητας, θα χρησιμοποιούσε την τιμή ειδικότητας για την οποία ο λόγος πιθανοφάνειας είναι περίπου ίσος με 1. Στην περίπτωση αυτή η πιθανότητα υποέκφρασης για ένα γονίδιο META την εφαρμογή της κατανομής Behrens Fisher είναι ίση με την πιθανότητα υποέκφρασης (prior πιθανότητα) πριν την εφαρμογή της κατανομής (δηλαδή προτού διεξαχθεί το

πείραμα).

Εικόνα 26 Ρυθμός μεταβολής καμπύλης ROC για διαφορετικές τιμές της πιθανότητας υποέκφρασης (χρωματικός κώδικας). Οι καμπύλες αντιστοιχούν στον εκτιμητή Bezier

Εικόνα 27 Ρυθμός μεταβολής καμπύλης ROC για διαφορετικές τιμές της πιθανότητας υποέκφρασης (χρωματικός κώδικας). Οι καμπύλες αντιστοιχούν στον πολυωνυμικό εκτιμητή

Τα αποτελέσματα της ανάλυσης ROC επιβεβαιώνουν την υπόθεση που προτάθηκε στην αρχή της ενότητας αυτή ότι δηλαδή, η «βέλτιστη» τιμή κατωφλίου υποέκφρασης βρίσκεται στο διάστημα 0.88-0.91 . Για τους σκοπούς της παρούσης

εργασίας υιοθετούμε το σαν κατώφλι υποεκφρασιμότητας την τιμή εκείνη που αντιστοιχεί σε ειδικότητα 0.895. Από εδώ και στο εξής οι αναλύσεις θα περιοριστούν σε γονίδια των οποίων οι πεποιθήσεις υποέκφρασης είναι μεγαλύτερες από την τιμή 0.89 που αντιστοιχεί στην τιμή αυτή ειδικότητας.

Αποτελέσματα πειραμάτων διαφορικής έκφρασης

Στο τμήμα αυτό θα αναλύσουμε τα αποτελέσματα των πειραμάτων υποεκφρασιμότητας για το σύνολο δεδομένων του Alizadeh. Όπως αναφέρθηκε στο κεφάλαιο που εισήγαγε το βιολογικό ερώτημα, οι συγκρίσεις που έχουν νόημα στα πλαίσια του υπό εξέταση συστήματος είναι οι ακόλουθες :

1. παρθενικά T με ενεργοποιημένα T
2. παρθενικά T με παρθενικά B
3. παρθενικά T με ενεργοποιημένα B

Τα αποτελέσματα των συγκρίσεων δίνονται με τη μορφή πινάκων όπου αναγράφουν το όνομα του γονιδίου, τον μοναδιαίο κωδικό του στην υποβάση αλληλουχιών αναφοράς της GENBANK, την αντίστοιχη πιθανότητα υποέκφρασης και τη σειρά της πιθανότητας (rank) για όλα τα γονίδια για τα οποία είχαμε δεδομένα. Επιπλέον δίνουμε τα αποτελέσματα της 3-πλής σύγκρισης, που αντιστοιχεί στο ερώτημα ποιοι είναι οι παράγοντες εκείνοι που υποεκφράζονται στη σύγκριση 1 **ΚΑΙ** υποεκφράζονται στη σύγκριση 2 **ΚΑΙ** υποεκφράζονται στη σύγκριση 3. Λόγω του βασικού ερωτήματος μόνο τα αποτελέσματα για μεταγραφικούς παράγοντες και πρωτεΐνες που σχετίζονται με τη διαδικασία της μεταγραφής παραθέτονται εδώ.

Η σύγκριση ένα έχει παρουσιασθεί σε προηγούμενη εργασία σε σχέση με την πρώτη αυτή απόπειρα εφαρμογής Bayesian μεθόδων για τη διεξαγωγή στατιστικών πειραμάτων διαφορικής έκφρασης, τα αποτελέσματα της παρούσης περιέχουν μεγαλύτερο αριθμό μεταγραφικών παραγόντων λόγω της απόφασης να διενεργηθεί το βήμα αναζήτησης ομολογίας κατά BLAST στην υποβάση της GENBANK αλληλουχιών αναφοράς. Με αυτό τον τρόπο αρκετά ESTs που δεν είχαν αντιστοιχηθεί σε γνωστά γονίδια ταυτοποιήθηκαν σαν ιχνηθέτες γνωστών πλέον γονιδίων. Επιπλέον κάποιες από τις αριθμητικές τιμές των πιθανοτήτων υποέκφρασης είναι διαφορετικές λόγω της κατασκευής των ανταλλάξιμων δειγμάτων με τον τρόπο που αναλύθηκε. Εξαιτίας της ύπαρξης πολλαπλών ιχνηθετών για κάποια από τα γονίδια, ορισμένοι από τους μεταγραφικούς παράγοντες που βάσει ενός ιχνηθέτη είχαν προταθεί ότι υποεκφράζονται, δεν εμφανίζονται πλέον στον κατάλογο των υποεκφρασμένων γονιδίων με πιθανότητα > 0.89 μια και η πιθανότητα υποέκφρασης βάσει ΟΛΩΝ των διαθέσιμων μετρήσεων ήταν μικρότερη του κατωφλίου. Αυτή είναι η περίπτωση του παράγοντα NFAT4 που στην προηγούμενη σύγκριση είχε δωθεί με πιθανότητα > 0.95 ενώ στην σύγκριση με την παρούσα μεθοδολογία έχει πιθανότητα 0.88 περίπου.

Στον Πίνακα 2 παρουσιάζονται τα αποτελέσματα για τη σύγκριση 1, ο πίνακας 3 δίνει τα αποτελέσματα της σύγκρισης 2 ενώ ο πίνακας 4 τα αποτελέσματα της σύγκρισης 3. Με την εφαρμογή της στρατηγικής του αφελούς Bayes παίρνουμε τα αποτελέσματα του πίνακα 5. Η εφαρμογή του προσέγγισης του αφελούς Bayes μειώνει δραστικά τον αριθμό των υποψηφίων από 30 (σύγκριση 1), 54 (σύγκριση 2), 79 (σύγκριση 3) σε 11 υποψηφίους.

Κωδικός γονιδίου	Πιθανότητα υποέκφρασης	Σειρά (rank) πιθανότητας	Όνομα γονιδίου
NM_181782	0.999396850090775	2	nuclear receptor coactivator 7 (NCOA7)
NM_005900	0.994935375966449	9	SMAD mothers against DPP homolog 1 (Drosophila) (SMAD1)
NM_003150	0.982942904361867	25	signal transducer and activator of transcription 3 (STAT3) transcript variant 2
NM_005195	0.980770381846295	29	CCAAT/enhancer binding protein (C/EBP) delta (CEBPD)
NM_018703	0.980303251835275	31	retinoblastoma binding protein 6 (RBBP6) transcript variant 2
NM_005517	0.979301271224567	32	high-mobility group nucleosomal binding domain 2 (HMGN2)
NM_005241	0.972992841246581	45	ecotropic viral integration site 1 (EVI1)
NM_138931	0.970530696128012	51	B-cell CLL/lymphoma 6 (zinc finger protein 51) (BCL6) transcript variant 2
NM_006197	0.968490330959933	55	pericentriolar material 1 (PCM1)
NM_006084	0.967553299705014	57	interferon-stimulated transcription factor 3 gamma 48kDa (ISGF3G)
NM_198392	0.962476198430888	68	transcription factor 21 (TCF21) transcript variant 1
NM_002357	0.942922144780255	115	MAX dimerization protein 1 (MAD)
NM_005243	0.940313176359363	121	Ewing sarcoma breakpoint region 1 (EWSR1) transcript variant EWS
NM_199420	0.932877399759571	141	polymerase (DNA directed) theta (POLQ) transcript variant 2
NM_032226	0.929380560770545	147	zinc finger CCHC domain containing 7 (ZCCHC7)
NM_145113	0.927063100645915	153	MAX protein (MAX) transcript variant 3
NM_021145	0.926579763247061	156	cyclin D binding myb-like transcription factor 1 (DMTF1)
NM_007358	0.925274359125629	161	likely ortholog of mouse metal response element binding transcription factor 2 (M96)
NM_001527	0.921399037075494	168	histone deacetylase 2 (HDAC2)
NM_005319	0.911516393742110	194	histone 1 H1c (HIST1H1C)
NM_005901	0.902090477200261	218	SMAD mothers against DPP homolog 2 (Drosophila) (SMAD2)
NM_002166	0.899923960258078	223	inhibitor of DNA binding 2 dominant negative helix-loop-helix protein (ID2)
NM_005239	0.898211726719624	231	v-ets erythroblastosis virus E26 oncogene homolog 2 (avian) (ETS2)
NM_181458	0.896685640464066	234	paired box gene 3 (Waardenburg syndrome 1) (PAX3) transcript variant PAX3D
NM_018660	0.895396912572073	240	zinc finger protein 395 (ZNF395)
NM_138559	0.893562636902930	247	B-cell CLL/lymphoma 11A (zinc finger protein) (BCL11A) transcript variant 3
NM_175931	0.893361740037556	249	core-binding factor runt domain; translocated to 3 (CBFA2T3) transcript variant 2
NM_000176	0.892978492482659	250	nuclear receptor subfamily 3 group C member 1 (glucocorticoid receptor) (NR3C1)
NM_002895	0.892758076294564	251	retinoblastoma-like 1 (p107) (RBL1) transcript variant 1
NM_015565	0.892498838866435	254	zinc finger protein 294 (ZNF294)

Πίνακας 3 Υποεκφρασμένοι μεταγραφικοί παράγοντες ή πρωτεΐνες σχετιζόμενες με τη μεταγραφή στα ενεργοποιημένα T λεμφοκύτταρα σε σχέση με τα παρθενικά κύτταρα.

Κωδικός γονιδίου	Πιθανότητα υποέκφρασης	Σειρά (rank) πιθανότητας	Όνομα γονιδίου
NM_024503	0.999995880816794	4	human immunodeficiency virus type I enhancer binding protein 3 (HIVEP3)
NM_181782	0.999916251734985	10	nuclear receptor coactivator 7 (NCOA7)
NM_032682	0.999712930051747	12	forkhead box P1 (FOXP1)
NM_003150	0.997593176472001	24	signal transducer and activator of transcription 3 (STAT3) transcript variant 2
NM_002479	0.99744569528209	26	myogenin (myogenic factor 4) (MYOG)
NM_005657	0.996965530638257	30	tumor protein p53 binding protein 1 (TP53BP1)
NM_018675	0.996953506208653	31	zinc finger protein 302 (ZNF302)
NM_005517	0.979301271224567	32	Homo sapiens high-mobility group nucleosomal binding domain 2 (HMG2)
NM_022463	0.995079734364351	37	nucleoredoxin (NXN)
NM_153812	0.991512328587826	57	PHD finger protein 13 (PHF13)
NM_014468	0.991030042778804	62	VENT-like homeobox 2 (VENTX2)
NM_014724	0.99083459214555	65	zinc finger protein 305 (ZNF305)
NM_022466	0.990819265615766	66	zinc finger protein subfamily 1A 5 (ZNFN1A5)
NM_138931	0.990330722946428	70	B-cell CLL/lymphoma 6 (zinc finger protein 51) (BCL6) transcript variant 2
NM_148169	0.989157973273433	80	F-box protein 17 (FBXO17) transcript variant 1
NM_052996	0.98868544308385	87	PR domain containing 7 (PRDM7)
NM_032679	0.987961912220017	90	zinc finger protein 577 (ZNF577)
NM_005378	0.985455800131088	98	v-myc myelocytomatosis viral related oncogene neuroblastoma derived (avian) (MYCN)
NM_001178	0.979154479643884	141	aryl hydrocarbon receptor nuclear translocator-like (ARNTL)
NM_021191	0.978547431806487	146	neurogenic differentiation 4 (NEUROD4)
NM_005131	0.973257039456696	173	THO complex 1 (THOC1)
NM_182918	0.968284608502042	198	v-ets erythroblastosis virus E26 oncogene like (avian) (ERG) transcript variant 1
NM_004830	0.966604733630767	208	cofactor required for Sp1 transcriptional activation subunit 3 130kDa (CRSP3)
NM_004405	0.965728880180481	214	distal-less homeo box 2 (DLX2)
NM_005967	0.964782587127204	218	NGFI-A binding protein 2 (EGR1 binding protein 2) (NAB2)
NM_018703	0.962135763712904	233	retinoblastoma binding protein 6 (RBBP6) transcript variant 2
NM_017410	0.960451574565443	240	homeo box C13 (HOXC13)
NM_018555	0.94835916832576	286	zinc finger protein 331 (ZNF331)
NM_014210	0.947604848704116	293	ecotropic viral integration site 2A (EVI2A)
NM_005962	0.942663516592186	314	MAX interactor 1 (MXI1) transcript variant 1

NM_006442	0.94122119073583	323	DR1-associated protein 1 (negative cofactor 2 alpha) (DRAP1)
NM_198893	0.938736417343937	336	zinc finger protein 160 (ZNF160) transcript variant 2
NM_022845	0.937368603240197	341	core-binding factor beta subunit (CBFB) transcript variant 1
NM_178864	0.928825900728436	387	HLH-PAS transcription factor NXF (NXF)
NM_020828	0.925122515754606	401	zinc finger protein 28 homolog (mouse) (ZFP28)
NM_003902	0.924772115264012	404	far upstream element (FUSE) binding protein 1 (FUBP1)
NM_005197	0.924117598694358	408	checkpoint suppressor 1 (CHES1)
NM_000966	0.923905106963148	410	retinoic acid receptor gamma (RARG)
NM_003189	0.923829127082637	411	T-cell acute lymphocytic leukemia 1 (TAL1)
NM_005392	0.923630683311191	414	PHD finger protein 2 (PHF2) transcript variant 1
NM_005243	0.922434402265073	420	Ewing sarcoma breakpoint region 1 (EWSR1) transcript variant EWS
NM_152943	0.920866469357959	434	zinc finger protein 268 (ZNF268) transcript variant B
NM_003430	0.919535823323301	445	zinc finger protein 91 (HPF7 HTF10) (ZNF91)
NM_004454	0.915181977006808	470	ets variant gene 5 (ets-related molecule) (ETV5)
NM_145897	0.9121943038198	489	prefoldin 5 (PFDN5) transcript variant 3
NM_002165	0.912162968807822	490	inhibitor of DNA binding 1 dominant negative helix-loop-helix protein (ID1)
NM_013361	0.910610340086499	496	zinc finger protein 223 (ZNF223)
NM_006197	0.910287574662829	498	pericentriolar material 1 (PCM1)
NM_017742	0.905513921265017	527	zinc finger CCHC domain containing 2 (ZCCHC2)
NM_001938	0.904199877054046	533	down-regulator of transcription 1 TBP-binding (negative cofactor 2) (DR1)
NM_006608	0.90055217437275	554	putative homeodomain transcription factor 1 (PHTF1)
NM_001986	0.896569638648578	568	ets variant gene 4 (E1A enhancer binding protein E1AF) (ETV4)
NM_005239	0.894850086665583	573	v-ets erythroblastosis virus E26 oncogene homolog 2 (avian) (ETS2)
NM_005994	0.892472985467713	585	T-box 2 (TBX2)
NM_004350	0.892429207156144	586	runt-related transcription factor 3 (RUNX3)

Πίνακας 4 Υποεκφρασμένοι μεταγραφικοί παράγοντες ή πρωτεΐνες σχετιζόμενες με τη μεταγραφή στα παρθενικά Β λεμφοκύτταρα σε σχέση με τα παρθενικά Τ κύτταρα.

Κωδικός γονιδίου	Πιθανότητα υποέκφρασης	Σειρά (rank) πιθανότητας	Όνομα γονιδίου
NM_181782	0.99998940599879	16	nuclear receptor coactivator 7 (NCOA7)
NM_022463	0.999952035689186	21	nucleoredoxin (NXN)
NM_020706	0.99994733328388	22	splicing factor arginine/serine-rich 15 (SFRS15)
NM_005373	0.999916574849363	23	myeloproliferative leukemia virus oncogene (MPL)
XM_499313	0.999894526880401	24	PREDICTED: similar to 60S ribosomal protein L35 (LOC442569)
NM_024503	0.999774460616023	30	human immunodeficiency virus type I enhancer binding protein 3 (HIVEP3)
NM_001270	0.998417285264787	40	chromodomain helicase DNA binding protein 1 (CHD1)
NM_005414	0.997597384549399	52	SKI-like (SKIL)
NM_014724	0.995950369747537	69	zinc finger protein 305 (ZNF305)
NM_005900	0.995600939781769	73	SMAD mothers against DPP homolog 1 (Drosophila) (SMAD1)
NM_005521	0.995498960381340	76	T-cell leukemia homeobox 1 (TLX1)
NM_021191	0.993659311968715	95	neurogenic differentiation 4 (NEUROD4)
NM_032679	0.993598715805078	96	zinc finger protein 577 (ZNF577)
NM_175709	0.991918444883713	110	chromobox homolog 7 (CBX7)
NM_052996	0.990989731403731	120	PR domain containing 7 (PRDM7)
NM_138931	0.990941730104390	122	B-cell CLL/lymphoma 6 (zinc finger protein 51) (BCL6) transcript variant 2
NM_007348	0.990191512925451	124	activating transcription factor 6 (ATF6)
XM_045290	0.985243679810346	172	PREDICTED: similar to basic leucine zipper and W2 domains 1 (LOC151579)
NM_005007	0.983747815531926	182	nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor-like 1 (NFKBIL1)
NM_198120	0.983609385075523	183	estrogen receptor binding site associated antigen 9 (EBAG9) transcript variant 2
XM_036115	0.983113229480351	186	PREDICTED: zinc finger CCCH type domain containing 5 (ZC3HDC5)
NM_005748	0.982064599009305	196	YY1 associated factor 2 (YAF2)
NM_015559	0.981953519408775	197	SET binding protein 1 (SETBP1)
NM_005178	0.980161068853945	209	B-cell CLL/lymphoma 3 (BCL3)
NM_144778	0.978804552159057	215	muscleblind-like 2 (Drosophila) (MBNL2) transcript variant 1
NM_007315	0.974910708516002	230	signal transducer and activator of transcription 1 91kDa (STAT1) transcript variant alpha
NM_005057	0.972988610049937	242	retinoblastoma binding protein 5 (RBBP5)
NM_014468	0.971562601473124	248	VENT-like homeobox 2 (VENTX2)
NM_005239	0.971413633046798	250	v-ets erythroblastosis virus E26 oncogene homolog 2 (avian) (ETS2)
NM_015478	0.968883246223585	263	l(3)mbt-like (Drosophila) (L3MBTL) transcript variant I

NM_145897	0.968508554043346	266	prefoldin 5 (PFDN5) transcript variant 3
NM_006442	0.965966960866952	281	DR1-associated protein 1 (negative cofactor 2 alpha) (DRAP1)
NM_005197	0.965554174790170	284	checkpoint suppressor 1 (CHES1)
NM_006902	0.960286525587951	313	paired related homeobox 1 (PRRX1) transcript variant pmx-1a
XM_376776	0.960150519331640	314	PREDICTED: thymus high mobility group box protein TOX (TOX)
NM_182918	0.956517347996458	332	v-ets erythroblastosis virus E26 oncogene like (avian) (ERG) transcript variant 1
NM_006163	0.953863169756005	344	nuclear factor (erythroid-derived 2) 45kDa (NFE2)
NM_032195	0.953398250131743	349	SON DNA binding protein (SON) transcript variant b
NM_032682	0.950215541798995	359	forkhead box P1 (FOXP1)
NM_016331	0.949239812113913	365	zinc finger protein 639 (ZNF639)
NM_080593	0.948815561529329	366	histone 1 H2bk (HIST1H2BK)
NM_001300	0.947141556726606	376	core promoter element binding protein (COPEB)
NM_138610	0.946266200513129	380	H2A histone family member Y (H2AFY) transcript variant 3
NM_005426	0.945166504452407	386	tumor protein p53 binding protein 2 (TP53BP2)
NM_004473	0.944521415580942	389	forkhead box E1 (thyroid transcription factor 2) (FOXE1)
NM_004496	0.944105783788476	393	forkhead box A1 (FOXA1)
NM_030634	0.943670764301309	396	zinc finger protein 436 (ZNF436)
NM_006197	0.936715290976312	428	pericentriolar material 1 (PCM1)
NM_003455	0.932789802372206	448	zinc finger protein 202 (ZNF202)
XM_375261	0.931198315510114	454	PREDICTED: similar to High-mobility group nucleosome binding domain 1 (LOC400452)
NM_005131	0.93078380796243	457	THO complex 1 (THOC1)
NM_198219	0.926579402055057	478	inhibitor of growth family member 1 (ING1) transcript variant 1
NM_003430	0.923053375577287	492	zinc finger protein 91 (HPF7 HTF10) (ZNF91)
NM_020828	0.921911363404529	499	zinc finger protein 28 homolog (mouse) (ZFP28)
NM_012245	0.921864877287942	500	SKI interacting protein (SKIIP)
XM_371197	0.919794263834064	508	PREDICTED: similar to zinc finger protein 111 (LOC388565)
NM_005257	0.918883747855179	515	GATA binding protein 6 (GATA6)
NM_007147	0.918628154091901	516	zinc finger protein 175 (ZNF175)
NM_022466	0.91770692313739	524	zinc finger protein subfamily 1A 5 (ZNFN1A5)
NM_001982	0.917472074962728	526	v-erb-b2 erythroblastic leukemia viral oncogene homolog 3 (avian) (ERBB3)
NM_153715	0.916253724752359	533	homeo box A10 (HOXA10) transcript variant 2
NM_001754	0.916175463828013	534	runt-related transcription factor 1 (acute myeloid leukemia 1; aml1 oncogene) (RUNX1)
NM_000966	0.915880055505268	539	retinoic acid receptor gamma (RARG)

NM_005966	0.91587472733837	540	NGFI-A binding protein 1 (EGR1 binding protein 1) (NAB1)
NM_174900	0.915025496608606	553	zinc finger protein 42 (ZFP42)
NM_002052	0.9145982685076	558	GATA binding protein 4 (GATA4)
NM_000635	0.913499763970901	564	regulatory factor X 2 (influences HLA class II expression) (RFX2) transcript variant 1
NM_002165	0.911736336220236	575	inhibitor of DNA binding 1 dominant negative helix-loop-helix protein (ID1)
NM_016423	0.905721368976696	603	zinc finger protein 219 (ZNF219)
NM_005994	0.90362661995345	612	T-box 2 (TBX2)
NM_014210	0.900689409345095	623	ecotropic viral integration site 2A (EVI2A)
NM_182547	0.90023149842348	628	putative NFkB activating protein HNLF (HNLF)
NM_024840	0.898246347915941	636	zinc finger protein 613 (ZNF613)
XM_372753	0.897637969891162	639	PREDICTED: similar to Zinc finger protein 264 (LOC390980)
NM_003428	0.896139712319479	647	zinc finger protein 84 (HPF2) (ZNF84)
NM_000376	0.895346412604766	650	vitamin D (125- dihydroxyvitamin D3) receptor (VDR)
NM_005392	0.894711822054769	652	PHD finger protein 2 (PHF2) transcript variant 1
NM_005243	0.891559921031958	677	Ewing sarcoma breakpoint region 1 (EWSR1) transcript variant EWS
NM_052923	0.890608569113829	683	zinc finger protein 452 (ZNF452)

Πίνακας 5 Υποεκφρασμένοι μεταγραφικοί παράγοντες ή πρωτεΐνες σχετιζόμενες με τη μεταγραφή στα ενεργοποιημένα Β λεμφοκύτταρα σε σχέση με τα παρθενικά Τ κύτταρα.

Κωδικός γονιδίου	Πιθανότητα υποέκφρασης	Σειρά (rank) πιθανότητας	Όνομα γονιδίου
NM_181782	0.999302565613	1	nuclear receptor coactivator 7 (NCOA7)
NM_138931	0.952440042746	8	B-cell CLL/lymphoma 6 (zinc finger protein 51) (BCL6) transcript variant 2
NM_005414	0.858985114781	17	SKI-like (SKIL)
NM_006197	0.825812616625	26	pericentriolar material 1 (PCM1)
NM_022466	0.802994653471	35	zinc finger protein subfamily 1A 5 (ZNFN1A5)
NM_005239	0.780788124795	41	v-ets erythroblastosis virus E26 oncogene homolog 2 (avian) (ETS2)
NM_005243	0.773318768244	44	Ewing sarcoma breakpoint region 1 (EWSR1) transcript variant EWS
NM_014468	0.760315221143	48	VENT-like homeobox 2 (VENTX2)
NM_005197	0.740735511559	55	checkpoint suppressor 1 (CHES1)
NM_002944	0.735963881395	60	v-ros UR2 sarcoma virus oncogene homolog 1 (avian) (ROS1)
NM_199420	0.729639664551	64	polymerase (DNA directed) theta (POLQ) transcript variant 2

Πίνακας 6 Μεταγραφικοί παράγοντες ή πρωτεΐνες σχετιζόμενες με τη μεταγραφή που υποεκφράζονται στα παρθενικά B λεμφοκύτταρα σε σχέση με τα παρθενικά T κύτταρα ΚΑΙ υποεκφράζονται στα ενεργοποιημένα B σε σχέση με τα παρθενικά T ΚΑΙ υποεκφράζονται στα ενεργοποιημένα T σε σχέση με τα παρθενικά T.

Από ανάγνωση των πινάκων 2-4 προκύπτουν μερικά ενδιαφέροντα στοιχεία. Πρώτα παρατηρούμε ότι οι 3 συγκρίσεις διαφέρουν στον αριθμό των υποεκφρασμένων γονιδίων: μόνο 30 (γνωστοί) παράγοντες υποεκφράζονται στα ενεργοποιημένα σε σχέση με τα παρθενικά T λεμφοκύτταρα. Αντίθετα περίπου διπλάσιος είναι ο αριθμός διαφορών στις συγκρίσεις μεταξύ T και B κυττάρων, εύρημα αναμενόμενο λόγω της μεγαλύτερης ανομοιογένειας των δύο ειδών κυττάρων. Ένας μεγάλος αριθμός μεταγραφικών της οικογένειας «δακτύλων ψευδαργύρου» (zinc fingers) είναι διαφορετικά εκφρασμένοι μεταξύ των T και B κυττάρων (είτε ενεργοποιημένων είτε παρθενικών), ενώ οι διαφορές μεταξύ των ενεργοποιημένων και παρθενικών T κυττάρων φαίνεται ότι πιο ομοιογενώς κατανομημένες.

Γνωρίζοντας ότι ο καταστολέας είναι μια πρωτεΐνη με ειδικότητα έκφρασης στα T παρθενικά λεμφοκύτταρα σε σχέση με τα B (παρθενικά ή ενεργοποιημένα) αποφασίσαμε να επικεντρώσουμε την προσοχή μας στις πρωτεΐνες εκείνες που είναι διαφορετικά εκφρασμένες και στις 3 συγκρίσεις.. Οι πρωτεΐνες αυτές δίνονται στον Πίνακα 5 από τον οποίο προκύπτει ότι μόνο 11 πρωτεΐνες που δεσμεύονται στο DNA περνούν το κριτήριο πιθανότητας υποεκφρασιμότητας (0.89^3) για την τριπλή σύγκριση. Από τις πρωτεΐνες αυτές δεδομένα για τις consensus θέσεις δέσμευσης στο DNA έχουμε για τις ακόλουθες:

- ZNFN1A5 (Pegasus) GXXTGTXG¹⁶¹
- ETS-2 GGA(A/T)¹⁶²
- BCL6 TTCCT(A/C)GAA(A/T) G¹⁶³

Συγκρίνοντας τις θέσεις δέσμευσης των παραγόντων αυτών με τις ακολουθίες της IL-2 και του HIV-1 LTR καταλήγουμε στο συμπέρασμα ότι η επιλογή του παράγοντα ETS-2 οδηγεί στον καλύτερο συνδυασμό πεποιθήσεων υποέκφρασης και πιθανότητας δέσμευσης στα υπό εξέταση στοιχεία. Ο παράγοντας αυτός επιλέχθηκε για τα επιβεβαιωτικά πειράματα του επόμενου κεφαλαίου.

6 Εκτέλεση και αποτελέσματα επιβεβαιωτικών πειραμάτων

Στο τμήμα αυτό θα παρουσιάσουμε τα αποτελέσματα των επιβεβαιωτικών πειραμάτων που σχεδιάστηκαν με βάση τα αποτελέσματα της «υπολογιστικής» διαφορικής έκφρασης.²² Τα αρχικά πειράματα EMSA εναντίον του υποκινητή της IL-2 είχαν σχεδιασθεί στο προηγούμενο στάδιο της παρούσης εργασίας και περιγράφονται εδώ για λόγους πληρότητας. Τα πειράματα αυτά παρείχαν την απαραίτητη επικύρωση των υπολογιστικών ευρημάτων η οποία κρίνεται απαραίτητη για την επιτυχία οποιουδήποτε ερευνητικού σχεδίου, υπολογιστικού η μή.

6.1 Μεθοδολογία πειραμάτων

Κυτταροκαλλιέργειες

Ηπαρινισμένο ολικό αίμα από υγιείς εθελοντές αιμοδότες συλλέχθηκε στο Κέντρο αιμοδοσίας του Περιφερειακού Πανεπιστημιακού Νοσοκομείου Πατρών. Στη συνέχεια τα μονοπύρρηνα κύτταρα του περιφερειακού αίματος απομονώθηκαν με τη βοήθεια ηλεκτροφόρησης βαθμίδωσης Φικόλης (Pharmacia, Upsalla, Sweden). Τα κύτταρα (10^6 /σημείο) επώασθηκαν για 8 ώρες στο καλλιεργητικό μέσο RPMI 1640 (Gibco BRL, Gaithersburg, MD) που περιείχει 10% ορό από έμβρυα βοοειδούς (FCS; Gibco BRL) καθώς και άλλα απαραίτητα θρεπτικά στοιχεία όπως έχει περιγραφεί προηγούμενα. Τα κύτταρα καλλιεργήθηκαν απουσία και παρουσία εστέρα φορβόλης σε συγκέντρωση 10 ng/ml (PMA; Sigma, Buchs, Switzerland καθώς και 5 μg/ml φυτοαιμαγλουτίνης (PHA; Sigma). Τα περιφερικά μονοκύτταρα χρησίμευσαν για τα περαιτέρω πειράματα EMSA χωρίς να υποβληθούν σε επιπλέον βήμα εμπλουτισμού με κυτταρομετρία ροής, καθώς προηγούμενα πειράματα είχαν αποδείξει ότι ο υπο διερεύνηση καταστολέας βρίσκεται σε ενεργή μορφή μόνο στα CD4 + κύτταρα¹⁶⁴. Στην περίπτωση που τα κύτταρα πρόκειται να ενεργοποιηθούν για αλυσιδωτή αντίδραση πολυμεράσης, τότε ακολουθήθηκε μεθοδολογία αρνητικής επιλογής των CD14+, CD19+ με ανοσομαγνητικά σωματίδια Dynabeads.

Απομόνωση πυρηνικών εκχυλίσμάτων από T λεμφοκύτταρα

Οι πρωτεΐνες απομονώθηκαν από ολικά εκχυλίσματα κυττάρων όπως είχε περιγραφεί προηγούμενα¹⁶⁵. Εν συντομία, τα κύτταρα υποβλήθηκαν σε υποτονικό σοκ, ακολουθούμενο από διάρρηξη των μεμβρανών με τη βοήθεια υπερηχητικών κυμάτων. Οι αντιδράσεις αυτές έγιναν σε ρυθμιστικό διάλυμα συγκέντρωσης KCl 300 mM ικανό να απομονώσει τις πρωτεΐνες από τη χρωματίνη.

Ακολούθως τα εκχυλίσματα επώαστηκαν σε πάγο για 20 λεπτά και υποβλήθηκαν σε φυγοκέντρωση σε 100000 g για 1 hr στους 4°C. Το υπερκείμενο συλλέχθηκε σε ρυθμιστικό διάλυμα που περιείχε 20 mM HEPES, 0.02 mM EDTA, pH 7.6, 0.5 mM 1,4 dithio-DL-threitol. Η συγκέντρωση των πρωτεϊνών μετρήθηκε με τη μέθοδο Bradford. Πριν από τη διενέργεια των πειραμάτων EMSA τα πρωτεϊνικά εκχυλίσματα κατακρημνίστηκαν με 10% PolyEtheleneGlycol (PEG8000) με συνεχή ανάδευση ΓΙΑ 20 min στους 4°C. Το διάλυμα αυτό στη συνέχεια φυγοκεντρήθηκε στις 10000g για 20 λεπτά και το καθίζημα επαναδιαλύθηκε σε ρυθμιστικό διάλυμα KIB buffer και αποθηκεύτηκε. Το πρωτόκολλο αυτό είχε δειχθεί ότι οδηγεί σε εμπλουτισμό

²² Τα πειράματα αυτά διεξήχθησαν από τους Παύλο Αδαμόπουλο και Τάσσο Γεωργακόπουλο στα πλαίσια της συνεργατικής προσπάθειας για τον εντοπισμό μεταγραφικών καταστολέων του υποκινητή της ιντερλευκίνης 2 και του LTR του ιού HIV

τουλάχιστον κατά 10 φορές της πρωτεΐνης του μεταγραφικού καταστολέα σε προηγούμενες δημοσιεύσεις¹⁶⁶.

Πειράματα EMSA

Στήλες πρωτεΐνης A-σεφαρόζης (Amersham) προετοιμάστηκαν σε μείγμα 1:1 κατ'όγκο σε διάλυμα KIB (PAS-KIB). 10 μl PAS-KIB αναμείχθηκαν με 10 μl of πρωτεϊνικού εκχυλίσματος και επωάστηκαν στους 4⁰ C για 1 hr. Ακολούθως το μείγμα της αντίδρασης φυγοκεντρήθηκε στις 27000g και το υπερκείμενο προεπωάστηκε με 1 μg of αντισώματος για δεδομένο μεταγραφικό παράγοντα TF-στους 4⁰ C για μία ώρα. Τα αντισώματα που χρησιμοποιήθηκαν είχαν ειδικότητα για τους ακόλουθους μεταγραφικούς παράγοντες : c-fos p62, fos-B, fra-1, fra-2 p46, c-jun p 49, jun-B p 39, jun-D p 39, NFκB p65, NFκB p50, c-rel p 75, ets-1 p54, πολυκλωνικό anti ets1/ets2, p53. Όλα τα αντισώματα προήλθαν από τους καταλόγους της Santa Cruz biotechnology Inc (Santa Cruz CA). Το υπερκείμενο κεκαθαρμένο από τον αντίστοιχο μεταγραφικό παράγοντα χρησιμοποιήθηκε σε EMSA πειράματα εναντίον ολιγονουκλεοτιδικού ιχνηθέτη ειδικού ως προς τη θέση δέσμευσης του NFAT: AAGAAAGGAGGAAAAACTGTTT

DNA κλώνοι

Ένα όχημα έκφρασης του ets-2 , pcDNA3-ETS2, κατασκευάστηκε με την εισαγωγή τμήματος του ανθρώπινου ets-2 cDNA μετά από ενίσχυση με PCR των προϊόντων της πέψη του τελευταίου με το περιοριστικό ένζυμο HindIII/BamHI . Οι οδηγοί για την αντίδραση της PCR ήταν οι ακόλουθοι: 5'-CCCAAGCTTGGCAGGATGAATGATTTTCGG και 5'-CGGGATCCTCAGTCCTCCGTGTCG . Επιπλέον δημιουργήθηκαν τα αντίστοιχα πλασμίδια-στόχοι: (1) Το πλασμίδιο IL-2-LUC που περιέχει το γονίδιο της λουσιφεράσης υπό τον έλεγχο τμήματος του υποκινητή του γονιδίου της ιντερλευκίνης 2 (θέσεις -326 to +45) (2) πλασμίδιο NFAT-CAT που περιέχει 3 αντίγραφα της θέσης δέσμευσης του NFAT στον υποκινητή της IL-2(θέσεις -285 μέχρι -255) εν σειρά με τις θέσεις -72 μέχρι +45 του υποκινητή του γονιδίου της ιντερλευκίνης 2. Το σύνθετο αυτό στοιχείο ελέγχει την έκφραση του γονιδίου της ακετυλάσης της χλωραμφενικόλης (CAT) (3) το πλασμίδιο CMV-CAT που χρησιμοποιεί τον υποκινητή του ιού CMV για να ελέγξει την έκφραση του ενζύμου CAT χρησίμευσε σαν πλασμίδιο αναφοράς.

Διαμολύνσεις

Κύτταρα Jurkat διαμολύνθηκαν με ηλεκτροπόρηση (electroporation) με τη βοήθεια συσκευής Gene Pulser (ECM399, BTX, San Diego, CA) στα 260V/28msec. 8x10⁶ κύτταρα/σημείο αναμείχθηκαν με 10μg πλασμιδιακού γονιδίου στόχου με και χωρίς 5, 10, 20 or 50μg pcDNA3-ETS2, σε 0.35 ml μέσου RPMI1640 εντός κυβέττες χάσματος 0.4cm. Μετά τη διαμόλυνση τα κύτταρα αφέθηκαν να ανακάμψουν εντός καλλιεργητικού μέσου RPMI1640 εμπλουτισμένου με 5% FCS, penicillin and streptomycin για 24h εντός επωαστή (37⁰C, 5%CO₂) . Στη συνέχεια ενεργοποιήθηκαν με τα μιτογόνα ιονομυκίνη (1μg/ml) και εστέρα φορβολης (PMA, 25ng/ml) και επωάστηκαν για επιπλέον 24h. Εναλλακτικά 5x10⁶ Jurkat κύτταρα/σημείο διαμολύνθηκαν με 70 μl του λιποσωματικού παράγοντα (LipoTAXI, Stratagene, Cedar Creek, TX) ο οποίος είχε προ-αναμειχθεί με 25 μg DNA στόχου με ή χωρίς 5, 10, 20 or 50μg pcDNA3-ETS2 σε όγκο 1 ml RPMI1640. Στη συνέχεια επωάστηκαν για 2h,

και αφού αφέθηκαν να ανακάμψουν για άλλες 22h εντός του επωαστή, ενεργοποιήθηκαν με μιτογόνα ως ανωτέρω και επώαστηκαν για άλλες 12h.

CD14-CD19- κύτταρα (8×10^6 /σημείο) διαμολύνθηκαν με 15 μl λιποσωματικού παράγοντα (LipoTAXI, Stratagene, Cedar Creek, TX) που είχε προ-αναμειχθεί με 10 μg πλασμιδιακού DNA (pcDNA3-ETS-2 ή pUC-BENN-CAT που περιέχει ολόκληρο το LTR του ιού HIV) για 2 hr. Μετά τη διαμόλυνση τα κύτταρα αφέθηκαν να ανακάμψουν εντός καλλιεργητικού μέσου RPMI1640 εμπλουτισμένου με 5% FCS, penicillin and streptomycin για 24h εντός επωαστή (37°C , 5% CO_2). Στη συνέχεια ενεργοποιήθηκαν με τα μιτογόνα ιονομυκίνη (1μg/ml) και εστέρα φορβολής (PMA, 25ng/ml) και επώαστηκαν για επιπλέον 24h.

Δοκιμασίες λουσιφεράσης και ακετυλάσης της χλωραμφενικόλης

Η ενεργότητα της λουσιφεράσης μετρήθηκε σε εκχυλίσματα όγκου 30 μl εκφραζόμενες σαν σχετικές μονάδες φωτός με τη βοήθεια του TD 20/20 λουμινόμετρου. Η αποτελεσματικότητα της αντίδρασης διαμόλυνσης ελέγχθηκε με τη συν-διαμόλυνση κυττάρων με 1μg pRL-TK που χρησιμοποιεί τον υποκινητή της θυμικινικής κινάσης το ιού του απλού έρπητα για να ρυθμίσει την έκφραση της of Renilla Luciferase. Η ενεργότητα της Renilla luciferase μετρήθηκε με τη διπλή δοκιμασία Dual-Luciferase reporter assay system (Promega, Madison, WI). Οι δοκιμασίες ενεργότητας CAT έγιναν σε κυτταρικά εκχυλίσματα 40 μl

Αντιδράσεις αλυσσιδωτής αντίδρασης πολυμέρασης

ETS-2 sense 5'-cgc agc ggc agg atg aat g-3' (που αντιστοιχεί στις θέσεις 151-169 του cDNA) antisense 5'-agg aac gga ggt gag gtg tg-3' (νουκλεοτίδια 707-726). Step 1) 95C 4min 2) 95C for 45sec 3) 62C 1min 4) 72C 1min (ολικός αριθμός κύκλων s 38) 5) 72C 10min.

IL-2 sense 5'-gca act cct gtc ttg cat tg-3' antisense 5'-aat gtg agc atc ctg gtg ag-3'. Πρόγραμμα step 1) 94C 5min 2) 94C for 45sec 3) 58C 30sec 4) 72C 1min (total number of cycles 36) 5) 72C 10min.

6.2 Πειραματική επιβεβαίωση της ειδική έκφρασης του ETS-2 στα παρθενικά T λεμφοκύτταρα

Η εφαρμογή του πλαισίου αναζήτησης υποεκφρασμένων γονιδίων με βάση την κατανομή Behrens Fisher υπέδειξε ότι το γονίδιο του μεταγραφικού παράγοντα ETS-2 υποεκφράζεται στα ενεργοποιημένα T κύτταρα (όσο και στα B κύτταρα ανεξάρτητα από την κατάσταση της ενεργοποίησης) σε σχέση με τα παρθενικά T κύτταρα. Η επιβεβαίωση της εγκυρότητας του ευρήματος αυτού βασίζεται στην ανάλυση της έκφρασης του αντίστοιχου γονιδίου με RT-PCR στους αντίστοιχους κυτταρικούς πληθυσμούς.

Όταν εμπλουτισμένοι πληθυσμοί ex-vivo εμπύρηνων κυττάρων περιφερικού αίματος σε T λεμφοκύτταρα αναλύθηκαν για τα επίπεδα της μεταγραφής των IL-2 και ETS-2 κατέστη εμφανής η αντίστροφη σχέση μεταξύ της έκφρασης του mRNA του IL-2 και ETS-2 (Εικόνα 28 στήλη 1 και στις δύο φωτογραφίες). Όταν τα κύτταρα παρέμειναν σε καλλιεργητικό μέσο η σχέση αυτή παρέμεινε αμετάβλητη (Εικόνα 28 στήλη 2) και αντιστράφηκε μόνο κατά την ενεργοποίηση των T λεμφοκυττάρων (Εικόνα 28 στήλη 3). Το γεγονός ότι τα επίπεδα του mRNA του ETS-2 μειώνονται κατά την ενεργοποίηση των T κυττάρων όχι μόνο επιβεβαιώνει το αποτέλεσμα της στατιστικής ανάλυσης των μικροσυστοιχειών που παρουσιάστηκε σε προηγούμενο τμήμα (σύγκριση 1) , αλλά αποτελεί ένδειξη της συμμετοχής του παράγοντα αυτού στη ρύθμιση του γονιδίου της ιντερλευκίνης 2.

Εικόνα 28 Ποσοτικοποίηση IL-2 (αριστερή φωτογραφία) και ETS-2 (φωτογραφία δεξιά) σε ex vivo T κύτταρα (1^η στήλη), T κύτταρα σε καλλιεργητικό μέσο (2^η στήλη) και μετά από ενεργοποίηση (3^η στήλη). Παρατηρούμε ότι η έκφραση του ETS-2 μειώνεται μετά την ενεργοποίηση των T λεμφοκυττάρων, ενώ η αντίστροφη σχέση παρατηρείται (όπως είναι αναμενόμενο) για την ιντερλευκίνη 2. Τα αποτελέσματα αυτά επικυρώνουν τα αποτελέσματα της σύγκρισης 1 που περιγράφηκε στο κεφάλαιο ανάλυσης των πειραμάτων μικροσυστοιχειών

Η ειδικότητα της έκφρασης του γονιδίου του ETS-2 μελετήθηκε περαιτέρω στις υποκατηγορίες T και B λεμφοκυττάρων των εμπύρηνων κυττάρων του περιφερικού αίματος και ορισμένων λευχαιμικών σειρών. Παρατηρούμε ότι η έκφραση του ETS-2 περιορίζεται στα εν ηρεμία T λεμφοκύτταρα, και εξαφανίζεται μετά την ενεργοποίηση των λεμφοκυττάρων αυτών (Εικόνα 29). Η ανάλυση της έκφρασης της ιντερλευκίνης 2 στους ίδιους πληθυσμούς κυττάρων (Εικόνα 31) πιστοποιεί επίσης την

- ο ειδικότητα των πρωτοκόλλων απομόνωσης των διαφόρων υποπληθυσμών κυττάρων (ο πληθυσμός των CD19+ κυττάρων που αντιστοιχεί στα B λεμφοκύτταρα δεν εκφράζει όπως και αναμένεται το γονίδιο της ιντερλευκίνης 2)
- ο επιτυχή ενεργοποίηση των T λεμφοκυττάρων (η έκφραση της ιντερλευκίνης 2 συμβαίνει μόνο μετά την ενεργοποίηση των κυττάρων αυτών)

Εικόνα 29 Η έκφραση του γονιδίου ETS-2 είναι χαρακτηριστική των εν ηρεμία T λεμφοκυττάρων. Πειράματα RT-PCT δείχνουν ότι τα B λεμφοκύτταρα (CD19+) δεν εκφράζουν το ETS-2 όταν απομονώνονται ex-vivo (19e) αλλά ούτε κατά την παραμονή τους σε καλλιεργητικό μέσο (19c) και την επακόλουθη ενεργοποίησή τους με μιτογόνα (19i). Αντίθετα η έκφραση του ETS-2 είναι χαρακτηριστική των T κυττάρων (είτε ex-vivo (Te) είτε κατά την παραμονή τους σε καλλιεργητικό μέσο (Tc). Μετά την ενεργοποίηση των T λεμφοκυττάρων (19i) το ETS-2 σταματά να εκφράζεται.

Εικόνα 30 Ανάλυση της έκφρασης του γονιδίου της ιντερλευκίνης 2 με RT-PCT στα δείγματα του προηγούμενου σχήματος. Τα Β λεμφοκύτταρα (CD19+) δεν εκφράζουν την IL-2 όταν απομονώνονται ex-vivo (19e) αλλά ούτε κατά την παραμονή τους σε καλλιεργητικό μέσο (19c) και την επακόλουθη ενεργοποίηση τους με μιτογόνα (19i). Αντίστοιχα η έκφραση της IL-2 είναι ανύπαρκτη στα εν ηρεμία Τ κυττάρων (είτε ex-vivo (Te) είτε κατά την παραμονή τους σε καλλιεργητικό μέσο (Tc). Μετά την ενεργοποίηση των Τ λεμφοκυττάρων (19i) η IL-2 εκφράζεται με πολύ έντονο τρόπο. Τα αποτελέσματα της εικόνας αυτής επιβεβαιώνουν την ειδικότητα των πρωτοκόλλων απομόνωσης που επιτυγχάνουν την απομόνωση Τ και Β κυττάρων χωρίς να επάγουν την ενεργοποίηση των πληθυσμών αυτών.

Έτσι και μέσω PCR επιβεβαιώνονται τα αποτελέσματα των 3 συγκρίσεων μικροσυστοιχειών που παρουσιάστηκαν σε προηγούμενο κεφάλαιο. Η σύγκριση 1 (παρθενικά Τ με ενεργοποιημένα Β) αντιστοιχεί στην σύγκριση της έντασης της ζώνης του ETS-2 στη στήλη (Lane) Ti με τις στήλες Te/Tc. Κατά ανάλογο τρόπο η σύγκριση 2 (παρθενικά Τ με παρθενικά Β) αντιστοιχεί στη σύγκριση της έντασης της ζώνης του ETS-2 mRNA μεταξύ των στηλών Te/Tc με τις στήλες 19e/19c, ενώ η σύγκριση 3 (παρθενικά Τ με ενεργοποιημένα Β) στις στήλες Te/Tc και 19i. Η διαφορετική έκφραση του ETS-2 στα εν ηρεμία παρθενικά Τ λεμφοκύτταρα επιβεβαιώνει τις αναλύσεις στατιστικής διαφορικής έκφρασης με τη χρήση της κατανομής Behrens Fisher και στις 3 συγκρίσεις που πραγματοποιήθηκαν.

6.3 Ο ETS-2 δρα ως καταστολέας του υποκινητή της ιντερλευκίνης-2

Σε προηγούμενο κεφάλαιο είδαμε ότι ο μεταγραφικός παράγοντας ETS-2 εμφανίζεται να έχει μεγάλη πιθανότητα υποέκφρασης, όπως αυτή αποτυπώθηκε με τη σύγκριση των παρθενικών με τα ενεργοποιημένα Τ λεμφοκύτταρα, αλλά και στην «τριπλή» σύγκριση (παρθενικά Τ vs ενεργοποιημένα κύτταρα) x (παρθενικά Τ vs ενεργοποιημένα κύτταρα) x (παρθενικά Τ vs ενεργοποιημένα κύτταρα). Ακολουθώντας τη μεθοδολογία που παρουσιάστηκε στο κεφάλαιο της μεθοδολογίας ανίχνευσης μεταγραφικών καταστολέων, προχωρήσαμε στον έλεγχο της δέσμευσης του στο στοιχείο NFAT που βρίσκεται στον υποκινητή της ιντερλευκίνης 2.

Αποτελέσματα πειραμάτων EMSA

Σε προηγούμενη μελέτη είχε δείχθει ότι ο παράγοντας που ψάχνουμε είναι μια μικρή πρωτεΐνη μοριακού βάρους 50-60 kDa (μέγεθος συμβατό με τον παράγοντα ETS-2). Με βάση τα ανωτέρω αποφασίστηκε να ελεγχθεί ο παράγοντας ETS-2 σε πειράματα

depletion EMSA. Ποικιλία αντισωμάτων εναντίον άλλων μεταγραφικών παραγόντων χρησιμοποιήθηκαν σαν αρνητικά controls. Οι αντίστοιχοι μεταγραφικοί παράγοντες (c-fos p62, fos-B, fra-1, fra-2 p46, c-jun p 49, jun-B p 39, jun-D p 39, NFκB p65, NFκB p50, c-rel p 75, p53 ογκογονίδιο)έχουν δειχθεί ότι είτε δεν δεσμεύονται με το προς εξέταση στοιχείο, είτε ότι δεν βρίσκονται σε παρθενικά T λεμφοκύτταρα. «Αφαίρεση» των παραγόντων αυτών δεν οδήγησε σε μείωση της έντασης της ζώνης στο πείραμα EMSA (επόμενη εικόνα). Αντίθετα χρήση anti-ets-1 και anti-ets-1/2 αντισωμάτων οδήγησε σε μείωση της έντασης της ζώνης που αντιστοιχεί στο σύμπλοκο ολιγονουκλεοτιδίου – μεταγραφικού παράγοντα υποδηλώνοντας ότι η αντίστοιχη πρωτεΐνη είτε ταυτίζεται, είτε έχει σημαντική δομική ομοιότητα με το ETS-2. Το πείραμα αυτό ενώ επιβεβαιώνει ότι ο παράγοντας ETS-2 1) υπάρχει στα εκχυλίσματα παρθενικών T λεμφοκυττάρων και 2) ότι δεσμεύεται στο υπό εξέταση στοιχείο ΔΕΝ παρέχει στοιχεία για τη λειτουργική σημασία (αν υπάρχει) της δέσμευσης αυτής. Για την απόδειξη αυτή απαιτούνται τα πειράματα συνδιαμόλυνσης τα αποτελέσματα των οποίων αναλύονται παρακάτω.

Αποτελέσματα πειράματων διαμόλυνσης

Για να δείξουμε τη λειτουργική σημασία του ETS-2 στην έκφραση στοιχείων που προέρχονται από τον υποκινητή της ιντερλευκίνης 2 σχεδιάσαμε πειράματα συνδιαμόλυνσης του γονιδίου του ETS-2 σε συνδυασμό με οχήματα που χρησιμοποιούσαν στοιχεία του υποκινητή της ιντερλευκίνης 2 για να ρυθμίσουν την έκφραση των ενζύμων CAT και λουσιφεράσης. Το πρώτο πείραμα, (που χρησιμοποίησε την περιοχή -326 to +45 του υποκινητή της IL-2) δείχνει μια δόσοεξαρτώμενη πτώση της ενεργότητας του ενζύμου της λουσιφεράσης όταν (συν-) διαμολύνεται με μεταβλητή ποσότητα πλασμιδίου που φέρει το ETS-2. Η μείωση της ενεργότητας της λουσιφεράσης εκπεφρασμένης σαν ποσοστό της ενεργότητας της IL-2/ πλασμίδιο αναφοράς, κυμαινόταν σε εύρος 25 μέχρι 45% όταν η ποσότητα του ETS-2 πλασμιδίου που (συν-) διαμολύνθηκε αυξήθηκε από 5 σε 20μg 8x10⁶ κύτταρα (Εικόνα 32). Παρόμοια μείωση παρατηρήθηκε όταν χρησιμοποιήθηκε ένα διαφορετικό

Εικόνα 31 Πειράματα «αφαιρετικής» EMSA εναντίον πληθώρας αντισωμάτων ειδικών προς τους μεταγραφικούς παράγοντες που αναφέρονται στο κείμενο. Τα αντισώματα εναντίον των παραγόντων ETS1 and ETS 1/2 «αφαιρούν» τον άγνωστο παράγοντα.

πλασμίδιο που έφερε τρία αντίγραφα του στοιχείου IL-2 PRRE (θέσεις -285 μέχρι -255 του υποκινητή της ιντερλευκίνης 2) στο 5' άκρο του γονιδίου της ακετυλάσης της χλωραμφαινικόλης. Η απόλυτη μείωση της ενεργότητας του ενζύμου CAT (εκπεφρασμένη σε % μετατροπή της χλωραμφαινικόλης στις ακετυλιωμένες της μορφές κυμάνθηκε από 25 μέχρι 89% (Εικόνα 33, Εικόνα 34). Η επανάληψη των πειραμάτων με ένα διαφορετικό πλασμίδιο στόχο που ρύθμιζε την έκφραση του γονιδίου της CAT μέσω του υποκινητή του κυτταρομεγαλοϊού δεν έδειξε παρόμοια δοσο-εξαρτώμενη σχέση. Αν μη τι άλλο, η συν-διαμόλυνση του ETS-2 με το CMV-CAT αύξησε την έκφραση του τελευταίου Εικόνα 35.

IL-2, ETS-2 Cotransfection

Εικόνα 32 Δοσοεξαρτώμενη καταστολή πλασμιδίου που φέρει περιοχή του υποκινητή της IL-2 από το γονίδιο του μεταγραφικού παράγοντα ETS-2. Η λευχαιμική κυτταρική σειρά Jurkat συνδιαμολύνθηκε με πλασμίδια που έφεραν το γονίδιο του μεταγραφικού παράγοντα ETS-2 και πλασμίδια με περιοχές του υποκινητή της IL-2 έμπροσθεν του γονιδίου της λουσιφεράσης. Ακολούθως τα κύτταρα ενεργοποιήθηκαν, και η ενεργότητα της λουσιφεράσης μετρήθηκε συναρτήσει της ποσότητας ETS-2 που συνδιαμολύνθηκε. Η μείωση της ενεργότητας της λουσιφεράσης κυμαίνεται από 25% (5 µg ETS-2 plasmid) σε 45% (για 20 µg συνδιαμολυνόμενου πλασμιδίου ETS-2)

Εικόνα 33 Δοσοεξαρτώμενη καταστολή της ενεργότητας της ακετυλάσης της χλωραμφαινικόλης (CAT) ρυθμιζόμενης από 3 αντίγραφα του στοιχείου IL-2 PRRE όταν συν-διαμολύνεται με μεταβαλλόμενη ποσότητα πλασμιδίου ETS-2 (αριθμοί στο κάτω άκρο της χρωματογραφίας

λεπτής στοιβάδας: ποσότητα ETS-2 σε $\mu\text{g}/8 \times 10^6$ κύτταρα). Το ίδιο φαινόμενο παρατηρήθηκε και με πλασμίδιο λουσιφεράσης.

Εικόνα 34 Δοσοεξαρτώμενη καταστολή πλασμιδίου που φέρει περιοχή του υποκινητή της IL-2 από το γονίδιο του μεταγραφικού παράγοντα ETS-2. Η λευχαιμική κυτταρική σειρά Jurkat συνδιαμολύνθηκε με πλασμίδια που έφεραν το γονίδιο του μεταγραφικού παράγοντα ETS-2 και πλασμίδια με περιοχές του υποκινητή της IL-2 (3 x IL-2 PRRE) έμπροσθεν του γονιδίου της λουσιφεράσης. Ακολούθως τα κύτταρα ενεργοποιήθηκαν, και η ενεργότητα της λουσιφεράσης μετρήθηκε συναρτήσει της ποσότητας ETS-2 που συνδιαμολύνθηκε. Η μείωση της ενεργότητας της λουσιφεράσης κυμαίνεται από 30% (5 μg ETS-2 plasmid) σε 80% (για 20 μg συνδιαμολυνόμενου πλασμιδίου ETS-2)

Εικόνα 35 Πειράματα συνδιαμολύνσης με πλασμίδιο αναφοράς (CMV-CAT) που δείχνουν την ειδικότητα του κατασταλτικού ρόλου του ETS-2 σε στοιχεία του υποκινητή της ιντερλευκίνης 2.

Ποσοτικοποίηση της σχέσης ιντερλευκίνης 2 και ETS-2 σε φυσιολογικά T λεμφοκύτταρα

Τα αποτελέσματα των πειραμάτων EMSA και συνδιαμόλυνσης αποδεικνύουν την κατασταλτική επίδραση του μεταγραφικού παράγοντα ETS-2 στην έκφραση του γονιδίου της ιντερλευκίνης 2. Έχοντας αφετηρία αυτό το εύρημα προχωρήσαμε σε μια πιο ενδελεχή ημιποσοτική μελέτη του φαινομένου της καταστολής αυτής στα T κύτταρα με την εκτέλεση κινητικών μελετών RT-PCR και πειραμάτων Western Blotting.

Η ανάλυση της κινητικής της έκφρασης των γονιδίων της ιντερλευκίνης 2 και ETS-2 σε διαφορετικά χρονικά σημεία μετά την ενεργοποίηση των T λεμφοκυττάρων αναδεικνύει επίσης την αντίστροφη σχέση μεταξύ της έκφρασης των δύο γονιδίων. Από την κινητική αυτή μελέτη βλέπουμε ότι το mRNA της ιντερλευκίνης-2 αρχίζει να πρωτο-εκφράζεται 4 ώρες μετά την μιτογονική ενεργοποίηση των T κυττάρων χρονικό σημείο στο οποίο η ζώνη του ETS-2 αρχίζει να μειώνεται. Αντίστοιχα το μέγιστο της έκφρασης της ιντερλευκίνης 2 συμπίπτει με το χρονικό σημείο που η έκφραση του γονιδίου ETS-2 πέφτει κάτω από τα όρια της ανίχνευσης της μεθόδου της RT-PCR. Το γεγονός ότι στα πειράματα μικροσυστοιχειών που χρησιμοποιήθηκαν στα πλαίσια της παρούσης εργασίας η αποτύπωση της γονιδιακής έκφρασης έγινε 3 ώρες μετά την ενεργοποίηση των T κυττάρων παρέχει περαιτέρω στήριξη στην εγκυρότητα της χρήσης των Bayesian μεθόδων σαν τεχνικών διεξαγωγής πειραμάτων στατιστικής διαφορικής έκφρασης.

Εικόνα 36 Κινητική μελέτη της έκφρασης του πρωτο-ογκογονιδίου ETS-2 μετά την ενεργοποίηση των T λεμφοκυττάρων. Εμπύρηνα κύτταρα περιφερικού αίματος εμπλουτίστηκαν σε T κύτταρα μετά από χρήση ινών νάυλον. Ο τελικός κυτταρικός πληθυσμός (>95% T κύτταρα) ενεργοποιήθηκε με μιτογόνα και η έκφραση του γονιδίου ETS-2 μελετήθηκε με RT-PCR σαν συνάρτηση του χρόνου σε ώρες (αριθμοί πάνω από τις στήλες της αντίδρασης PRC) μετά την ενεργοποίηση των T κυττάρων. Παρατηρούμε ότι η έκφραση του γονιδίου ETS-2 αρχίζει να μειώνεται από την 4^η ώρα και πέφτει κάτω από τα όρια ανιχνευσιμότητας της μεθόδου μετά την 10^η ώρα.

Εικόνα 37 Κινητική μελέτη της έκφρασης της ιντερλευκίνης 2 μετά την ενεργοποίηση των T λεμφοκυττάρων. Εμπύρηνα κύτταρα περιφερικού αίματος εμπλουτίστηκαν σε T κύτταρα μετά από χρήση ινών νάυλον. Ο τελικός κυτταρικός πληθυσμός (>95% T κύτταρα) ενεργοποιήθηκε με μιτογόνα και η έκφραση του γονιδίου IL-2 μελετήθηκε με RT-PCR σαν συνάρτηση του χρόνου σε ώρες (αριθμοί πάνω από τις στήλες της αντίδρασης PRC) μετά την ενεργοποίηση των T

κυττάρων. Παρατηρούμε ότι η έκφραση του γονιδίου της ιντερλευκίνης 2 αρχίζει να αυξάνεται από την 4^η ώρα και φτάνει τη μέγιστη ένταση μετά την 10^η ώρα ακριβώς το αντίστροφο .

Η κινητική αυτή ανάλυση μελετήθηκε στην περίπτωση του ETS-2 και σε πρωτεϊνικό επίπεδο σε κύτταρα αίματος ομφάλιου λώρου, πηγή εξαιρετικά πλούσια σε παρθενικά T λεμφοκύτταρα (Εικόνα 38) .

Εικόνα 38 Κινητική μελέτη των επιπέδων της πρωτεΐνης του γονιδίου ETS-2 μετά από την ενεργοποίηση κυττάρων ομφάλιου λώρου (πηγή εμπλουτισμένη σε παρθενικά T κύτταρα). Τα επίπεδα της πρωτεΐνης μειώνονται μετά την ενεργοποίηση των κυττάρων και σχεδόν εξαφανίζεται 12-18 ώρες μετά την ενεργοποίηση των T κυττάρων, χρονική κλίμακα που συμφωνεί με την κινητική του αντίστοιχου mRNA

Παρατηρούμε ότι τα επίπεδα της πρωτεΐνης σχεδόν μηδενίζονται μεταξύ της 12^{ης} και 18^{ης} ώρας μετά την ενεργοποίηση των T κυττάρων, χρονική κλίμακα συμβατή με την κινητική του αντίστοιχου mRNA.

6.4 Ο ETS-2 δρα ως καταστολέας του HIV-LTR

Έχοντας δείξει την κατασταλτική επίδραση του ETS-2 στον υποκινητή της ιντερλευκίνης 2 εξετάζουμε την επίδραση του παράγοντα ETS-2 στον υποκινητή του ιού HIV. Λόγω της ομοιότητας των δύο στοιχείων (δηλαδή του PU στον υποκινητή της ιντερλευκίνης 2 και του αντίστοιχου τμήματος στον υποκινητή του HIV) και προηγούμενων εργασιών ¹⁶⁷ παραλείπουμε το πείραμα δέσμευσης (EMSA) .

Στα πειράματα συνδιαμόλυνσης χρησιμοποιήθηκε ένα εύρος συγκεντρώσεων του πλασμιδίου ETS-2 και δύο διαφορετικά πλασμίδια που φέρουν στοιχεία του υποκινητή του ιού HIV. Το πλασμίδιο PUCBEN περιέχει όλο το LTR του ιού ενώ το πλασμίδιο 2xRATS φέρει την αλληλουχία που δεσμεύεται ο καταστολέας. Η ενεργότητα CAT μειώνεται από 37% σε 26% καθώς η ποσότητα του συνδιαμολυνμένου ETS-2 πλασμιδίου αυξάνεται από 0 σε 40 pg , μεταβολή στατιστικά σημαντική ($P < 0.002$, two-sided T test ανεξάρτητων δειγμάτων) στην περίπτωση του πλασμιδίου που φέρει τον υποκινητή του γονιδίου της ιντερλευκίνης 2. Ενδιάμεσες συγκεντρώσεις πλασμιδίου ETS-2 οδηγούν σε μικρότερα επίπεδα καταστολής τα οποία είναι επίσης στατιστικά σημαντικά (π.χ. για συγκέντρωση 20 pg η μείωση της ενεργότητας από 35% σε 23% έχει τιμή πιθανότητας $P < 0.05$). Στην περίπτωση του πλασμιδίου 2xRATS η καταστολή της ενεργότητας CAT είναι στατιστικά σημαντική μόνο για τις υψηλότερες συγκεντρώσεις του πλασμιδίου ETS-2 (μείωση από 34% σε 26% με $P < 0.009$) αν και οι μετρήσεις για το στοιχείο αυτό χαρακτηρίζονταν από μεγαλύτερη μεταβλητότητα γεγονός που μειώνει την ευαισθησία ανίχνευσης σημαντικών ποσοτικών διαφορών.

Εικόνα 39 Μείωση της ενεργότητας CAT οδηγούμενη από το στοιχείο PUCBEN. Η ενεργότητα CAT (μετρούμενη σε % μετατροπής της χλωραμφενικόλης στις ακετυλιωμένες μορφές της) μειώνεται από 37% (όταν δεν χρησιμοποιείται ETS-2) σε 26% όταν συνδιαμολώνεται με 40 µg πλασμιδίου που φέρει το ETS-2. Οι μετρήσεις εκφράζονται σαν μέση τιμή τριών μετρήσεων $\pm$ δειγματικό σφάλμα της μέσης τιμής

Εικόνα 40 Μείωση της ενεργότητας CAT οδηγούμενη από το στοιχείο 2xRATS. Η ενεργότητα CAT (μετρούμενη σε % μετατροπής της χλωραμφενικόλης στις ακετυλιωμένες μορφές της) μειώνεται από 35% (όταν δεν χρησιμοποιείται ETS-2) σε 23% όταν συνδιαμολώνεται με 40 µg πλασμιδίου που φέρει το ETS-2. Οι μετρήσεις εκφράζονται σαν μέση τιμή τριών μετρήσεων $\pm$ δειγματικό σφάλμα της μέσης τιμής

Αντίστοιχη είναι και η συμπεριφορά των δύο πλασμιδίων που προέρχονται από στοιχεία του υποκινητή της ιντερλευκίνης 2. Για να το δείξουμε αυτό εκτελέσαμε μια νέα σειρά πειραμάτων συνδιαμόλυνσης σε λευχαιμικά κύτταρα Jurkat χρησιμοποιώντας το ίδιο εύρος συγκεντρώσεων πλασμιδίου ETS-2 συνδιαμολυνμένου με πλασμίδιο που έφερε ολόκληρο τον υποκινητή της ιντερλευκίνης 2 ή το στοιχείο 4PU. Στα πειράματα αυτά η ενεργότητα CAT μειώνεται από 37% σε 26% καθώς η ποσότητα του συνδιαμολυνμένου ETS-2 πλασμιδίου αυξάνεται από 0 σε 40 pg, μεταβολή στατιστικώς σημαντική ($P < 0.002$, two-sided T test ανεξάρτητων δειγμάτων) στην περίπτωση του πλασμιδίου που φέρει τον

υποκινητή του γονιδίου της ιντερλευκίνης 2. Ενδιάμεσες συγκεντρώσεις πλασμιδίου ETS-2 οδηγούν σε μικρότερα επίπεδα καταστολής τα οποία είναι επίσης στατιστικά σημαντικά (π.χ. για συγκέντρωση 20 pg η μείωση της ενεργότητας από 37% σε 29% έχει τιμή πιθανότητας $P < 0.008$). Στην περίπτωση του πλασμιδίου 4PU η καταστολή της ενεργότητας CAT είναι στατιστικά σημαντική μόνο για την υψηλότερη συγκέντρωση του πλασμιδίου ETS-2 (μείωση από 42% σε 31% με $P < 0.025$) αν και οι μετρήσεις για το στοιχείο αυτό χαρακτηρίζονταν από μεγαλύτερη μεταβλητότητα γεγονός που μειώνει την ευαισθησία ανίχνευσης σημαντικών ποσοτικών διαφορών.

Εικόνα 41 Μείωση της ενεργότητας CAT οδηγούμενη από το στοιχείο 4PU. Η ενεργότητα CAT (μετρούμενη σε % μετατροπής της χλωραμφενικόλης στις ακετυλιωμένες μορφές της) μειώνεται από 37% (όταν δεν χρησιμοποιείται ETS-2) σε 26% όταν συνδιαμολύνεται με 40 µg πλασμιδίου που φέρει το ETS-2. Οι μετρήσεις εκφράζονται σαν μέση τιμή τριών μετρήσεων $\pm$ δειγματικό σφάλμα της μέσης τιμής

Εικόνα 42 Μείωση της ενεργότητας CAT οδηγούμενη από τον υποκινητή του γονιδίου της ιντερλευκίνης 2. Η ενεργότητα CAT (μετρούμενη σε % μετατροπής της χλωραμφενικόλης στις ακετυλιωμένες μορφές της) μειώνεται από 42% (όταν δεν χρησιμοποιείται ETS-2) σε 31% όταν συνδιαμολύνεται με 40 µg πλασμιδίου που φέρει το ETS-2. Οι μετρήσεις εκφράζονται σαν μέση τιμή τριών μετρήσεων $\pm$ δειγματικό σφάλμα της μέσης τιμής

Η πρωτεΐνη ETS-2 όχι μόνο εκφράζεται στα κύτταρα Jurkat μετά από τη διαμόλυνση τους με το αντίστοιχο πλασμίδιο αλλά και ανταποκρίνεται στην κατάσταση ενεργοποίησης του κυττάρου, αποκλείοντας έτσι την περίπτωση ένα έμμεσο αποτέλεσμα της συνδιαμόλυνσης να οφείλεται για τις παρατηρούμενες μεταβολές της ενεργότητας CAT στα κύτταρα αυτά.

Εικόνα 43 Ανίχνευση πρωτεΐνης ETS-2 μετά από συνδιαμόλυνση κυττάρων Jurkat. Western blot μη διαμολυνμένων εν ηρεμία (J), και ενεργοποιημένων (Ji) Jurkat κυττάρων · η ποσοτικοποίηση της πρωτεΐνης ETS-2 έγινε επίσης και σε ενεργοποιημένα μη διαμολυσμένα (JT), και διαμολυσμένα (JTi) Jurkat κύτταρα. Η αριστερή γέλη αντιστοιχεί στην πρωτεΐνη ETS-2, ενώ η γέλη δεξιά είναι η αποτύπωση κατά Western της tubulin (θετικό κοντρόλ μεθόδου). Παρατηρούμε ότι όχι μόνο το ETS-2 εκφράζεται μετά από την διαμόλυνση των κυττάρων αλλά και ανταποκρίνεται στην κατάσταση ενεργοποίησης των.

7 Συζήτηση και Επίλογος

7.1 Μεθοδολογικό πλαίσιο ανίχνευσης μεταγραφικών καταστολέων

Όπως αναφέρθηκε στην αρχή του κειμένου, σκοπός της παρούσας εργασίας ήταν η πρόταση ενός πλαισίου το οποίο συνδιάζοντας δεδομένα από πειράματα μαζικής αποτύπωσης της γονιδιακής έκφρασης και θέσεων δέσμευσης μεταγραφικών παραγόντων θα μπορούσε να χρησιμοποιηθεί για την ανίχνευση μεταγραφικών καταστολέων που δρουν δεσμευόμενοι σε μια βραχεία αλληλουχία Y που ανευρίσκεται στον υποκινητή ενός ή περισσότερων γονιδίων. Το πλαίσιο αυτό θα έπρεπε να λαμβάνει υπόψη τις 3 βασικές ιδιότητες που πρέπει να έχει ένας υποψήφιος μεταγραφικός παράγοντας για να χαρακτηριστεί σαν καταστολέας: 1) να εκφράζεται διαφορετικά σε μια κυτταρική κατάσταση ως προς άλλες καταστάσεις στις οποίες το γονίδιο(-α) δεν τελεί υπό καταστολή 2) να δεσμεύεται στο *cis*-στοιχείο μέσω του οποίου γνωρίζουμε (ή υποθέτουμε) ότι διαμεσολαβείται η καταστολή 3) η δέσμευση να οδηγεί σε μεταγραφική καταστολή σε συστήματα όπου η εκφράζουμε με τεχνητό τρόπο τόσο την πρωτεΐνη του υποτιθέμενου καταστολέα, όσο το στοιχείο στο οποίο δεσμεύεται. Πιστεύουμε ότι καταφέραμε να προσδιορίσουμε τα χαρακτηριστικά που πρέπει να έχει ένα τέτοιο πλαίσιο, τόσο σε ποιοτικό όσο και σε ποσοτικό επίπεδο. Επιπλέον με την εφαρμογή του πλαισίου αυτού σε ένα πραγματικό πρόβλημα βασικής ανοσολογικής έρευνας, διερευνήσαμε λύσεις σε συγκεκριμένα τεχνικά ζητήματα που έχουν πολύ γενικότερες προεκτάσεις από την συγκεκριμένη εφαρμογή. Κομβικό ρόλο στην ανάπτυξη της προτεινόμενης μεθοδολογίας, έπαιξε η υιοθέτηση της Bayesian θεωρίας πιθανοτήτων, ενός συνεκτικού πλαισίου για την ποσοτικοποίηση της «αλήθειας» σε καταστάσεις αβεβαιότητας. Εντός του πλαισίου της Bayesian θεωρίας πιθανοτήτων, οι τελευταίες δεν είναι τίποτα άλλο από *μέτρα* της βεβαιότητας του χρήστη της θεωρίας ότι ένα συγκεκριμένο αβέβαιο ενδεχόμενο είναι αληθές. Έτσι για την λύση πρακτικών προβλημάτων, η Bayesian θεωρία απαιτεί πρώτα να εντοπισθούν τα αβέβαια ενδεχόμενα που παίζουν καθοριστικό ρόλο στην επίλυση του προβλήματος και στη συνέχεια να δοθούν τα αντίστοιχα *ποσοτικά* μέτρα πεποίθησης.

Πως όμως μπορεί να εφαρμοσθεί η Bayesian θεωρία για να κατασκευασθεί ένα μεθοδολογικό πλαίσιο ανίχνευσης μεταγραφικών καταστολέων; Με άλλα λόγια ποια είναι τα αβέβαια ενδεχόμενα του προβλήματος μας για τα οποία μπορούμε να μιλήσουμε με πιθανοκρατικά μέτρα αβεβαιότητας; Ξεκινώντας από τα αφετηριακά τρία ερωτήματα παρατηρούμε ότι το καθένα από αυτά μπορεί να απαντηθεί μέσω διακριτών πειραματικών διαδικασιών (ποσοτικοποίηση γονιδιακής έκφρασης, πειράματα EMSA, δοκιμασίες συν-διαμόλυνσης), η *έκβαση* των οποίων έχει τα χαρακτηριστικά αβέβαιων ενδεχομένων. Προτού διενεργηθεί το κάθε πείραμα, η έκβαση του δεν μπορεί να είναι γνωστή με βεβαιότητα · με τα δεδομένα του πειράματος στη διάθεση μας, μπορούμε να έχουμε μια πιο σαφή *πεποίθηση* του αν ένας δεδομένος υποψήφιος *εκφράζεται* με διαφορετικό τρόπο, *δεσμεύεται* στην υπό εξέταση ακολουθία και *καταστέλλει* την έκφραση γονιδίων. Εφόσον δεχτούμε να *ποσοτικοποιήσουμε* την πεποίθησή μας για καθένα από αυτά τα αβέβαια ενδεχόμενα (δηλαδή τις εκβάσεις των πειραμάτων) σε μία κλίμακα η οποία σε κάθε ενδεχόμενο αντιστοιχεί έναν αριθμό μεταξύ μηδέν και ένα, δηλαδή ένα μέτρο πιθανότητας, τότε εργαλεία της Bayesian θεωρίας πιθανοτήτων μπορούν να χρησιμοποιηθούν για να σχεδιασθεί ένα ποσοτικό μεθοδολογικό πλαίσιο το οποίο στη συνέχεια εφαρμόζεται για τη λύση του υπό εξέταση προβλήματος.

Δεχόμενοι να αντιμετωπίσουμε τα αποτελέσματα των διαθέσιμων πειραμάτων σαν αβέβαια ενδεχόμενα, τότε η ανίχνευση του μεταγραφικού καταστολέα που δρα δεσμευόμενος σε μια βραχεία αλληλουχία DNA αντιστοιχεί στο βέλτιστο σύνθετο πείραμα ποσοτικοποίησης, δέσμευσης, καταστολής. Για την ταυτοποίηση του βέλτιστου πειράματος απαιτείται ο ορισμός μιας *συνάρτησης ωφελιμότητας ή χρησιμότητας* (utility function) μέσω της οποίας ο σχεδιάζων το πείραμα εκφράζει με ποσοτικό τρόπο τις προτιμήσεις του για τα διάφορα αβέβαια ενδεχόμενα. Έχοντας καθορίσει τόσο τα πιθανοκρατικά μέτρα, όσα και τις αντίστοιχες συναρτήσεις χρησιμότητας το βέλτιστο πείραμα είναι αυτό που μεγιστοποιεί την *αναμενόμενη ωφελιμότητα*, ποσότητα που ορίζεται σαν την αναμενόμενη τιμή της χρησιμότητας ως προς τα πιθανοκρατικά μέτρα που εμφανίζονται στη δένδροειδή αναπαράσταση του πειράματος. Η έννοια της προτίμησης (*preference*) είναι καθαρά υποκειμενική και γενικώς θα είναι διαφορετική μεταξύ των διαφόρων χρηστών του ίδιου πειράματος, αλλά και μεταξύ διαφορετικών *χρήσεων* του ίδιου πειράματος. Το χαρακτηριστικό αυτό της Bayesian προσέγγισης με κανένα τρόπο δεν μπορεί να εκληφθεί σαν αδυναμία · αντίθετα αποτελεί την δύναμη της θεωρίας καθώς με την υιοθέτηση διαφορετικών συναρτήσεων χρησιμότητας μπορούμε να προσαρμόσουμε τα ίδια πειράματα σε διαφορετικές χρήσεις. Για τους σκοπούς της παρούσας εργασίας, υιοθετήθηκε μια πολύ απλή συνάρτηση χρησιμότητας που έδινε τη μέγιστη τιμή (ίση με ένα) στο αβέβαιο ενδεχόμενο του σύνθετου πειράματος κατά το οποίο και τα τρία αβέβαια ενδεχόμενα (διαφορικής έκφρασης, δέσμευσης και καταστολής) είχαν καταφατική απάντηση, ενώ έδινε μηδενική αξία αν έστω και ένα από αυτά τα ενδεχόμενα είχε αρνητική απάντηση. Σε σημειολογικό επίπεδο ο μαθηματικός αυτός φορμαλισμός εκφράζει την *απόλυτη* προτίμηση μας για την διεξαγωγή πειραμάτων που απαντούν καταφατικά και στα τρία ερωτήματα και την *απόλυτη* έλλειψη προτίμησης σε πειράματα (και άρα μεταγραφικούς παράγοντες) για τους οποίους έστω και ένα από τα βασικά ερωτήματα επιδέχεται αρνητικής απάντησης. Για παράδειγμα έστω ένας παράγοντας που υποεκφράζεται σε δύο καταστάσεις, και ο οποίος όταν συν-διαμολύνεται με ένα πλασμίδιο που φέρει ένζυμο ανταποκριτή υπό το μεταγραφικό έλεγχο του στοιχείου Y οδηγεί σε καταστολή. Αν στο πείραμα EMSA δεν καταφέρουμε να δείξουμε άμεση φυσική αλληλεπίδραση μεταξύ του στοιχείου Y και του μεταγραφικού παράγοντα τότε η παρατηρούμενη καταστολή κατά πάσα πιθανότητα δεν διαμεσολαβείται μέσω άμεσου μηχανισμού δέσμευσης του παράγοντα στο στοιχείο Y. Αλλά αν ο αρχικός σκοπός μας ήταν να βρούμε παράγοντες που δρουν με αυτό τον άμεσο τρόπο, τότε έχουμε αποτύχει και το πείραμα είχε ελάχιστη (στην ουσία μηδενική) χρησιμότητα για το σκοπό μας. Αν ο αρχικός σκοπός ήταν να βρούμε έναν άμεσο είτε έμμεσο καταστολέα (π.χ. που δρα αυξάνοντας την έκφραση ενός δεύτερου μεταγραφικού παράγοντα που αλληλεπιδρά με το στοιχείο Y) τότε το πείραμα ήταν επιτυχές, άρα απαιτείται μια διαφορετική συνάρτηση χρησιμότητας.

Με την χρήση αυτής της συνάρτησης χρησιμότητας προκύπτει ότι το βέλτιστο πείραμα (το οποίο προτείνει και το μεταγραφικό καταστολέα) είναι αυτό που μεγιστοποιεί την πιθανότητα των τριών ενδεχομένων. Χωρίς επιπλέον παραδοχές το αποτέλεσμα αυτό δεν έχει μεγάλη πρακτική αξία, γιατί απαιτεί την διενέργεια των 3 πειραμάτων για ΟΛΟΥΣ τους μεταγραφικούς παράγοντες, διαδικασία εξαιρετικά χρονοβόρα και μεγάλου κόστους. Αν και η διενέργεια του πειράματος ποσοτικοποίησης της (διαφορικής) έκφρασης των μεταγραφικών παραγόντων με τη βοήθεια κάποιας μαζικής τεχνικής (π.χ. μικροσυστοιχείες, σειριακή ανάλυση γονιδιακής έκφρασης) μπορεί να παράσχει δεδομένα για τον υπολογισμό των πιθανοτήτων υποέκφρασης ταυτόχρονα για όλους τους παράγοντες, τα πειράματα

δέσμευσης και καταστολής πρέπει να γίνουν για καθέναν από τους παράγοντες χωριστά, εξαιρετικά μη αποδοτική προσέγγιση. Ένας απλός μηχανιστικός συλλογισμός για τους παράγοντες από τους οποίους εξαρτάται το ενδεχόμενο υπερμετακίνησης, και καταστολής υποδεικνύει, ότι τα δύο αυτά ενδεχόμενα (και άρα η πιθανότητα τους) πρέπει να εξαρτάται από την πιθανότητα δέσμευσης του μεταγραφικού παράγοντα στο στοιχείο Y. Η τελευταία είναι υπολογίσιμη εξετάζοντας την ομοιότητα του Y με γνωστές (πειραματικά επιβεβαιωμένες) αλληλουχίες στις οποίες δεσμεύεται ο μεταγραφικός παράγοντας. Ο ευρετικός τρόπος με τον οποίο αξιοποιήθηκε αυτή η παρατήρηση βασίστηκε σε δύο στοιχεία:

1. η εξάρτηση αυτή πρέπει να εκφράζεται σε μαθηματικούς όρους με μια μονοτονικώς αύξουσα συνάρτηση
2. στην ανάπτυξη κατά Taylor της μαθηματικής μορφής που παίρνει η εξάρτηση αυτή και στην αντικατάσταση της από την προσέγγιση πρώτης τάξης

Τότε, η μεγιστοποίηση της αναμενόμενης χρησιμότητας δίνεται από τον παράγοντα εκείνο με τη μεγαλύτερο γινόμενο πιθανότητας υποέκφρασης επί πιθανότητας δέσμευσης στο στοιχείο Y. Οι δύο αυτές ποσότητες είναι δυνητικά υπολογίσιμες από βάσεις δεδομένων γονιδιακής έκφρασης και θέσεων δέσμευσης μεταγραφικών παραγόντων (π.χ. TRANSFAC), παρακάμπτοντας την ανάγκη εκτέλεσης όλων των προαναφερθέντων πειραμάτων. Σε μαθηματικό επίπεδο, το πρόβλημα μετασχηματίζεται στον υπολογισμό δύο (ανεξάρτητων) πιθανοτήτων, κάθε μία από τις οποίες αποτελεί ξεχωριστό πρόβλημα στο χώρο της βιοπληροφορικής (διαφορική έκφραση γονιδίων με μαζικές τεχνικές γονιδιακής αποτύπωσης και ανίχνευση βραχέων μοτίβων σε αλληλουχίες DNA).

Στο σημείο αυτό πρέπει να τονισθεί ότι ο τρόπος με τον οποίο γίνεται η συλλογή των δεδομένων που αξιοποιείται για τον υπολογισμό των αντίστοιχων πιθανοτήτων είναι εντελώς αδιάφορος. Για παράδειγμα, η ποσοτικοποίηση της πιθανότητας διαφορικής έκφρασης του μεταγραφικού παράγοντα δεν είναι ανάγκη να γίνει με τη βοήθεια μικροσυστοιχειών · διαφορικές βιβλιοθήκες, σειριακή ανάλυση γονιδιακής έκφρασης, PCR πραγματικού χρόνου θα μπορούσαν εξίσου καλά να παράσχουν τα αντίστοιχα ποσοτικά δεδομένα που μορφοποιούν τις πεποιθήσεις υποέκφρασης. Επιπλέον, δεδομένα από ετερογενείς μετρητικές τεχνικές (εφαρμοσμένες βέβαια στα ίδια βιολογικά συστήματα) είναι συγκρίσιμα από τη στιγμή που η ποσότητα ενδιαφέροντος δεν είναι οι πρωτογενείς μετρήσεις αυτές καθε αυτές αλλά η πιθανότητα διαφορικής έκφρασης με βάση τις πρωτογενείς μετρήσεις. Παρά τη διαφορετική φύση (και κλίμακα) των πρωτογενών μετρήσεων, οι αντίστοιχες πεποιθήσεις εκφράζονται στην ίδια κλίμακα, αυτή της συνάρτησης πιθανότητας και πιθανότητες για αβέβαια ενδεχόμενα είναι πάντα συγκρίσιμες μεταξύ τους, άσχετα από την ερμηνεία των αντίστοιχων ενδεχομένων.

Η δυνατότητα χρήσης των πιθανοτήτων με ανταλλάξιμο τρόπο, αν και σημαντικός στη μεγιστοποίηση της χρήσης της διαθέσιμης πληροφορίας συνδυάζοντας δεδομένα έκφρασης από ετερογενείς πηγές, ενέχει τον κίνδυνο εμφάνισης λογικών παραδόξων. Η αποφυγή των τελευταίων απαιτεί την χρήση επιπλέον δικλείδων ασφαλείας που εισαγάγονται στο πλαίσιο μέσω περιορισμών στις μορφές των συναρτήσεων χρησιμότητας. Ο επιπλέον περιορισμός αφορά την ανάγκη αντιμετώπισης με ασύμμετρο τρόπο των δύο προαναφερθέντων πιθανοτήτων, ώστε να δίνεται μεγαλύτερη βαρύτητα στην πιθανότητα υποέκφρασης. Ο περιορισμός αυτός μας κατεύθυνε να αναζητήσουμε τρόπους αυξομείωσης του «βάθους» της αναζήτησης μας σε παράγοντες των οποίων η πιθανότητα διαφορικής υποέκφρασης είναι μεγαλύτερη από ένα κατώφλι, διαδικασία που θυμίζει τον καθορισμό επιπέδων σημαντικότητας (significance levels) στη συχνοτική στατιστική θεωρία. Αντίθετα

όμως με την τελευταία δεν υπάρχει τίποτα *σημαντικό* στα στατιστικά αυτά κατώφλια · υπάρχουν μόνο για να αποφευχθούν ορισμένα παράδοξα και βρίσκονται πλήρως υπό τον έλεγχο του ερευνητή όπως δείξαμε με τα πειράματα δειγματοληψίας. Στα πλαίσια της παρούσας εργασίας μας επέτρεψαν να επικεντρωθούμε στις μία από τις δύο στατιστικές συνιστώσες του προβλήματος, δηλαδή στην πιθανότητα διαφορικής (υπό-)έκφρασης, και το πρόβλημα Behrens Fisher.

7.2 Το πρόβλημα και η κατανομή Behrens Fisher

Όπως αναφέρθηκε στην εισαγωγή του αντίστοιχου κεφαλαίου, ξεκινώντας από τις πρωταρχικές μετρήσεις των επιπέδων mRNA των μεταγραφικών παραγόντων κατασκευάζουμε τις κατανομές πιθανοτήτων για τη διαφορά των επιπέδων σε δύο καταστάσεις. Αυτή είναι η γενική διατύπωση του (γενικευμένου) προβλήματος Behrens Fisher, το οποίο αναφέρεται στη διαφορά των παραμέτρων θέσης μεταξύ δύο κατανομών. Εδώ θα πρέπει να τονισθεί και πάλι ότι το πλαίσιο που προτάθηκε στο πρώτο βήμα αυτής της εργασίας είναι ανεξάρτητο τόσο από τα πειράματα που παρέχουν τα δεδομένα έκφρασης όσο και από τα μοντέλα που χρησιμοποιούνται για τον υπολογισμό των πιθανοτήτων αυτών. Παρόλα αυτά, έστω και για λόγους επίδειξης έπρεπε να υιοθετηθεί ένα τέτοιο μοντέλο, απαιτώντας έτσι την υιοθέτηση κάποιων βασικών παραδοχών.

Η πρώτη και βασική παραδοχή που έγινε στα πλαίσια της παρούσας εργασίας είναι ότι οι πεποιθήσεις για τη διαφορά έκφρασης των επιπέδων mRNA, μπορούν να εκφραστούν με το κανονικό μοντέλο πιθανότητας, αναγάγοντας έτσι το πρόβλημα υπολογισμού των πιθανοτήτων στο κλασσικό πρόβλημα Behrens Fisher. Η παραδοχή της κανονικότητας θεωρείται δικαιολογημένη για τους ακόλουθους λόγους:

1. η κλίμακα μετρήσεων των κανονικοποιημένων μικροσυστοιχειών πρέπει να είναι συμμετρική ως προς το μηδέν
2. ο μικρός αριθμός των δειγμάτων κάνει προβληματική την υιοθέτηση πιο σύνθετων μορφών κατανομών πεποιθήσεων. Ξεκινώντας μόνο με 3 μετρήσεις σε κάθε κατηγορία, κανείς είναι υποχρεωμένος να κάνει ισχυρές παραμετρικές υποθέσεις διαφορετικά οι υπολογισμοί πιθανοτήτων καθίσταται προβληματικές
3. διεξαγωγή στατιστικών δοκιμασιών μη κανονικότητας (το τροποποιημένο κατά Lilliefors στατιστικό Kolmogorov-Smirnov) έδειξαν ότι η πληθώρα των μετρήσεων (80% ακολουθεί την κανονική κατανομή).
4. η κατανομή των *διαφορών* των μετρήσεων (δηλαδή η κατανομή Behrens Fisher) η οποία αποτελεί και το αντικείμενο μέσω των οποίων έγιναν οι υπολογισμοί των πιθανοτήτων υποέκφρασης, είναι εξαιρετικά σταθερή (robust) ακόμα και σε περιπτώσεις που οι υποθέσεις της κανονικότητας παραβιάζονται¹⁰¹
5. Το κανονικό μοντέλο πιθανότητας, αποτελεί τη βάση πιο σύνθετων (και λιγότερο «παραμετρικών») μοντέλων πεποιθήσεων όπως τα μοντέλα μίξεως (mixture models)^{168;169} τα οποία κατασκευάζονται από τη στάθμιση κανονικών μοντέλων. Στην περίπτωση όπου υιοθετούσαν ένα τέτοιο μοντέλο, η μορφή των πεποιθήσεων για τη διαφορά των παραμέτρων θέσης, θα έπαιρνε τη μορφή ενός μοντέλου μίξεως κατανομών Behrens Fisher και τα τεχνικά ζητήματα που θα έπρεπε να λυθούν στην περίπτωση αυτή είναι απλή προέκταση των ζητημάτων που αντιμετωπίσαμε στην παρούσα εργασία

Η δεύτερη ουσιαστική παραδοχή σχετίζεται με την επιλογή του κατωφλίου υποέκφρασης (δηλαδή το όριο C στο ολοκλήρωμα ορισμού της πιθανότητας

υποέκφρασης). Η τιμή αυτή είναι που στην ουσία κατατμεί την πιθανότητα για τα επίπεδα των μεταγραμμάτων σε τρεις περιοχές: την περιοχή υποέκφραση, υπερέκφρασης και αυτή της μη-μεταβολής. Επιλέγοντας σαν τιμή του ορίου αυτού την τιμή μηδέν, ελέγχουμε στην ουσία την υπόθεση (ή ισοδύναμα εκφράζουμε την πεποίθησή μας), ότι ο λόγος των επιπέδων των μεταγραφικών παραγόντων είναι ίσος με ένα. Αυτό προκύπτει άμεσα από το γεγονός ότι οι μετρήσεις μικροσυστοιχειών (που είναι ανάλογες των συγκεντρώσεων των mRNA στην αντίδραση υβριδοποίησης) εκφράζονται σε λογαριθμική κλίμακα. Έτσι διαφορά στην κλίμακα αυτή ίση με μηδέν, αντιστοιχεί σε λόγο (των μη μετασχηματισμένων δεδομένων) ίση με ένα. Επιλέγοντας να ορίσουμε την υποέκφραση σαν οποιαδήποτε μεταβολή από την τιμή συγκεντρώσεων ίση με ένα, είναι ταυτόσημο με μία ελάχιστη συντηρητική προσέγγιση, η οποία θεωρεί κάθε μεταβολή (έστω και μικρή) στα επίπεδα των μεταγραμμάτων σημαντική. Αν κάποιος ήθελε να υιοθετήσει μια περισσότερη συντηρητική προσέγγιση, θα έπρεπε στην τιμή της σταθμισμένης διαφοράς των μεταγραμμάτων δ να προσθέσει την τιμή εκείνη C που εκφράζει το κατώφλι υποεκφρασιμότητας. Αν και οι πιθανότητες που θα προκύψουν είναι διαφορετικές, ο υπολογιστικός μηχανισμός που αναπτύχθηκε δεν αλλάζει. Με δεδομένο το γεγονός ότι αντίστοιχες εργασίες στην περιοχή των μικροσυστοιχειών χρησιμοποιούν σαν τιμή του C το μηδέν ¹⁷⁰ υιοθετήσαμε την προσέγγιση αυτή, γνωρίζοντας ότι επιλογή διαφορετικών τιμών (π.χ. $C = \log 2$) έχει μεγαλύτερες πιθανότητες να δώσει αποτελέσματα περισσότερο συμβατά με τεχνικές Northern (που επιλέγουν σαν κατώφλι του λόγου υπο- ή υπερ-εκφρασιμότητας την τιμή 2 ή 3). Πέρα της διάθεσης να ακολουθήσουμε μια ευρέως παραδεκτή σύμβαση, η επιλογή αυτή υπαγορεύεται και από έναν ακόμα λόγο: την παντελή απουσία δεδομένων για το ποιες διαφορές επιπέδων mRNA έχουν βιολογική σημασία. Εφόσον αυτή η πληροφορία αυτή δεν είναι διαθέσιμη, η αντιστοίχιση στατιστικής και βιολογικής σημαντικότητας είναι προτιμότερο να γίνεται με τη λιγότερο συντηρητικό τρόπο: το τίμημα βέβαια είναι η αύξηση του αριθμού των «ψευδών» ανακαλύψεων που επιτάσσει την ανάγκη εκτέλεσης εξουθενωτικών επιβεβαιωτικών πειραμάτων.

Η μείωση του αριθμού των «ψευδών» ανακαλύψεων μπορεί να γίνει και πριν από το στάδιο της ανάλυσης με την επιλογή του κατάλληλου κατωφλίου πιθανότητας υποέκφρασης C_p . Εντός του πλαισίου που παρουσιάστηκε στην αρχή της διατριβής, ο ρόλος αυτής της σταθεράς είναι να εισαγάγει μια ασυμμετρία στην διαδικασία αναζήτησης μεταγραφικών παραγόντων, ώστε η πιθανότητα υποέκφρασης να επικυριαρχεί της πιθανότητας δέσμευσης στο υπό εξέταση μοτίβο. Από μια διαφορετική σκοπιά όμως, ο ορισμός ενός κατωφλίου υποέκφρασης, ορίζει ένα διχοτομικό κανόνα ο οποίος μπορεί να χρησιμοποιηθεί για να εμπλουτίσει το σύνολο των υποψηφίων προς ανάλυση σε «αληθώς» υποεκφρασμένους μεταγραφικούς παράγοντες. Σε κάθε σύγκριση αποτύπωσης γονιδιακής έκφρασης μεταξύ δύο καταστάσεων, είναι γενικώς παραδεχτό ότι θα υπάρχουν γονίδια τα οποία θα υποεκφράζονται, και άλλα που δεν υποεκφράζονται (είτε θα υπερεκφράζονται είτε η έκφραση τους θα μένει σταθερή). Εφαρμόζοντας μια διχοτομική δοκιμασία σε ένα τέτοιο σύνολο δεδομένων έχουμε τα ενδεχόμενα του ακόλουθου πίνακα (Πίνακας 7). Αν σε μια δεδομένη σύγκριση, το ποσοστό των υποεκφρασμένων γονιδίων είναι f , τότε πριν τον υπολογισμό της πιθανότητας υποέκφρασης καθένα από τα γονίδια έχει λόγο πιθανοτήτων $f / (1-f)$ να είναι υποεκφρασμένο. Γνωρίζοντας ότι το συγκεκριμένο γονίδιο έχει πιθανότητα υποέκφρασης (όπως αυτή δίνεται από τη κατανομή Behrens Fisher) $> C_p$ τότε η πεποίθησή μας για το αν είναι αληθώς

υποεκφρασμένο δίνεται από τη σχέση: $\frac{f}{1-f} \times \frac{LR^+(C_p)}{LR^-(C_p)}$, όπου LR^+ ο λόγος πιθανοφάνειας της θετικής δοκιμασίας και LR^- ο λόγος πιθανοφάνειας (likelihood ratio) της αρνητικής δοκιμασίας.

Γονίδιο				
Πιθανότητα υποέκφρασης > C_p		Υποεκφρασμένο	Μη υποεκφρασμένο	
	Ναι	AΘ Αληθώς θετική	ΨΘ Ψευδώς θετική	$\Theta = A\Theta + \Psi\Theta$
	Όχι	ΨΑ Ψευδώς αρνητική	ΑΑ Αληθώς αρνητική	$A = A\Lambda + \Psi\Lambda$
		$Y = A\Theta + \Psi\Lambda$	$MY = \Psi\Theta + A\Lambda$	

Πίνακας 7 Πιθανά ενδεχόμενα εφαρμογής της διχοτομικής δοκιμασίας (Πιθανότητα υποέκφρασης > C_p) σε ένα σύνολο μετρήσεων μικροσυστοιχειών. Ο κανόνας αυτός για κάθε γονίδιο έχει 4 πιθανά ενδεχόμενα ανάλογα με το αν το γονίδιο είναι υποεκφρασμένο ή όχι. Η ευαισθησία του κανόνα δίνεται από το λόγο : $A\Theta/Y$, και η ειδικότητα από τη σχέση $A\Lambda/A$. Ο λόγος πιθανοφάνειας μιας θετικής δοκιμασίας δίνεται από το λόγο Ευαισθησίας/(1-Ειδικότητα), ενώ ο λόγος πιθανοφάνειας μιας αρνητικής δοκιμασίας δίνεται από τη σχέση: $(1-\text{Ευαισθησία})/\text{Ειδικότητα}$

Επιλέγοντας σαν τιμή του C_p τη μικρότερη τιμή για την οποία ο λόγος πιθανοφάνειας (LR^+ / LR^-) είναι μόλις μεγαλύτερος από ένα, έχει ως αποτέλεσμα το σύνολο των παραγόντων που προκύπτει μετά από τη εφαρμογή του διχοτομικού κανόνα να έχει μεγαλύτερο ποσοστό αληθώς υποεκφρασμένων γονιδίων από το αρχικό σύνολο. Με βάση τα παραπάνω, η ανάλυση των καμπυλών ROC που παρουσιάστηκε στο κεφάλαιο της ανάλυσης των δεδομένων των μικροσυστοιχειών δεν είναι παρά ένας αριθμητικός τρόπος για τον υπολογισμό του λόγου πιθανοφάνειας σαν συνάρτηση του κατωφλίου πιθανότητας υποέκφρασης · η παράγωγος της καμπύλης ROC σε κάθε σημείο είναι ίση με την τιμή του λόγου πιθανοφάνειας. Το σημείο στο οποίο η παράγωγος αυτή ήταν λίγο μεγαλύτερο από ένα, επιλέχθηκε σαν το κατώφλι για να φιλτράρει τους υπό εξέταση υποψηφίους, με τη βοήθεια «τεχνητών» συνόλων μετρήσεων για την εκτίμηση της ειδικότητας (δείγματα καναλιού αναφοράς) και ευαισθησίας (δείγματα πειραματικού καναλιού).

Η εξερεύνηση εναλλακτικών μαθηματικών φορμαλισμών και αριθμητικών μεθόδων για τον βέλτιστο υπολογισμό της πιθανότητας υποέκφρασης με βάση την κατανομή Behrens Fisher προέκυψε από την ανάγκη ανάλυσης των καμπυλών ROC που κατασκευάστηκαν με βάση «τεχνητά» πειράματα διαφορικής έκφρασης. Καθώς ο υπολογισμός 12000²³ πιθανοτήτων υποέκφρασης της κατανομής Behrens Fisher εντός του *Mathematica* βασίζόμενοι στην μορφή της συνάρτησης πιθανότητας της κατανομής Behrens Fisher απαιτεί περίπου 30 λεπτά²⁴ πραγματικού χρόνου, η ανάλυση των καμπυλών ROC (που βασίστηκαν σε 15000 διαφορικές συγκρίσεις, κάθε μία από τις οποίες περιλαμβάνει μετρήσεις για 12000 γονίδια) θα ήθελε λίγο παραπάνω από 10 μήνες να ολοκληρωθεί. Η υιοθέτηση μία εκ των εκφράσεων 1 ή 2, μειώνει το χρόνο αυτό στον 1 ½ μήνα, ενώ τα προγράμματα που αναπτύχθηκαν στη γλώσσα προγραμματισμού C μπορούν να κάνουν τους αντίστοιχους υπολογισμούς σε λιγότερο από 12 ώρες.

²³ Μέσος αριθμός γονιδίων με ανιχνεύσιμα σήματα μετά την προ-επεξεργασία των πειραμάτων μικροσυστοιχειών

²⁴ Pentium 4 με ταχύτητα χρονισμού 3.2 GHz και 2 GB μνήμη RAM , Mathematica 4.2, Windows XP Pro

7.3 Προεκτάσεις των βιολογικών ευρημάτων

Όπως αναφέρθηκε στην αρχή του κειμένου, το έναυσμα για την εργασία αυτή ήταν η αποκάλυψη ενός (σημαντικού) μηχανισμού μεταγραφικού ελέγχου της ιντερλευκίνης -2 και του υποκινητή του ιού HIV. Μέσω των στατιστικών μεθόδων που εκτέθηκαν στα προηγούμενα κεφάλαια προτάθηκε ότι ο μεταγραφικός παράγοντας ETS-2 παίζει τον κατασταλτικό αυτό ρόλο στα παρθενικά T λεμφοκύτταρα, μια κατηγορία υποκυττάρων σημαντική για την ανοσο-απόκριση. Πως όμως εντάσσεται το εύρημα αυτό σε μια γενικότερη εποπτεία του φυσιολογικού ελέγχου και των διαταραχών της ανοσο-απόκρισης;

Από την πλευρά της φυσιολογίας, η οικογένεια ets εμπλέκεται στη διαφορική ρύθμιση μιας μεγάλης ποικιλίας γονιδίων του ανοσοποιητικού. Φυλογενετικά είναι μια αρχαία οικογένεια μεταγραφικών παραγόντων και δεκάδες μέλη της οικογένειας από οργανισμούς που καλύπτουν όλο το φάσμα της εξέλιξης έχουν απομονωθεί¹⁷¹. Όλα τα μέλη της οικογένειας μοιράζονται ένα κοινό πρωτεϊνικό μοτίβο, την περιοχή (domain) ets¹⁷²· το αντιπροσωπευτικό μέλος της οικογένειας παράγοντας ets-1 αλλά και άλλα κρυσταλλογραφικά μελετημένα μέλη της οικογένειας δεσμεύονται σε αλληλουχίες DNA 20 βάσεων, στο κέντρο των οποίων υπάρχει το μοτίβο GGA. Μεταγραφικοί παράγοντες της οικογένειας εκφράζονται κατά τη διάρκεια της οντογένεσης του ανοσοποιητικού· μια πρόσφατη μελέτη της έκφρασης μελών της οικογένειας αποκάλυψε δυναμικά πρότυπα έκφρασης σε όλα τα κύτταρα που ελέγχθησαν¹⁷³. Χαρακτηριστικά των μελών της οικογένειας όπως για παράδειγμα η εκφυλισμένη περιοχή δέσμευσης, το σχετικά μικρό μοριακό βάρος τα καθιστούν ιδανικά για το ρόλο του καταστολέα της μεταγραφής. Πραγματικά σε διάφορα συστήματα¹⁷⁴⁻¹⁸³ έχει παρατηρηθεί ο κατασταλτικός ρόλος που παίζουν στη μεταγραφή, ρόλος που δεν περιορίζεται στη λειτουργία του ανοσοποιητικού.

Το ερώτημα που ανακύπτει στο σημείο αυτό σχετίζεται με το πιθανό ρόλο που παίζει η (δυσ-)ρύθμιση της έκφρασης μελών της οικογένειας και συγκεκριμένα του ETS-2 σε παθολογικές καταστάσεις. Μια σημαντική παρατήρηση έρχεται από τη μελέτη αυτοάνοσων φαινομένων χρησιμοποιώντας εκχυλίσματα ασθενών με Ιδιοπαθή Θρομβοπενική Πορφύρα (Immune Thrombocytopenic Purpura – ITP)¹⁸⁴, νόσημα που χαρακτηρίζεται από την ανοσολογικά διαμεσολαβούμενη μείωση των αιμοπεταλίων στο περιφερικό αίμα. Για τη *in vivo* διερεύνηση πιθανών διαταραχών στον έλεγχο της ανοσοαπόκρισης χρησιμοποιήθηκαν ιχνηθέτες με την αλληλουχία AAGGAG (αλληλουχία που βρίσκεται στον υποκινητή της IL-2) σε πειράματα EMSA με εκχυλίσματα λεμφοκυττάρων ασθενών τόσο στην ενεργή φάση του νοσήματος, όσο και κατά τη διάρκεια ύφεσης. Στα πειράματα αυτά, σύμπλοκα πρωτεΐνης-ιχνηθέτη ήταν εμφανή σε εκχυλίσματα από φυσιολογικά άτομα, αλλά δεν ανευρίσκονταν σε ασθενείς με ITP. Το γεγονός ότι η πρωτεΐνη – καταστολέας δεν ανευρίσκεται ούτε κατά τη φάση ενεργότητας της νόσου, ούτε και κατά την ύφεση της υποδηλώνει ότι αυτή η βιοχημική προηγείται του νοσήματος και δεν είναι επιφανόμενο. Πιθανολογείται ότι κατά τη διάρκεια της θυμικής ωρίμανσης των λεμφοκυττάρων η αποτυχία της έκφρασης του καταστολέα από ένα πληθυσμό ώριμων κατά τα άλλα λεμφοκυττάρων, θα μπορούσε να δώσει γένεση σε αυξημένο αριθμό παρθενικών T κυττάρων με προδιάθεση για ανάπτυξη αυτοανοσίας.

Η ανακάλυψη της ταυτότητας του άγνωστου καταστολέα του υποκινητή του ιού HIV στα παρθενικά T κύτταρα, και η ταυτοποίηση του σαν το ίδιο μόριο που καταστέλλει την έκφραση του υποκινητή της ιντερλευκίνης - 2 , θέτει ορισμένα ενδιαφέροντα ερωτήματα για την ανοσοπαθολογία της νόσου του AIDS. Από τη μία, η ύπαρξη του καταστολέα, μπορεί να εξηγήσει την παρατήρηση γιατί το ιός ενώ μολύνει τα

παρθενικά λεμφοκύτταρα, δεν εκφράζεται σε αυτά. Όντας μεταγραφικά σιωπηλός ο ιός δεν εκφράζει τα ένζυμα που είναι απαραίτητα για το πολλαπλασιασμό του και άρα δεν είναι ευάλωτος στα αντι-ρετροϊκά φάρμακα, που στοχεύουν τις διάφορες φάσεις του πολλαπλασιασμού του ιού. Η άμεση συνέπεια του παραπάνω είναι ότι τα παρθενικά κύτταρα αποτελούν μία από τις δεξαμενές του ιού που αιτιολογικά ευθύνονται για την υποκλινική φάση της νόσου και την αντοχή του ιού στη μακροχρόνια θεραπεία με διαθέσιμα κοκτέιλ φαρμάκων.

Κατά τη διάρκεια της λοίμωξης με τον ιό HIV ο ανταγωνισμός μεταξύ του ιϊκού και του ενδογενούς στοιχείου του υποκινητή της ιντερλευκίνης-2 για τη δέσμευση της πρωτεΐνης του καταστολέα θα μπορούσε να εξηγήσει μια σειρά κλινικών παρατηρήσεων σχετικά με τη φυσική πορεία της νόσου και συγκεκριμένα την ποικιλία αυτοάνοσων φαινομένων¹⁸⁵⁻¹⁸⁷ αλλά και την επίδραση άλλων λοιμώξεων στην κινητική του ιού. Ένα πιθανό μοντέλο είναι το ακόλουθο: Λόγω της δυνατότητας δέσμευσης του καταστολέα στον υποκινητή του HIV, τα μόρια που διαθέσιμα για την καταστολή των ενδογενών στόχων του μειώνονται με αποτέλεσμα την ενεργοποίηση των κυττάρων με λιγότερο ισχυρά αντιγονικά ερεθίσματα (π.χ. αυτο-αντισώματα). Αφού τα κύτταρα διεγερθούν η άρση της καταστολής και η παραγωγή κυττοκινών και ιντερλευκίνης-2 αλλάζουν το ρεπερτόριο των εκφραζόμενων κυτταρικών παραγόντων που με τη σειρά τους διεγείρουν τη μεταγραφή των γονιδίων του ιού. Αντίστοιχο ρόλο, μπορούν να παίξουν εξωγενή αντιγόνα (μυκοβακτήρια, βακτήρια, άλλοι ιοί, παράσιτα) φαινόμενο με σαφείς κλινικές προεκτάσεις. Και στις δύο περιπτώσεις, η αρχική ενεργοποίηση των παρθενικών T κυττάρων ακολουθείται από την πυροδότηση αντιδράσεων των εκτελεστικών (effector) λειτουργιών του ανοσοποιητικού με αποτέλεσμα την καταστροφή τόσο των αυτο-αντιγόνων αλλά και των T κυττάρων (που εκφράζουν στην επιφάνεια τους ιϊκά αντιγόνα). Έτσι η αρχική φάση της ενεργοποίησης ακολουθείται από τη «διαγραφή» των κυττάρων που ανταποκρίθηκαν, οδηγώντας έτσι στη δραστική μείωση του ρεπερτορίου των T κυττάρων και την εμφάνιση ευκαιριακών λοιμώξεων.

8 Βιβλιογραφία

1. Bishop JO, Morton JG, Rosbash M, Richardson M. Three abundance classes in HeLa cell messenger RNA. *Nature* 1974; **250**: 199-204
2. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 1995; **270**: 467-470
3. Mouzaki A, Rungger D, Tucci A, Doucet A, Zubler RH. Occurrence of a silencer of the interleukin-2 gene in naive but not in memory resting T helper lymphocytes. *Eur.J.Immunol.* 1993; **23**: 1469-1474
4. Mouzaki A, Rungger D. Properties of transcription factors regulating interleukin-2 gene transcription through the NFAT binding site in untreated or drug-treated naive and memory T-helper cells. *Blood* 1994; **84**: 2612-2621
5. Alizadeh A, Eisen M, Davis RE *et al.* The lymphochip: a specialized cDNA microarray for the genomic-scale analysis of gene expression in normal and malignant lymphocytes. *Cold Spring Harb.Symp.Quant.Biol.* 1999; **64**: 71-78
6. Alizadeh AA, Eisen MB, Davis RE *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 2000; **403**: 503-511
7. Valitutti S, Muller S, Cella M, Padovan E, Lanzavecchia A. Serial triggering of many T-cell receptors by a few peptide-MHC complexes. *Nature* 1995; **375**: 148-151
8. Rogers PR, Huston G, Swain SL. High antigen density and IL-2 are required for generation of CD4 effectors secreting Th1 rather than Th0 cytokines. *J.Immunol.* 1998; **161**: 3844-3852
9. Iezzi G, Karjalainen K, Lanzavecchia A. The duration of antigenic stimulation determines the fate of naive and effector T cells. *Immunity.* 1998; **8**: 89-95
10. Crabtree GR, Clipstone NA. Signal transmission between the plasma membrane and nucleus of T lymphocytes. *Annu Rev Biochem* 1994; **63**: 1045-1083
11. Loh C, Shaw KT, Carew J *et al.* Calcineurin binds the transcription factor NFAT1 and reversibly regulates its activity. *J Biol Chem* 1996; **271**: 10884-10891
12. Rao A. NFATp, a cyclosporin-sensitive transcription factor implicated in cytokine gene induction. *J.Leukoc.Biol.* 1995; **57**: 536-542
13. Rao A. NF-ATp: a transcription factor required for the co-ordinate induction of several cytokine genes. *Immunol.Today* 1994; **15**: 274-281

14. Chuvpilo S, Avots A, Berberich-Siebelt F *et al.* Multiple NF-ATc isoforms with individual transcriptional properties are synthesized in T lymphocytes. *J.Immunol.* 1999; **162**: 7294-7301
15. Chuvpilo S, Schomberg C, Gerwig R *et al.* Multiple closely-linked NFAT/octamer and HMG I(Y) binding sites are part of the interleukin-4 promoter. *Nucleic Acids Res.* 1993; **21**: 5694-5704
16. Klein-Hessling S, Schneider G, Heinfling A, Chuvpilo S, Serfling E. HMG I(Y) interferes with the DNA binding of NF-AT factors and the induction of the interleukin 4 promoter in T cells. *Proc.Natl.Acad.Sci.U.S.A* 1996; **93**: 15311-15316
17. Szabo SJ, Gold JS, Murphy TL, Murphy KM. Identification of cis-acting regulatory elements controlling interleukin-4 gene expression in T cells: roles for NF-Y and NF-ATc. *Mol.Cell Biol.* 1993; **13**: 4793-4805
18. Rooney JW, Sun YL, Glimcher LH, Hoey T. Novel NFAT sites that mediate activation of the interleukin-2 promoter in response to T-cell receptor stimulation. *Mol.Cell Biol.* 1995; **15**: 6299-6310
19. Rooney JW, Hoey T, Glimcher LH. Coordinate and cooperative roles for NF-AT and AP-1 in the regulation of the murine IL-4 gene. *Immunity.* 1995; **2**: 473-483
20. Goldfeld AE, Tsai E, Kincaid R *et al.* Calcineurin mediates human tumor necrosis factor alpha gene induction in stimulated T and B cells. *J.Exp.Med.* 1994; **180**: 763-768
21. McCaffrey PG, Goldfeld AE, Rao A. The role of NFATp in cyclosporin A-sensitive tumor necrosis factor-alpha gene transcription. *J.Biol.Chem.* 1994; **269**: 30445-30450
22. Masuda ES, Tokumitsu H, Tsuboi A *et al.* The granulocyte-macrophage colony-stimulating factor promoter cis-acting element CLE0 mediates induction signals in T cells and is recognized by factors related to AP1 and NFAT. *Mol.Cell Biol.* 1993; **13**: 7399-7407
23. Cockerill PN, Bert AG, Jenkins F *et al.* Human granulocyte-macrophage colony-stimulating factor enhancer function is associated with cooperative interactions between AP-1 and NFATp/c. *Mol.Cell Biol.* 1995; **15**: 2071-2079
24. Serfling E, Avots A, Neumann M. The architecture of the interleukin-2 promoter: a reflection of T lymphocyte activation. *Biochim.Biophys Acta* 1995; **1263**: 181-200
25. Loh C, Carew JA, Kim J, Hogan PG, Rao A. T-cell receptor stimulation elicits an early phase of activation and a later phase of deactivation of the transcription factor NFAT1. *Mol Cell Biol* 1996; **16**: 3945-3954

26. Serfling E, Berberich-Siebelt F, Chuvpilo S *et al.* The role of NF-AT transcription factors in T cell activation and differentiation. *Biochim.Biophys.Acta* 2000; **1498**: 1-18
27. Shores EW, Love PE. TCR zeta chain in T cell development and selection. *Curr.Opin.Immunol.* 1997; **9**: 380-389
28. Shores EW, Love PE. TCR zeta chain in T cell development and selection. *Curr.Opin.Immunol.* 1997; **9**: 380-389
29. Avots A, Buttmann M, Chuvpilo S *et al.* CBP/p300 integrates Raf/Rac-signaling pathways in the transcriptional induction of NF-ATc during T cell activation. *Immunity.* 1999; **10**: 515-524
30. Chuvpilo S, Jankevics E, Tyrsin D *et al.* Autoregulation of NFATc1/A expression facilitates effector T cells to escape from rapid apoptosis. *Immunity.* 2002; **16**: 881-895
31. Chuvpilo S, Avots A, Berberich-Siebelt F *et al.* Multiple NF-ATc isoforms with individual transcriptional properties are synthesized in T lymphocytes. *J.Immunol.* 1999; **162**: 7294-7301
32. Marienfeld R, Neumann M, Chuvpilo S *et al.* Cyclosporin A interferes with the inducible degradation of NF-kappa B inhibitors, but not with the processing of p105/NF-kappa B1 in T cells. *Eur.J.Immunol.* 1997; **27**: 1601-1609
33. Neumann M, Wohlleben G, Chuvpilo S *et al.* CD40, but not lipopolysaccharide and anti-IgM stimulation of primary B lymphocytes, leads to a persistent nuclear accumulation of RelB. *J.Immunol.* 1996; **157**: 4862-4869
34. Schneider G, Heinfling A, Klein-Hessling S *et al.* The inducible transcription factor NF-AT plays an important role in the activation of the murine interleukin-4 promoter. *Immunobiology* 1995; **193**: 268-272
35. Serfling E, Avots A, Neumann M. The architecture of the interleukin-2 promoter: a reflection of T lymphocyte activation. *Biochim.Biophys Acta* 1995; **1263**: 181-200
36. Pfeuffer I, Klein-Hessling S, Heinfling A *et al.* Octamer factors exert a dual effect on the IL-2 and IL-4 promoters. *J.Immunol.* 1994; **153**: 5572-5585
37. Chuvpilo S, Avots A, Berberich-Siebelt F *et al.* Multiple NF-ATc isoforms with individual transcriptional properties are synthesized in T lymphocytes. *J.Immunol.* 1999; **162**: 7294-7301
38. Serfling E, Berberich-Siebelt F, Chuvpilo S *et al.* The role of NF-AT transcription factors in T cell activation and differentiation. *Biochim.Biophys.Acta* 2000; **1498**: 1-18

39. Jain J, Loh C, Rao A. Transcriptional regulation of the IL-2 gene. *Curr.Opin.Immunol.* 1995; **7**: 333-342
40. Clements JL, Cooper SM, Budd RC. Abnormal regulation of the IL-2 promoter in lpr CD4-CD8- T lymphocytes results in constitutive expression of a novel nuclear factor of activated T cells-binding factor. *J Immunol* 1995; **154**: 6372-6381
41. Serfling E, Avots A, Neumann M. The architecture of the interleukin-2 promoter: a reflection of T lymphocyte activation. *Biochim.Biophys Acta* 1995; **1263**: 181-200
42. Mouzaki A, Weil R, Muster L, Rungger D. Silencing and trans-activation of the mouse IL-2 gene in *Xenopus* oocytes by proteins from resting and mitogen-induced primary T-lymphocytes. *EMBO J.* 1991; **10**: 1399-1406
43. Serfling E, Avots A, Neumann M. The architecture of the interleukin-2 promoter: a reflection of T lymphocyte activation. *Biochim.Biophys Acta* 1995; **1263**: 181-200
44. Mouzaki A, Rungger D. Properties of transcription factors regulating interleukin-2 gene transcription through the NFAT binding site in untreated or drug-treated naive and memory T-helper cells. *Blood* 1994; **84**: 2612-2621
45. Romano-Spica V, Georgiou P, Suzuki H, Papas TS, Bhat NK. Role of ETS1 in IL-2 gene expression. *J Immunol* 1995; **154**: 2724-2732
46. Mouzaki A, Rungger D. Properties of transcription factors regulating interleukin-2 gene transcription through the NFAT binding site in untreated or drug-treated naive and memory T-helper cells. *Blood* 1994; **84**: 2612-2621
47. Spina CA, Guatelli JC, Richman DD. Establishment of a stable, inducible form of human immunodeficiency virus type 1 DNA in quiescent CD4 lymphocytes in vitro. *J Virol* 1995; **69**: 2977-2988
48. Spina CA, Prince HE, Richman DD. Preferential replication of HIV-1 in the CD45RO memory cell subset of primary CD4 lymphocytes in vitro. *J Clin.Invest* 1997; **99**: 1774-1785
49. Siekevitz M, Josephs SF, Dukovich M *et al.* Activation of the HIV-1 LTR by T cell mitogens and the trans-activator protein of HTLV-I. *Science* 1987; **238**: 1575-1578
50. Yuan R, Bohan C, Shiao FC *et al.* Activation of HIV LTR-directed expression: analysis with pseudorabies virus immediate early gene. *Virology* 1989; **172**: 92-99
51. Rohr O, Marban C, Aunis D, Schaeffer E. Regulation of HIV-1 gene transcription: from lymphocytes to microglial cells. *J Leukoc Biol* 2003; **74**: 736-749

52. Shaw JP, Utz PJ, Durand DB *et al.* Identification of a putative regulator of early T cell activation genes. *Science* 1988; **241**: 202-205
53. Cullen BR. Regulation of HIV-1 gene expression. *FASEB J* 1991; **5**: 2361-2368
54. Demarchi F, D'Agaro P, Falaschi A, Giacca M. In vivo footprinting analysis of constitutive and inducible protein-DNA interactions at the long terminal repeat of human immunodeficiency virus type 1. *J Virol* 1993; **67**: 7450-7460
55. Cullen BR, Malim MH. The HIV-1 Rev protein: prototype of a novel class of eukaryotic post-transcriptional regulators. *Trends Biochem Sci* 1991; **16**: 346-350
56. Cullen BR. Regulation of HIV-1 gene expression. *FASEB J* 1991; **5**: 2361-2368
57. Kinoshita S, Su L, Amano M *et al.* The T cell activation factor NF-ATc positively regulates HIV-1 replication and gene expression in T cells. *Immunity*. 1997; **6**: 235-244
58. Kinoshita S, Chen BK, Kaneshima H, Nolan GP. Host control of HIV-1 parasitism in T cells by the nuclear factor of activated T cells. *Cell* 1998; **95**: 595-604
59. Rosen FS. The acquired immunodeficiency syndrome (AIDS). *J Clin. Invest* 1985; **75**: 1-3
60. Sikder SK, Mitra D, Laurence J. Identification of a novel cell-type and context specific enhancer within the negative regulatory element of the human immunodeficiency virus type 1 long terminal repeat. *Arch.Virol.* 1994; **137**: 139-147
61. Mouzaki A, Doucet A, Mavroidis E, Muster L, Rungger D. A repression-derepression mechanism regulating the transcription of human immunodeficiency virus type 1 in primary T cells. *Mol Med* 2000; **6**: 377-390
62. Copeland KF, Hendrikx PJ, Haaksma AG *et al.* Comparison of the response to T-cell activation by integrated HIV-1 and HTLV-1 LTR-lacZ vectors. *Virology* 1995; **209**: 633-636
63. Mouzaki A, Doucet A, Mavroidis E, Muster L, Rungger D. A repression-derepression mechanism regulating the transcription of human immunodeficiency virus type 1 in primary T cells. *Mol Med* 2000; **6**: 377-390
64. Mouzaki A, Rungger D. Properties of transcription factors regulating interleukin-2 gene transcription through the NFAT binding site in untreated or drug-treated naive and memory T-helper cells. *Blood* 1994; **84**: 2612-2621
65. Papavassiliou AG. Transcription factors: structure, function, and implication in malignant growth. *Anticancer Res* 1995; **15**: 891-894

66. Papavassiliou AG. Molecular medicine. Transcription factors. *N.Engl.J Med* 1995; **332**: 45-47
67. Granger-Schnarr M, Lloubes R, de Murcia G, Schnarr M. Specific protein-DNA complexes: immunodetection of the protein component after gel electrophoresis and Western blotting. *Anal Biochem* 1988; **174**: 235-238
68. Lane D, Prentki P, Chandler M. Use of gel retardation to analyze protein-nucleic acid interactions. *Microbiol.Rev* 1992; **56**: 509-528
69. Dyer RB, Herzog NK. Immunodepletion EMSA: a novel method to identify proteins in a protein-DNA complex. *Nucleic Acids Res* 1995; **23**: 3345-3346
70. Groskreutz D, Schenborn ET. Reporter systems. *Methods Mol Biol* 1997; **63**: 11-30
71. Schenborn E, Groskreutz D. Reporter gene vectors and assays. *Mol Biotechnol* 1999; **13**: 29-44
72. Southern EM. DNA microarrays. History and overview. *Methods Mol.Biol.* 2001; **170**: 1-15
73. Southern E, Mir K, Shchepinov M. Molecular interactions on microarrays. *Nat.Genet.* 1999; **21**: 5-9
74. Chen JJ, Wu R, Yang PC *et al.* Profiling expression patterns and isolating differentially expressed genes by cDNA microarray system with colorimetry detection. *Genomics* 1998; **51**: 313-324
75. Granjeaud S, Bertucci F, Jordan BR. Expression profiling: DNA arrays in many guises. *Bioessays* 1999; **21**: 781-790
76. Bertucci F, Bernard K, Loriod B *et al.* Sensitivity issues in DNA array-based expression measurements and performance of nylon microarrays for small samples. *Hum.Mol.Genet.* 1999; **8**: 1715-1722
77. Bertucci F, Loriod B, Tagett R *et al.* [DNA arrays: technological aspects and applications.]. *Bull.Cancer* 2001; **88**: 243-252
78. Schuchhardt J, Beule D, Malik A *et al.* Normalization strategies for cDNA microarrays. *Nucleic Acids Res.* 2000; **28**: E47
79. Zien A, Aigner T, Zimmer R, Lengauer T. Centralization: a new method for the normalization of gene expression data. *Bioinformatics.* 2001; **17 Suppl 1**: S323-S331
80. Workman C, Jensen LJ, Jarmer H *et al.* A new non-linear normalization method for reducing variability in DNA microarray experiments. *Genome Biol.* 2002; **3**: research0048

81. Kepler TB, Crosby L, Morgan KT. Normalization and analysis of DNA microarray data by self-consistency and local regression. *Genome Biol.* 2002; **3**: RESEARCH0037
82. Colantuoni C, Henry G, Zeger S, Pevsner J. Local mean normalization of microarray element signal intensities across an array surface: quality control and correction of spatially systematic artifacts. *Biotechniques* 2002; **32**: 1316-1320
83. Wang Y, Lu J, Lee R, Gu Z, Clarke R. Iterative normalization of cDNA microarray data. *IEEE Trans.Inf.Technol.Biomed.* 2002; **6**: 29-37
84. Yang YH, Dudoit S, Luu P *et al.* Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.* 2002; **30**: e15
85. Chen YJ, Kodell R, Sistare F *et al.* Normalization methods for analysis of microarray gene-expression data. *J Biopharm.Stat.* 2003; **13**: 57-74
86. Taniguchi M, Miura K, Iwao H, Yamanaka S. Quantitative assessment of DNA microarrays--comparison with Northern blot analyses. *Genomics* 2001; **71**: 34-39
87. Bartosiewicz M, Trounstein M, Barker D, Johnston R, Buckpitt A. Development of a toxicological gene array and quantitative assessment of this technology. *Arch.Biochem.Biophys.* 2000; **376**: 66-73
88. Brazma A, Vilo J. Gene expression data analysis. *FEBS Lett.* 2000; **480**: 17-24
89. Brazma A, Vilo J. Gene expression data analysis. *Microbes.Infect.* 2001; **3**: 823-829
90. Soinov LA, Krestyaninova MA, Brazma A. Towards reconstruction of gene networks from expression data by supervised learning. *Genome Biol.* 2003; **4**: R6
91. Chen D, Toone WM, Mata J *et al.* Global transcriptional responses of fission yeast to environmental stress. *Mol.Biol.Cell* 2003; **14**: 214-229
92. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. Systematic determination of genetic network architecture. *Nat Genet* 1999; **22**: 281-285
93. Beer MA, Tavazoie S. Predicting gene expression from sequence. *Cell* 2004; **117**: 185-198
94. Pritsker M, Liu YC, Beer MA, Tavazoie S. Whole-genome discovery of transcription factor binding sites by network-level conservation. *Genome Res* 2004; **14**: 99-108
95. Loredo T. From Laplace To Supernova SN 1987a: Bayesian Inference In Astrophysics. In: Fougere F, ed. *Maximum Entropy and Bayesian Methods*. Kluwer Academic Publishers, 1990: 81-142

96. Bernardo JM, Smith AFM. *Bayesian theory*. John Wiley and Sons, 1994
97. Lieb JD. Genome-wide mapping of protein-DNA interactions by chromatin immunoprecipitation and DNA microarray hybridization. *Methods Mol.Biol.* 2003; **224**: 99-109
98. Lawrence CE, Altschul SF, Boguski MS *et al.* Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* 1993; **262**: 208-214
99. Hertz GZ, Hartzell GW, III, Stormo GD. Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Comput.Appl.Biosci.* 1990; **6**: 81-92
100. Stormo GD, Hartzell GW, III. Identifying protein-binding sites from unaligned DNA fragments. *Proc Natl Acad Sci U.S.A* 1989; **86**: 1183-1187
101. Robinson GK. Properties of the Student's t and the Behrens-Fisher solution to the two means problem. *The Annals of Statistics* 1976; **4**: 963-971
102. Kim SH, Cohen AS. On the Behrens-Fisher Problem: A Review. *Journal of Educational and Behavioral Statistics* 1998; **23**: 356-377
103. Welch BL. The Generalization of 'Student's' Problem when Several Different Population Variances are Involved. *Biometrika* 1947; **34**: 28-35
104. Scheffe H. On Solutions of the Behrens-Fisher Problem, Based on the t^2 -Distribution. *The Annals of Mathematical Statistics* 1943; **14**: 35-44
105. Scheffe H. A Note on the Behrens-Fisher Problem. *The Annals of Mathematical Statistics* 1944; **15**: 430-432
106. Walsh JE. On the Power Function of the "Best" t^2 -Test Solution of the Behrens-Fisher Problem. *The Annals of Mathematical Statistics* 1949; **20**: 616-618
107. Ray WD, Pitman AENT. An Exact Distribution of the Fisher-Behrens-Welch Statistic for Testing the Difference Between the Means of Two Normal Populations with Unknown Variance. *Journal of the Royal Statistical Society.Series B (Methodological)* 1961; **23**: 377-384
108. Scheffe H. Practical Solutions of the Behrens-Fisher Problem. *Journal of the American Statistical Association* 1970; **65**: 1501-1508
109. Mehta JS, Srinivasan R. On the Behrens-Fisher Problem. *Biometrika* 1970; **57**: 649-655
110. Eaton ML, Olshen RA. Random Quotients and the Behrens-Fisher Problem. *The Annals of Mathematical Statistics* 1972; **43**: 1852-1860
111. Lauer GN, Han CP. Power of Cochran's Test in the Behrens-Fisher Problem. *Technometrics* 1974; **16**: 545-549

112. Ghosh BK. A Two-Stage Procedure for the Behrens-Fisher Problem. *Journal of the American Statistical Association* 1975; **70**: 457-462
 113. Smit CF. On a Distribution-Free Behrens-Fisher Test by Hettmansperger & Malin. *Biometrika* 1980; **67**: 241-242
 114. Dalal SR, Mudholkar GS. A Conservative Test and Confidence Region for Comparing Heteroscedastic Regressions. *Biometrika* 1988; **75**: 149-152
 115. Sprott DA, Farewell VT. The Difference Between Two Normal Means. *American Statistician* 1993; **47**: 126-128
 116. Duong Q, Shorrock R. On Behrens-Fisher solutions. *The Statistician* 1996; **45**: 57-63
 117. Best DJ, Rayner JCW. Welch's Approximate Solution for the Behrens-Fisher Problem. *Technometrics* 1987; **29**: 205-210
 118. Wang YY. Probabilities of the Type I Errors of the Welch Tests for the Behrens-Fisher Problem. *Journal of the American Statistical Association* 1971; **66**: 605-608
 119. Rossi JAD. An Application of Welch's Approximate t-Solution of the Behrens-Fisher Problem to Confidence Intervals. *Technometrics* 1975; **17**: 57-60
 120. Kohr RL, Games PA. Testing Complex a Priori Contrasts on Means from Independent Samples. *Journal of Educational Statistics* 1977; **2**: 207-216
 121. Fenstad GU. A Comparison Between the U and V Tests in the Behrens-Fisher Problem. *Biometrika* 1983; **70**: 300-302
 122. Kim SH, Cohen AS. On the Behrens-Fisher Problem: A Review. *Journal of Educational and Behavioral Statistics* 1998; **23**: 356-377
 123. Robinson GK. Properties of Student's t and of the Behrens-Fisher Solution to the Two Means Problem. *The Annals of Statistics* 1976; **4**: 963-971
 124. Lee AFS, Gurland J. Size and Power of Tests for Equality of Means of Two Normal Populations with Unequal Variances. *Journal of the American Statistical Association* 1975; **70**: 933-941
 125. Ebenhardt, K and Guthie, W. Should $(X_1 - X_2)$ have larger uncertainty than X_1 ? Online Publications, Statistical Engineering Division, National Institutes of Standards and Technology . 2001. 10-5-2003.
- Ref Type: Electronic Citation
126. James GS. The Behrens-Fisher Distribution and Weighted Means. *Journal of the Royal Statistical Society. Series B (Methodological)* 1959; **21**: 73-90

127. Yates F. Fiducial Probability, Recognisable Sub-Sets and Behrens' Test. *Biometrics In Memoriam: Ronald Aylmer Fisher, 1890-1962* 1964; **20**: 343-360
128. Behrens WU. The Comparison of Means of Independent Normal Distributions with Different Variances. *Biometrics* 1964; **20**: 16-27
129. Robinson GK. Properties of Student's t and of the Behrens-Fisher Solution to the Two Means Problem. *The Annals of Statistics* 1976; **4**: 963-971
130. Box G, Tiao G. Inferences concerning the difference between two means. In: Box G, Tiao G, eds. *Bayesian Inference in Statistical Analysis*. Wiley Interscience, 1992: 102-109
131. Patil VH. Approximation to the Behrens-Fisher Distributions. *Biometrika* 1965; **52**: 267-271
132. Patil V. Approximation to the Generalized Behrens-Fisher Distribution Involving Three Variates. *Biometrika* 1969; **56**: 687-689
133. Lee PM. Hypothesis Testing. *Bayesian statistics an introduction*. Arnold, London: 1997: 117-138
134. Ruben H. On the Distribution of the Weighted Difference of Two Independent Student Variables. *Journal of the Royal Statistical Society. Series B (Methodological)* 1960; **22**: 188-194
135. Dickey JM. Expansions of t Densities and Related Complete Integrals. *The Annals of Mathematical Statistics* 1967; **38**: 503-510
136. Dickey JM. Three Multidimensional-integral Identities with Bayesian Applications. *The Annals of Mathematical Statistics* 1968; **39**: 1615-1628
137. Rahman M, Saleh A. Explicit form of the distribution of the Behrens-Fisher d statistic. *J R Statist Soc B* 1974; **36**: 54-60
138. ElGabali MA. Multiple-term approximations for Appell's F_1 function. *Journal of the Australian Mathematical Society Series B-Applied Mathematics* 1997; **39**: 135-148
139. Colavecchia FD, Gasaneo G, Miraglia JE. Numerical evaluation of Appell's F_1 hypergeometric function. *Computer Physics Communications* 2001; **138**: 29-43
140. Chu WC, Srivastava HM. Ordinary and basic bivariate hypergeometric transformations associated with the Appell and Kampé de Fériet functions. *Journal of Computational and Applied Mathematics* 2003; **156**: 355-370
141. Tarasov VF. W. Gordon's integral (1929) and its representations by means of Appell's functions F_2 , F_1 , and F_3 . *Journal of Mathematical Physics* 2003; **44**: 1449-1452

142. Colavecchia FD, Gasaneo G. fl: a code to compute Appell's F-1 hypergeometric function. *Computer Physics Communications* 2004; **157**: 32-38
143. Andrews GE, Askey R, Ranjan R. Special Functions. *The hypergeometric functions*. Cambridge University Press, New York: 1999: 61-113
144. Cuyt A, Driver K, Tan J, Verdonk B. A finite sum representation of the Appell series $F_1(a, b, b'; c; x, y)$. *Journal of Computational and Applied Mathematics* 1999; **105**: 213-219
145. Ismail MEH, Pitman J. Algebraic evaluations of some Euler integrals, duplication formulae for Appell's hypergeometric function F_1 , and Brownian variations. *Canadian Journal of Mathematics-Journal Canadien de Mathematiques* 2000; **52**: 961-981
146. Hutchinson JI. The Hypergeometric Functions of N Variables. *Transactions of the American Mathematical Society* 1909; **10**: 61-70
147. Mori M, Sugihara M. The double-exponential transformation in numerical analysis. *Journal of Computational and Applied Mathematics* 2001; **127**: 287-296
148. Sugihara M. Optimality of the double exponential formula - Functional analysis approach. *Numerische Mathematik* 1997; **75**: 379-395
149. Neuman E, Vanfleet PJ. Moments of Dirichlet Splines and Their Applications to Hypergeometric-Functions. *Journal of Computational and Applied Mathematics* 1994; **53**: 225-241
150. Lijoi A, Regazzini E. Means of a Dirichlet process and multiple hypergeometric functions. *Annals of Probability* 2004; **32**: 1469-1495
151. Medvedovic M, Sivaganesan S. Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics*. 2002; **18**: 1194-1206
152. Burr D, Doss H, Cooke GE, Goldschmidt-Clermont PJ. A meta-analysis of studies on the association of the platelet P1A polymorphism of glycoprotein IIIa and risk of coronary heart disease. *Stat.Med* 2003; **22**: 1741-1760
153. Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. *Journal of Machine Learning Research* 2003; **3**: 993-1022
154. Barnard K, Duygulu P, Forsyth D *et al*. Matching words and pictures. *Journal of Machine Learning Research* 2003; **3**: 1107-1135
155. Laval G, SanCristobal M, Chevalet C. Maximum-likelihood and markov chain monte carlo approaches to estimate inbreeding and effective size from allele frequency changes. *Genetics* 2003; **164**: 1189-1204

156. Muhammad M, Mori M. Double exponential formulas for numerical indefinite integration. *Journal of Computational and Applied Mathematics* 2003; **161**: 431-448
157. Zhang S, Ji J. *Computation of Special Functions*. John Wiley and Sons, 1996
158. Alizadeh A, Eisen M, Botstein D, Brown PO, Staudt LM. Probing lymphocyte biology by genomic-scale gene expression analysis. *J Clin.Immunol.* 1998; **18**: 373-379
159. Alizadeh AA, Eisen MB, Davis RE *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 2000; **403**: 503-511
160. Alizadeh AA, Staudt LM. Genomic-scale gene expression profiling of normal and malignant immune cells. *Curr.Opin.Immunol.* 2000; **12**: 219-225
161. Perdomo J, Holmes M, Chong B, Crossley M. Eos and pegasus, two members of the Ikaros family of proteins with distinct DNA binding activities. *J Biol Chem* 2000; **275**: 38347-38354
162. Tymms MJ, Kola I. Regulation of gene expression by transcription factors Ets-1 and Ets-2. *Mol Reprod.Dev* 1994; **39**: 208-214
163. Baron BW, Anastasi J, Thirman MJ *et al.* The human programmed cell death-2 (PDCD2) gene is a target of BCL6 repression: implications for a role of BCL6 in the down-regulation of apoptosis. *Proc Natl Acad Sci U.S.A* 2002; **99**: 2860-2865
164. Mouzaki A, Rungger D, Tucci A, Doucet A, Zubler RH. Occurrence of a silencer of the interleukin-2 gene in naive but not in memory resting T helper lymphocytes. *Eur.J.Immunol.* 1993; **23**: 1469-1474
165. Mouzaki A, Weil R, Muster L, Rungger D. Silencing and trans-activation of the mouse IL-2 gene in *Xenopus* oocytes by proteins from resting and mitogen-induced primary T-lymphocytes. *EMBO J.* 1991; **10**: 1399-1406
166. Mouzaki A, Rungger D. Properties of transcription factors regulating interleukin-2 gene transcription through the NFAT binding site in untreated or drug-treated naive and memory T-helper cells. *Blood* 1994; **84**: 2612-2621
167. Mouzaki A, Doucet A, Mavroidis E, Muster L, Rungger D. A repression-derepression mechanism regulating the transcription of human immunodeficiency virus type 1 in primary T cells. *Mol Med* 2000; **6**: 377-390
168. Pan W, Lin J, Le CT. How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach. *Genome Biol.* 2002; **3**: research0022
169. Ghosh D, Chinnaiyan AM. Mixture modelling of gene expression data from microarray experiments. *Bioinformatics.* 2002; **18**: 275-286

170. Baldi P, Long AD. A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. *Bioinformatics*. 2001; **17**: 509-519
171. Degnan BM, Degnan SM, Naganuma T, Morse DE. The ets multigene family is conserved throughout the Metazoa. *Nucleic Acids Res*. 1993; **21**: 3479-3484
172. Shore P, Whitmarsh AJ, Bhaskaran R *et al*. Determinants of DNA-binding specificity of ETS-domain transcription factors. *Mol.Cell Biol*. 1996; **16**: 3338-3349
173. Anderson MK, Hernandez-Hoyos G, Diamond RA, Rothenberg EV. Precise developmental regulation of Ets family transcription factors during specification and commitment to the T cell lineage. *Development* 1999; **126**: 3131-3148
174. Lopez RG, Carron C, Oury C *et al*. TEL is a sequence-specific transcriptional repressor. *J Biol Chem* 1999; **274**: 30132-30138
175. Chakrabarti SR, Nucifora G. The leukemia-associated gene TEL encodes a transcription repressor which associates with SMRT and mSin3A. *Biochem Biophys Res Commun* 1999; **264**: 871-877
176. Guidez F, Petrie K, Ford AM *et al*. Recruitment of the nuclear receptor corepressor N-CoR by the TEL moiety of the childhood leukemia-associated TEL-AML1 oncoprotein. *Blood* 2000; **96**: 2557-2561
177. Fenrick R, Wang L, Nip J *et al*. TEL, a putative tumor suppressor, modulates cell growth and cell morphology of ras-transformed cells while repressing the transcription of stromelysin-1. *Mol Cell Biol* 2000; **20**: 5828-5839
178. Wang L, Hiebert SW. TEL contacts multiple co-repressors and specifically associates with histone deacetylase-3. *Oncogene* 2001; **20**: 3716-3725
179. Irvin BJ, Wood LD, Wang L *et al*. TEL, a putative tumor suppressor, induces apoptosis and represses transcription of Bcl-XL. *J Biol Chem* 2003; **278**: 46378-46386
180. Wood LD, Irvin BJ, Nucifora G, Luce KS, Hiebert SW. Small ubiquitin-like modifier conjugation regulates nuclear export of TEL, a putative tumor suppressor. *Proc Natl Acad Sci U.S.A* 2003; **100**: 3257-3262
181. Kawagoe H, Potter M, Ellis J, Grosveld GC. TEL2, an ETS factor expressed in human leukemia, regulates monocytic differentiation of U937 Cells and blocks the inhibitory effect of TEL1 on ras-induced cellular transformation. *Cancer Res* 2004; **64**: 6091-6100
182. Kawagoe H, Potter M, Ellis J, Grosveld GC. TEL2, an ETS factor expressed in human leukemia, regulates monocytic differentiation of U937 Cells and blocks the inhibitory effect of TEL1 on ras-induced cellular transformation. *Cancer Res* 2004; **64**: 6091-6100

183. Dandre F, Owens GK. Platelet-derived growth factor-BB and Ets-1 transcription factor negatively regulate transcription of multiple smooth muscle cell differentiation marker genes. *Am J Physiol Heart Circ.Physiol* 2004; **286**: H2042-H2051
184. Mouzaki A, Theodoropoulou M, Gianakopoulos I *et al*. Expression patterns of Th1 and Th2 cytokine genes in childhood idiopathic thrombocytopenic purpura (ITP) at presentation and their modulation by intravenous immunoglobulin G (IVIg) treatment: their role in prognosis. *Blood* 2002; **100**: 1774-1779
185. Sekigawa I, Ogasawara H, Kaneko H, Hishikawa T, Hashimoto H. Retroviruses and autoimmunity. *Intern.Med* 2001; **40**: 80-86
186. Zandman-Goddard G, Shoenfeld Y. HIV and autoimmunity. *Autoimmun.Rev* 2002; **1**: 329-337
187. Saif MW. HIV-associated autoimmune hemolytic anemia: an update. *AIDS Patient.Care STDS*. 2001; **15**: 217-224