Μεταπτυχιακή Εργασία

Πανεπιστήμιο Ιωαννίνων
Τμήμα Πληροφορικής

ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ ΈΚΦΡΑΣΗΣ ΓΟΝΙΔΙΩΝ ΚΑΙ ΑΝΑΚΑΤΑΣΚΕΥΗ ΓΕΝΕΤΙΚΩΝ ΔΙΚΤΥΩΝ ΜΕ ΧΡΗΣΗ ΓΡΑΦΙΚΩΝ ΜΟΝΤΕΛΩΝ

Γεώργιος Χ. Σακελλάρης

Πανεπιστήμιο Ιωαννίνων

Ιούνιος 2004

Πρόλογος

Η παρούσα εργασία εκπονήθηκε από τον Γεώργιο Χ. Σακελλάρη ως μερική απαίτηση για την απόκτηση του μεταπτυχιακού διπλώματος ειδίκευσης στο τμήμα Πληροφορικής, του Πανεπιστημίου Ιωαννίνων. Σκοπός της παρούσας εργασίας ήταν η ανάλυση των δεδομένων έκφρασης γονιδίων και η ανακατασκευή της δομή των γενετικών δικτύων.

Επιβλέπων Καθηγητής: Δημήτριος Ι. Φωτιάδης
Επιτροπή Καθηγητών: Άριστείδης Λύκας
Κωνσταντίνος Μπλέκας

Πανεπιστήμιο Ιωαννίνων
Τμήμα Πληροφορικής
Ιούνιος 2004

Περίληψη

Σκοπός της παρούσας εργασίας είναι η ανακατασκευή της δομή των γενετικών δικτύων με ανάλυση δεδομένων έκφρασης γονιδίων. Η εργασία βασίζεται στην αναπαράσταση των γονιδιακών δικτύων με την μορφή μαθηματικών μοντέλων τα οποία χρησιμοποιούνται για να εκφράσουν το αίτιο και το αιτιατό σε ένα βιολογικό σύστημα. Τα βασικά μέρη της εργασίας είναι πρώτον η χρησιμοποίηση και η σύγκριση των αποτελεσμάτων τριών αλγορίθμων ομαδοποίησης (Kmean, Fuzzy KMeans, Mixture Models), και η χρήση γραφικών μοντέλων για την εξαγωγή του γενετικού δικτύου. Επίσης χρησιμοποιήθηκε εκ των προτέρων βιολογική γνώση στην διαδικασία ανακατασκευής του γενετικού δικτύου. Τα παραπάνω βήματα παρουσιάζονται με την μορφή μιας πλήρως αυτοματοποιημένης διαδικασίας που παίρνει ως είσοδο ένα microarray σύνολο δεδομένων και έχει ως έξοδο μια πιθανή δομή του γενετικού δικτύου.

Περιεχόμενα

Κεφάλαιο 1: Εισαγωγή

Σκοπός της εργασίας είναι η ανάλυση των δεδομένων έκφρασης γονιδίων και η ανακατασκευή των γενετικών δικτύων. Με τον όρο ανάλυση δεδομένων έκφρασης γονιδίων εννοούμε την εφαρμογή υπολογιστικών μεθόδων με σκοπό την εξαγωγή χρήσιμης πληροφορίας από τα δεδομένα που διαθέτουμε. Η ανάλυση στην παρούσα εργασία περιλαμβάνει τεχνικές προεπεξεργασίας για την αποβολή των δεδομένων που δεν είναι "σημαντικά" ή που περιέχουν θόρυβο, επίσης περιλαμβάνει την εφαρμογή υπολογιστικών μεθόδων για ομαδοποίηση εκείνων των γονιδίων που παρουσιάζουν όμοια συμπεριφορά κατά την διάρκεια των πειραματικών μετρήσεων. Αναλυτικά οι μέθοδοι καθώς και ο σκοπός για τον οποίο εφαρμόζονται παρουσιάζονται στο κεφάλαιο της μεθοδολογίας. Η ανακατασκευή των γενετικών δικτύων επιτυγχάνεται με χρήση γραφικών μοντέλων και σκοπός είναι η εύρεση σχέσεων αλληλεπίδρασης μεταξύ των γονιδίων και των παράγωγών τους από τα δεδομένα έκφρασης.

Στην παρούσα εργασία έχουν χρησιμοποιηθεί δεδομένα έκφρασης γονιδίων, τα τελευταία χρόνια έχουν αναπτυχθεί διάφορες τεχνικές (όπως για παράδειγμα οι τεχνικές μικροσυστοιχιών-microarrays) με σκοπό την καταγραφή των πληροφοριών που παράγονται όταν πραγματοποιούνται διάφορες βιολογικές διεργασίες σε ένα βιολογικό σύστημα. Οι τεχνικές αυτές επιτρέπουν την ανάπτυξη υπολογιστικών μεθόδων που στόχο έχουν την ανάλυση των συνιστωσών του συστήματος και την απόκτηση γνώσης για την λειτουργία των σύνθετων αυτών συστημάτων. Το πρόβλημα της ανακατασκευής των γενετικών δικτύων θεωρείται ένα από τα πιο σημαντικά προβλήματα που ανήκουν στην κατηγορία των "functional genomics", και αυτό γιατί η γνώση για το πως συμπεριφέρονται τα γονίδια μπορούν να μας βοηθήσουν στην ανακάλυψη νέων τρόπων αντιμετώπισης των ασθενειών αλλά και στην κατανόηση της συμπεριφοράς των βιολογικών συστημάτων.

Τα γονίδια σε ένα κύτταρο επηρεάζουν την κυτταρική δομή, την κίνηση, το μεταβολισμό, την ομοιόσταση, την μετάδοση σημάτων, την αναπαραγωγή και την αποκατάσταση. Οι πρωτεΐνες παίζουν σημαντικό ρόλο στον γενετικό έλεγχο: όπου ελέγχεται πιο γονίδιο μεταφράζεται σε πρωτεΐνη κάθε στιγμή. Μέσω αυτών των

γενετικών ρυθμιστικών μηχανισμών, οι πρωτεΐνες είναι υπεύθυνες για τον έλεγχο της ύπαρξής τους. Πολύ λίγα όμως είναι γνωστά για τα σήματα και τους ελέγχους που ενεργοποιούν και καταστέλλουν την έκφραση των συγκεκριμένων γονιδίων. Σε αυτή την εργασία παρουσιάζεται μεθοδολογία ανακατασκευής των γενετικών δικτύων με χρήση γραφικών μοντέλων και δεδομένα έκφρασης γονιδίων. Μελετάται επίσης η καταλληλότητα αυτής της προσέγγισης, και οι περιορισμοί της. Τα δεδομένα χρησιμοποιούνται παράχθηκαν από πειραματικές μετρήσεις στον οργανισμό S. Cerevisiae (γνωστό και ως ζύμη).

Η ανάλυση της έκφρασης των γονιδίων θεωρείται μια από τις πλέον αποτελεσματικές τεχνικές για την απόκτηση γνώσης σε ρυθμιστικά δίκτυα γονιδίων. Η γνώση αυτών των συστημάτων θεωρείται ότι μπορεί να δώσει νέα διάσταση στις τεχνικές παραγωγής φαρμάκων αλλά και αντιμετώπισης ασθενειών γενικότερα. Πρέπει επίσης να επισημανθεί ότι στην μελέτη των βιολογικών συστημάτων και πιο συγκεκριμένα στην μελέτη των γενετικών δικτύων οι πειραματικές μέθοδοι είναι δύσκολο να εφαρμοστούν εξαιτίας του πλήθους των συνιστωσών που υπεισέρχονται. Για παράδειγμα στην μελέτη ενός απλού οργανισμού (yeast) οι συνιστώσες (γονίδια, μακρομόρια, γονιδιακά παράγωγα, κτλ) που πρέπει να αναλυθούν είναι της τάξης των μερικών δεκάδων χιλιάδων.

Στο κεφάλαιο 2, παρουσιάζονται μερικές από τις βασικές αρχές των βιολογικών συστημάτων και των γενετικών δικτύων, στο κεφάλαιο 3 περιγράφονται ορισμένες από τις πιο σύγχρονες τεχνικές μέτρησης έκφρασης γονιδίων. Στο κεφάλαιο 4 αναφέρονται μερικές από τις μεθόδους και τεχνικές που έχουν ήδη προταθεί και αναφέρονται στην ανακατασκευή των γενετικών δικτύων, ενώ στο κεφάλαιο 5 επισημαίνονται οι προσπάθειες που γίνονται για την οργάνωση της βιολογικής γνώσης η οποία παράγεται καθημερινά από εργαστήρια σε όλον τον κόσμο. Η οργάνωση και διαχείριση αυτής της γνώσης επιτρέπει την συνεχή πρόοδο των μελετών πάνω στα βιολογικά συστήματα. Στο κεφάλαιο 6 αναφέρονται οι βασικοί ορισμοί των γραφικών μοντέλων, οι τρόποι με τους οποίους χρησιμοποιούνται για την επίλυση προβλημάτων αλλά και οι αδυναμίες τους. Στο κεφάλαιο 7 περιγράφεται η μέθοδος η οποία προτείνεται για την ανάλυση των δεδομένων έκφρασης και ανακατασκευής των γενετικών δικτύων. Τέλος στα κεφάλαια που ακολουθούν αναφέρονται τα αποτελέσματα της μεθόδου αλλά και τα συμπεράσματα.

Κεφάλαιο 2: Βιολογικά συστήματα

Στο κεφάλαιο αυτό περιγράφονται βασικές αρχές των βιολογικών συστημάτων και των γενετικών δικτύων. Αναλύεται η σημασία της γνώσης αυτών των συστημάτων σε διάφορους τομείς της ιατρικής και στην αντιμετώπιση ασθενειών.

Η βιοπληροφορική σχετίζεται με την μελέτη των βιολογικών συστημάτων με χρήση υπολογιστικών τεχνικών. Αντιπροσωπεύει έναν σχετικά νέο τομέα της πληροφορικής ο οποίος εξελίσσεται όσο παράγονται δεδομένα από τεχνολογίες που σχεδιάζονται για τη μέτρηση βιολογικών συστημάτων. Από την ανακάλυψη των τεχνικών μικροσυστοιχιών το 1995 μέχρι σήμερα υπάρχει μεγάλη ανάπτυξη στην παραγωγή δεδομένων έκφρασης γονιδίων από πολλούς οργανισμούς.

Τα δεδομένα αυτά επέτρεψαν στο να αναπτυχθεί η θεωρία με βάση την οποία γενετικά συστήματα είναι λογικά δίκτυα των οποίων οι κόμβοι αποτελούν τα γονίδια (ή άλλα στοιχεία του βιολογικού συστήματος) τα οποία επιδρούν στα επίπεδα έκφρασης των υπολοίπων γονιδίων. Ένας από τους πρώτους στόχους ήταν η κατασκευή των γενετικών δικτύων χρησιμοποιώντας δεδομένα έκφρασης τα οποία παράγονται από microarrays τεχνικές. Μέθοδοι και τεχνικές που υπάρχουν ήδη στην επιστήμη της πληροφορικής έχουν αποδείξει ότι ο συμπερασμός ρυθμιστικών δικτύων είναι πιθανός χρησιμοποιώντας μόνο σύνολα από δεδομένα έκφρασης γονιδίων. Ένα ανοιχτό ερώτημα είναι, το πόσα δεδομένα είναι απαραίτητα ώστε οι τεχνικές reverse engineering να έχουν ικανοποιητικά αποτελέσματα στην επίλυση αυτών των προβλημάτων. Παράλληλα, οι πειραματικοί βιολόγοι εργάζονται στα ίδια προβλήματα. Ωστόσο ο αριθμός των πειραμάτων που είναι απαραίτητος για την ανακατασκευή των γενετικών δικτύων είναι πολύ μεγάλος και ο χρόνος που χρειάζονται για να πραγματοποιηθούν είναι εκθετικά ανάλογος με τον αριθμό των γονιδίων που συμμετέχουν στο δίκτυο. Για τον λόγο αυτό εφαρμόζονται αλγόριθμοι που έχουν σκοπό την μείωση του αριθμού των γονιδίων. Αυτή η διαδικασία παρέχει μια χονδροειδή μοντελοποίηση των περίπλοκων εξαρτήσεων που υπάρχουν μεταξύ των γονιδίων. Τα δίκτυα που προκύπτουν χρησιμοποιώντας μόνο μέρος των κόμβων περιγράφουν όχι ικανοποιητικά τα πειραματικά αποτελέσματα, ενώ δύσκολα θα μπορούσαν να χρησιμοποιηθούν για πρόβλεψη.

2.1 Γενετικά δίκτυα

Ο κώδικας ακολουθίας του DNA καθορίζει τις βιοχημικές διαδικασίες και την έκφραση των γονιδίων. Τα γονίδια παρουσιάζουν μεταβλητότητα στην έκφρασή τους. Αυτή η διαφοροποίηση στην έκφραση έχει δείξει ότι είναι συνέπεια πολλών παραγόντων, όπως οι περιβαλλοντικές συνθήκες, η παρουσία ή η απουσία εξωτερικών ουσιών καθώς και άλλων γονιδίων. Τα προϊόντα των γονιδίων χρησιμεύουν ως μηχανισμός ώθησης για άλλα γονίδια. Η προκύπτουσα αντίδραση μπορεί να προκαλέσει αλληλεπιδράσεις με συνέπειες στους μεταγραφικούς παράγοντες που επηρεάζουν την έκφραση άλλων γονιδίων, η διαδικασία αυτή διαμορφώνει ένα εκτεταμένο γενετικό δίκτυο αλληλεπιδράσεων [Brazma et al. 1998] [Arnone et al. et al. 1997] [Arlinghaus et al. 1997].

Η δομή των γενετικών δικτύων μπορεί να εξαχθεί με ανάλυση δεδομένων έκφρασης γονιδίων. Η βασική ιδέα είναι να αναπαραστήσουμε τα γονίδια ως μαθηματικά μοντέλα και να εφαρμόσουμε αλγόριθμους μάθησης σε αυτό το μοντέλο, ώστε να μάθει τους κανόνες αλληλεπίδρασης του δικτύου των γονιδίων. Τέτοια μαθηματικά μοντέλα που θα ταίριαζαν τη φύση των γενετικών δικτύων είναι τα Boolean δίκτυα[Akutsu et al. 1999], τα Hidden Markov Models (HMMs), τα Bayesian μοντέλα [Wessels et al. 2001].

Η παραδοσιακή βιολογία εξετάζει τα γονίδια ή τις πρωτεΐνες μεμονωμένα. Αυτό όμως που δίνει σημαντικές πληροφορίες, είναι η αλληλεπίδραση αυτών των στοιχείων. Το λογικό επομένως βήμα είναι να συνδυαστούν τα δεδομένα από διάφορες πηγές για να δημιουργηθεί μια ιεραρχική αποτύπωση των αλληλεπιδράσεων των δεδομένων που συμμετέχουν στα γονιδιακά δίκτυα και στη συνέχεια στα pathways. Τα μαθηματικά μοντέλα είναι πολύ σημαντικά για την ολοκλήρωσης αυτών των πληροφοριών και στην κατανόηση της λειτουργίας των συστημάτων λόγω του μεγάλου όγκου των δεδομένων. Και αυτό γιατί μέσα από μια στοχαστική έρευνα πάνω σε ένα σύστημα, είναι δυνατό να καθοριστεί ένα μέρος του μοντέλου που είτε λείπει είτε δεν είναι καλά κατανοητό. Για παράδειγμα ανακαλύφθηκε μια μονάδα ελέγχου σε ένα γονίδιο αχινών που δεν ήταν γνωστή και παρατηρήθηκε πρόσφατα με τη χρήση ενός μαθηματικού μοντέλου [Yuh, et al. 2001].

Τα" βιολογικά δίκτυα **είναι ένα σύνολο βιοχημικών οντοτήτων (συμπεριλαμβανομένου του RNA αγγελιοφόρων, των πρωτεϊνών, του DNA, των ιόντων, ή άλλων μορίων, όπως οι ορμόνες), τα οποία αλληλεπιδρούν για να παράγουν βιολογικά προϊόντα.** Η ανάλυση αυτών των συστημάτων επιδιώκει να βελτιώσει την γνώση για τις αλληλεπιδράσεις μεταξύ των γονιδίων και των παραγώγων τους, και παρέχει πρόβλεψη για τη γενική συμπεριφορά του συστήματος. Αυτό ονομάζεται συνήθως βιολογία συστημάτων επειδή επιδιώκει να μελετήσει ταυτόχρονα τη σύνθετη αλληλεπίδραση πολλών επιπέδων βιολογικών πληροφοριών.

2.2 Βιολογικές διεργασίες

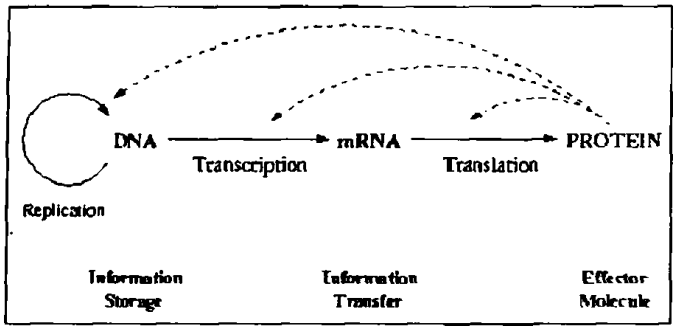
Στις παρακάτω παραγράφους επιχειρείται μια περιληπτική περιγραφή των βιολογικών διεργασιών οι οποίες παίρνουν μέρος σε ένα κύτταρο και αποτελούν μέρος του βιολογικού συστήματος κάθε κυτταρικού οργανισμού.

2.2.1 Γενετικός Έλεγχος

Μέσα σε κάθε οργανισμό υπάρχει αποθηκευτικός χώρος δεδομένων όπου περιέχονται όλες οι πληροφορίες που απαιτούνται για τον κύκλο ζωής του οργανισμού. Αυτές οι πληροφορίες κωδικοποιούνται στο DNA (G, C, A, T), υπό μορφή διαδοχικών blocks πληροφοριών, αποκαλούμενα γονίδια. Κάθε γονίδιο κωδικοποιεί μια συγκεκριμένη πρωτεΐνη που μπορεί να εκφραστεί (expressed) ή να κατασταλεί (repressed) σε περίπτωση ανάγκης. Για να μετατραπεί η ακολουθία DNA σε ένα πρωτεϊνικό μόριο, πρέπει να αντιγραφεί (transcribed) και έπειτα να μετατραπεί (translated) σε μια πρωτεΐνη. Κατά την μεταγραφή ενός γονιδίου παράγεται ένα αντίγραφο RNA αγγελιοφόρων (mRNA), το οποίο μπορεί έπειτα να μεταφραστεί σε πρωτεΐνη. Αυτό επιδρά στο DNA που περιέχει τις πληροφορίες για μια μεγάλη σειρά πρωτεϊνών (effectors molecules), αλλά μόνο εκείνα που εκφράζονται είναι παρόντες ως αντίγραφα mRNA.

Κάθε βήμα της μετατροπής από τις αποθηκευμένες πληροφορίες (DNA), μέσω mRNA (messenger), στην πρωτεϊνική σύνθεση (effectors) καταλύεται από μόρια επίδρασης (effectors molecules). Αυτά τα μόρια μπορούν να είναι ένζυμα, που είναι χημικά ενεργές πρωτεΐνες, ή άλλοι παράγοντες που απαιτούνται από μια διαδικασία

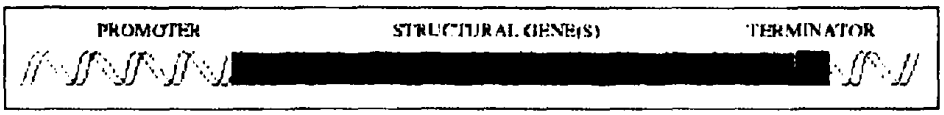
για να συνεχιστεί. Συνεπώς ένας βρόχος διαμορφώνεται όπου τα προϊόντα ενός γονιδίου απαιτούνται για να παραγάγουν τα περαιτέρω προϊόντα γονιδίων.



Εικόνα 2-1 Μετατροπή της αποθηκευμένης πληροφορίας (DNA), μέσω mRNA (messenger), σε πρωτεΐνες

2.2.1.1 Μεταγραφή

Τα γονίδια αποτελούνται από έναν αριθμό ευδιάκριτων περιοχών οι οποίες ελέγχουν το επιθυμητό προϊόν. Αυτές οι περιοχές είναι γενικά της μορφής: promoter - gene(s) - terminator, όπως παρουσιάζεται στην εικόνα 2.2. Ο έλεγχος της έκφρασης γονιδίων από την περιοχή γονιδίων-υποκινητών εμποδίζει τη σύνθεση περιττών προϊόντων, αν και ο έλεγχος της έκφρασης μπορεί να πραγματοποιηθεί και μετά από την έναρξη. Ανάλογα με τις ρυθμιστικές πιέσεις, οι promoters μπορούν να έχουν μια συχνότητα έναρξης πέρα από τα 10.000-fold range, με τους διαφορετικούς promoters να έχουν διαφορετικά βασικά επίπεδα μεταγραφής. Αυτή η ευελιξία επιτρέπει τον αποδοτικό έλεγχο ενός μεγάλου αριθμού κυτταρικών συστατικών.

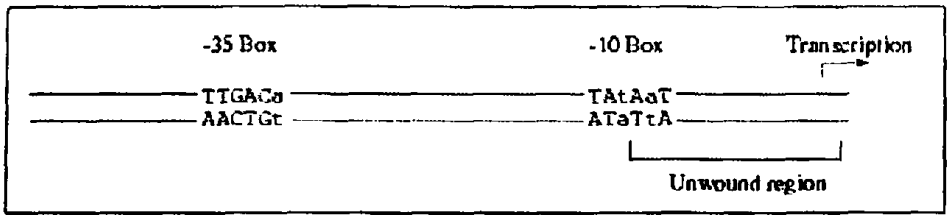


Εικόνα 2-2 Περιοχές είναι γενικά της μορφής: promoter - gene(s) - terminator

2.2.1.2 Promoters

Η περιοχή γονιδίων-υποκινητών (promoters) των προκαρυωρικών οργανισμών περιέχει σύντομες ακολουθίες που είναι σημαντικές για τον έλεγχο της έκφρασης γονιδίων (Εικόνα 2.3). Αυτές οι ακολουθίες εμφανίζονται προς τα πάνω από την περιοχή έναρξης μεταγραφής (+1) στις θέσεις -10 και -35. Η περιοχή -35 έχει

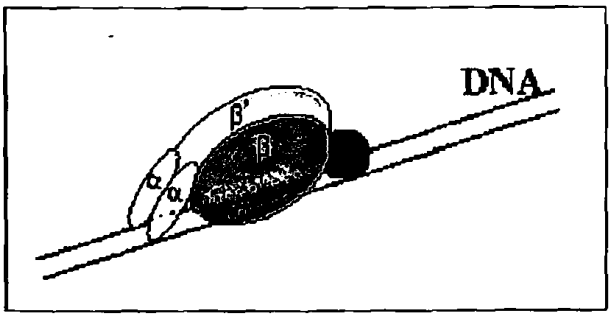
επιπτώσεις στη σύνδεση της πολυμεράσης RNA, ενώ η περιοχή -10 (αρχικά αποκαλούταν *Pribnow box*) στη μεταγενέστερη μεταγραφή.



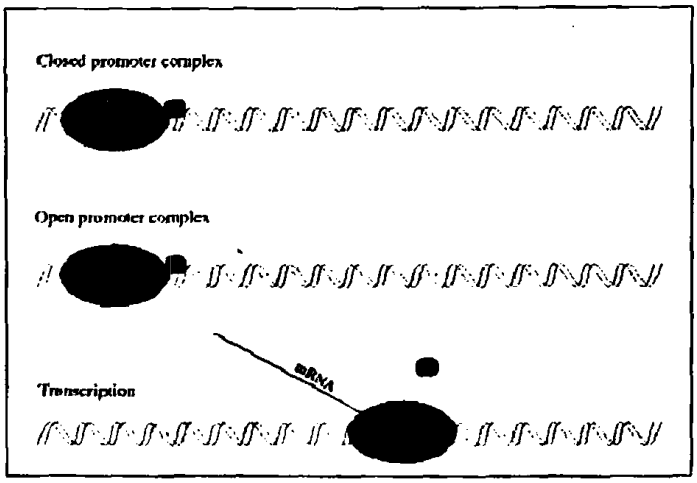
Εικόνα 2-3 Περιοχή γονιδίων-υποκινητών

2.2.1.3 RNA polymerase

Η RNA πολυμεράση Εικόνα 2.4, Εικόνα 2.5 είναι το ένζυμο που είναι αρμόδιο για τη μεταγραφή. Η πολυμεράση RNA μπορεί να υπάρξει σε μια από δύο καταστάσεις - τη κατάσταση έναρξης που σχετίζεται με τη σύνδεση με τους promoters (open promoter complex), και τη κατάσταση επιμήκυνσης για τη χαλαρή σύνδεση και τη σύνθεση mRNA (closed promoter complex).



Εικόνα 2-4 RNA πολυμεράση



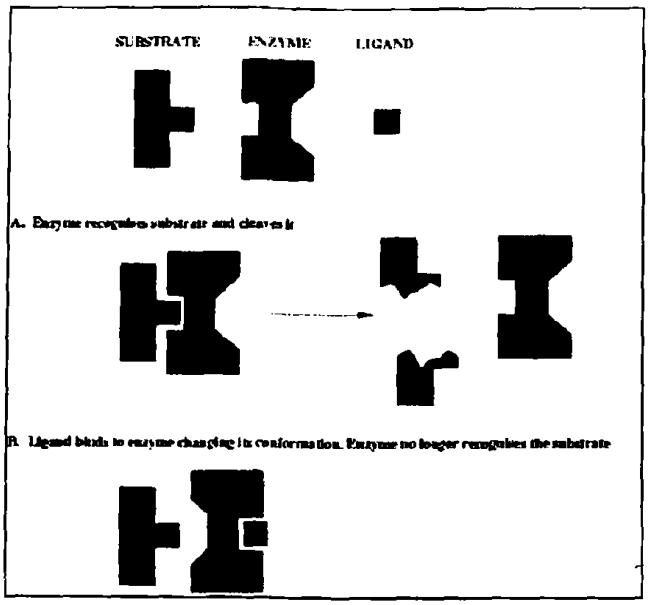
Εικόνα 2-5 mRNA υποκινητές

2.2.2 Μεταβολισμός

Η ανάπτυξη των βακτηρίων απαιτεί το συντονισμό πολυάριθμων ενζυμικών αντιδράσεων που συνδυάζονται, δημιουργώντας ένα μεταβολικό μονοπάτι (metabolic pathway). Αυτά τα μονοπάτια μπορούν να οδηγήσουν στη σύνθεση ουσιαστικών μεσαζόντων (π.χ. αμινοξέα για τις πρωτεΐνες ή νουκλεοτίδες για τα νουκλεϊνικά οξέα), ή μπορούν να αναμειχθούν στη διακοπή των υποστρωμάτων για να παρέχουν ενέργεια και προδρόμους ανεφοδιασμού. Ο έλεγχος των βιοσυνθετικών (δημιουργικών) και κατασταλτικών διαδικασιών απαιτείται για να αποτρέψει την ανεπαρκή χρησιμοποίηση των διαθέσιμων ενώσεων και για να ελαχιστοποιήσει την παραγωγή των ανεπιθύμητων ή περιττών προϊόντων.

2.2.2.1 Κανονισμός της ενζυμικής δραστηριότητας

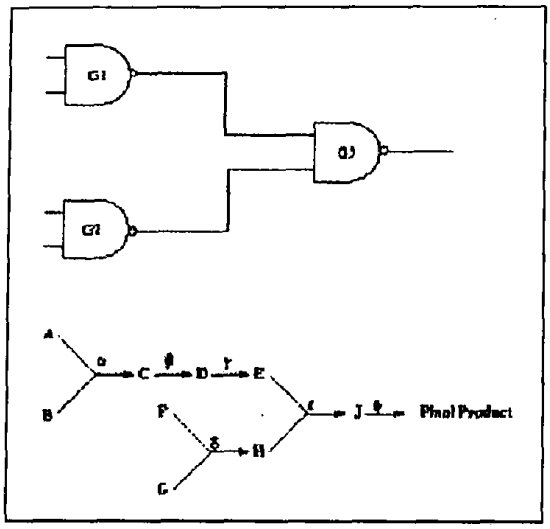
Όπως αναφέρθηκε προηγουμένως, αλλοστερικά ένζυμα (allosteric enzymes) μπορούν να υπάρξουν σε περισσότερους από έναν σχηματισμούς, πράγμα που εξαρτάται από την παρουσία ή την απουσία ενός ligand. Επομένως, εκτός από την ενεργό περιοχή των αλλοστερικών ενζύμων (η περιοχή όπου η αντίδραση υποστρωμάτων πραγματοποιείται) υπάρχει μια *περιοχή συνδέσεων ligand* που, όταν καταλαμβάνεται, αλλάζει τη διαμόρφωση και κατά συνέπεια και τις ιδιότητες της πρωτεΐνης. Για παράδειγμα, εάν ένα αμινοξύ είναι παρόν στο μέσο αύξησης, αυτό το μόριο μπορεί να ενεργήσει ως ligand σε ένα ένζυμο μέσα στη βιοσυνθετική διάβαση (pathway) για εκείνο το αμινοξύ, κλείνοντας αποτελεσματικά την παραγωγή. Συνεπώς το ligand μπορεί να είναι απολύτως ανεξάρτητο από το υπόστρωμα στο οποίο το ένζυμο λειτουργεί. Τα αλλοστερικά ένζυμα μπορούν είτε να αναπτύξουν τη δραστηριότητά τους που ενισχύεται παρουσία του ligand τους, είτε να την μειώσουν.



Εικόνα 2-6 Κανονισμός της ενζυμικής δραστηριότητας

2.2.2.2 Μεταβολικά μονοπάτια ως Boolean δίκτυα

Εάν χρησιμοποιείται ένα υποθετικό μονοπάτι όπως παρουσιάζεται πιο κάτω, μπορεί να χρησιμοποιηθεί ένα Boolean δίκτυο με βάση αυτήν την διάβαση. Μια εφικτή εφαρμογή θα περιελάμβανε το δευτεροβάθμιο μεταβολισμό, όπου τα προϊόντα της διάβασης δεν απαιτούνται πραγματικά για τη συνεχή αύξηση του οργανισμού. Επιπρόσθετα ο μεταβολισμός εμφανίζεται αργότερα στη φάση ανάπτυξης του οργανισμού σε σχέση με τον αρχικό μεταβολισμό, και επειδή τα προϊόντα του δεν είναι απολύτως απαιτούμενα, ο χειρισμός αυτών των διαδικασιών δεν θα ήταν καταστρεπτικός για το κύτταρο. Ένα παράδειγμα του δευτεροβάθμιου μεταβολισμού είναι η παραγωγή των αντιβιοτικών που, αν και δίνουν ένα ανταγωνιστικό πλεονέκτημα, δεν γίνεται απαραίτητα για την ανάπτυξη του οργανισμού.



Εικόνα 2-7 Μοντελοποίηση μεταβολικής διαδικασίας με χρήση Boolean δικτύου

Κεφάλαιο 3: Έκφραση γονιδίων και τεχνικές μέτρησης

Παρακάτω γίνεται αναφορά στις μεθόδους μέτρησης έκφρασης των γονιδίων και στον τρόπο που αυτά τα δεδομένα μπορούν να χρησιμοποιηθούν. Συζητούνται επίσης οι δυσκολίες που υπάρχουν όσον αφορά την εξαγωγή γονιδιακών δικτύων από πολλά δεδομένα.

Στη μηχανική μάθηση είναι ευρέως γνωστό ότι όσο περισσότερες μεταβλητές έχουμε να μοντελοποιήσουμε, τόσο δυσκολότερη είναι η μοντελοποίηση, επειδή το μέγεθος του χώρου αναζήτησης αυξάνεται εκθετικά με τον αριθμό των παραμέτρων του μοντέλου. Αυτό αναφέρεται συχνά ως **"Curse of Dimensionality"**. Αν και τις περισσότερες φορές θέλουμε να είμαστε σε θέση να εξετάσουμε όσο το δυνατόν περισσότερες μεταβλητές του προβλήματος, συχνά επιλέγουμε εκείνες που θεωρούνται πιο σημαντικές στο σύστημα, και αγνοούμε τις υπόλοιπες. Η χρήση της a priori γνώσης μπορεί να βοηθήσει στο περιορισμό του χώρου επιλογής μοντέλων.

Ποίες είναι όμως εκείνες οι μεταβλητές που μας ενδιαφέρουν στα γενετικά δίκτυα. Η κατάσταση ενός κυττάρου αποτελείται από πολλές εσωτερικές και εξωτερικές παραμέτρους που καθορίζουν τη συμπεριφορά του. Ένα κεντρικό δόγμα της μοριακή βιολογία αναφέρει ότι, η δραστηριότητα ενός κυττάρου καθορίζεται από το ποία γονίδια εκφράζονται και ποια όχι. Εάν ένα γονίδιο εκφράζεται, το DNA του μεταγράφεται στο συμπληρωματικό RNA "αγγελιοφόρων" (mRNA), το οποίο στην συνέχεια μεταφράζεται σε κάποια πρωτεΐνη. Μπορούμε να μετρήσουμε το επίπεδο έκφρασης κάθε γονιδίου με την μέτρηση του αριθμού των αντιγράφων του mRNA που είναι παρόντα στο κύτταρο. Εκτός από mRNA και τα πρωτεϊνικά επίπεδα, υπάρχουν και άλλες παράμετροι, συμπεριλαμβανομένου του όγκου των κυττάρων, του ποσοστού αύξησης των καταστάσεων μεθυλίωσης του DNA, της κατάστασης των πρωτεϊνών, των ιονικών επιπέδων, κλπ.

Υπάρχουν δύο σημαντικοί τρόποι για την ανάκτηση του προφίλ έκφρασης. Ο πρώτος τρόπος είναι με μέτρηση του mRNA και ο δεύτερος με μέτρηση των πρωτεϊνικών μορίων. Ο πιο συνήθης από τους δύο είναι ο πρώτος. Αρχικά τα

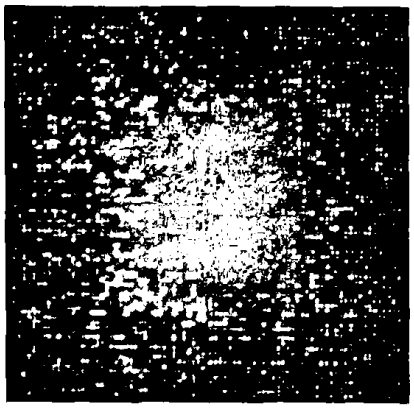
mRNA μόρια πρέπει να διαχωριστούν από όλα τα άλλα κυτταρικά μόρια. Κάθε mRNA έχει μια μοναδική ετικέτα προσκολλημένη στο ένα του άκρο, μια αλληλουχία μικρού μήκους από A βάσεις (polyA). Κατά την μέτρηση των mRNA γίνετε εκμετάλλευση αυτής της ιδιότητας. Δημιουργείται μια σταθερή μικρή αλυσίδα από polyT βάσεις. Καθώς το περιεχόμενο του κυττάρου περνάει από τη βάση οι άκρες polyA που είναι στο τέλος του mRNA κάνει ζεύγος με το polyT. Κατά αυτόν τον τρόπο απομονώνεται το mRNA. Δυστυχώς τα mRNA μόρια δεν είναι τόσο σταθερά και εμφανίζονται σε πολύ μικρές ποσότητες. Έτσι, έχοντας δεσμεύσει τα mRNAs, πρέπει να τα μετατρέψουν σε κάτι πολύ πιο σταθερό και εύκολα μετρήσιμο.

3.1 Επίπεδα mRNA

Μερικές τεχνικές που έχουν αναπτυχθεί αποσκοπούν στην μέτρηση των επιπέδων του mRNA κατά την διάρκεια συγκεκριμένων πειραμάτων. Μερικές από αυτές περιγράφονται παρακάτω.

3.1.1 cDNA Μικροσυστοιχίες

Δημιουργήθηκε στο Stanford University, οι μικροσυστοιχίες είναι slide γυαλιού πάνω στο οποίο το cDNA έχει εναποτεθεί από ειδικά ρομποτικά μηχανήματα εκτύπωσης. Είναι ιδανικά για ανάλυση έκφρασης μέχρι 10.000 cDNA κλώνων ανά πίνακα.



Εικόνα 3-1 cDNA microarray για μέτρηση των επιπέδων έκφρασης γονιδίων

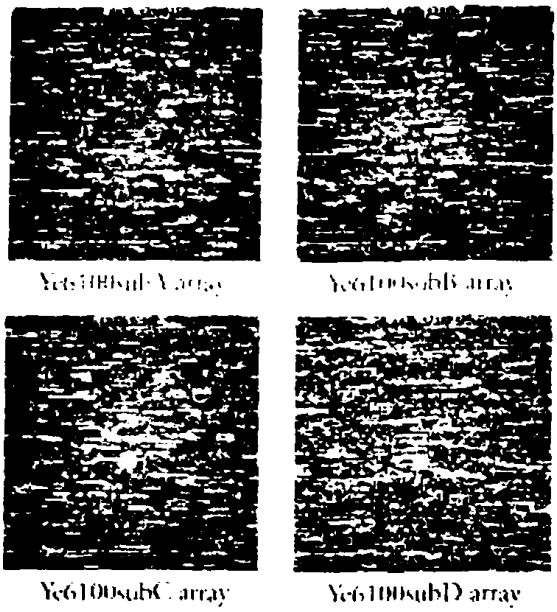
Οι μετρήσεις μικροσυστοιχιών πραγματοποιούνται ως διαφορικές υβριδοποιήσεις για να ελαχιστοποιήσουν τα λάθη που προέρχονται από τη μεταβλητότητα που

παρουσιάζεται στο spotting του cDNA: mRNA από δύο διαφορετικές πηγές, χαρακτηρισμένες με δύο διαφορετικές χρωστικές ουσίες φθορισμού, σαρώνεται πάνω στην πίνακα την ίδια στιγμή. Το σήμα φθορισμού από κάθε πληθυσμό mRNA αξιολογείται ανεξάρτητα, και έπειτα χρησιμοποιείται για τον υπολογισμό του λόγου πειραματικό / ελεγχόμενο.

Σε εργαστήρια στο Stanford έχουν χρησιμοποιηθεί microarrays για μετρήσεις των επίπεδων έκφρασης γονιδίων για ολόκληρο το γονιδίωμα σακχαροκυττάρων (περίπου 6400 ευδιάκριτες ακολουθίες DNA) κατά τη διάρκεια α) της diauxic μετατόπισης (μετάβαση από το μεταβολισμό ζάχαρης στο μεταβολισμό αιθανόλης), β) sporulation και γ) ολόκληρου του κύκλου των κυττάρων. Αυτά τα δεδομένα είναι δημόσια διαθέσιμα. Οι εκφράσεις γονιδίων από το Incyte είναι διαθέσιμες με πρότυπα για τον άνθρωπο, τον αρουραίο, το ποντίκι, φυτά και μικροβιακά γονιδιώματα.

3.1.2 Συστοιχίες Ολιγονουκλεοτιδίων

Παράγονται με την τεχνική Affymetrix, αποτελούνται από μικρά πλακίδια γυαλιού με εκατοντάδες μικρά 20-μερή ολιγονουκλεοτίδια probes προσαρτώμενα στην επιφάνειά τους. Τα ολιγονουκλεοτίδια συνθέτονται κατευθείαν στην επιφάνεια χρησιμοποιώντας συνδυασμό από ημιαγώγιμες φωτολιθογραφίες και χημικής φωτοσύνθεσης. Εξαιτίας της πολυπλοκότητας της διαδικασίας, μεγάλος αριθμός από mRNAs μπορούν να αποτυπωθούν παράλληλα. Παρόλα' αυτά η κατασκευή και η ανάγνωση των chips απαιτεί ακριβό εξοπλισμό. Τα τελευταία chips έχουν πάνω από 65,000 διαφορετικά probes, με πολλά probes να αντιστοιχούν σε κάθε mRNA. Εικόνα 3.2

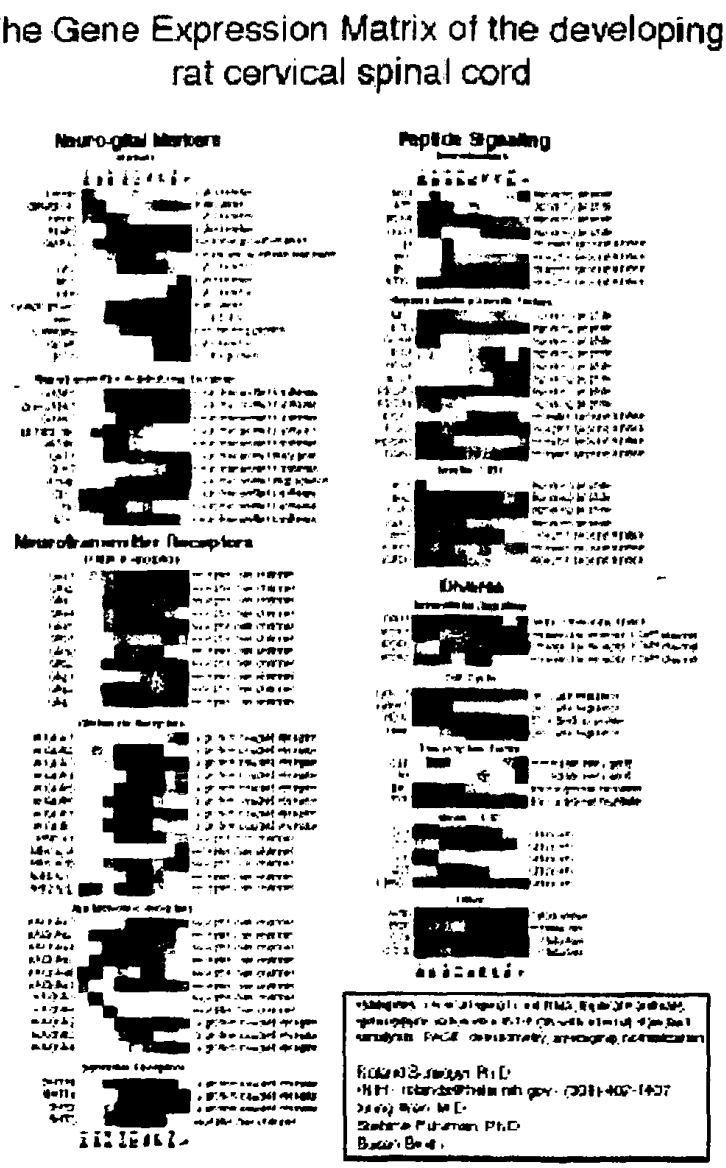


Εικόνα 3-2Oligonucleotide chips

Η Affymetrix κατασκευάζει GeneChips για 42,000 ανθρώπινα γονίδια και ESTs, για 30,000 murine (είδος ποντικιού) γονίδια και ESTs, και 6,100 ORFs μύκητα ζύμης (όλο το γονιδίωμα). Λίγα δεδομένα είναι διαθέσιμα, με εξαίρεση την βάση δεδομένων έκφρασης του S. Cerevisiae.

3.1.3 RT-PCR

Η μέτρηση της έκφρασης των γονιδίων χρησιμοποιώντας την RT-PCR (Reverse Transcriptase Polymerase Chain Reaction) τεχνική γίνεται ως εξής, το mRNA αρχικά μετατρέπεται σε cDNA, και το cDNA στη συνέχεια μετασχηματίζεται σε μετρήσιμα επίπεδα χρησιμοποιώντας PCR. Χρησιμοποιώντας calibration τεχνικές, το RT-PCR μπορεί να επιτύχει μεγάλη ακρίβεια σε συνδυασμό με εξαιρετική ευαισθησία. Η μέθοδος απαιτεί PCR primers για όλα τα γονίδια ενδιαφέροντος, επίσης δεν είναι παράλληλη όπως οι προηγούμενες τρεις μέθοδοι, γι αυτό η αυτοματοποίηση είναι σημαντική για την αύξηση της ταχύτητας.



Εικόνα 3-3 RT-PCR (Reverse Transcriptase Polymerase Chain Reaction) για μέτρηση της έκφρασης των γονιδίων

Ο Roland Somogyi έχει χρησιμοποιήσει αυτή τη μέθοδο για να μετρήσει τα επίπεδα έκφρασης 112 γονιδίων σε εννέα διαφορετικές χρονικές στιγμές κατά την διάρκεια ανάπτυξης καρκινικών κυττάρων στο νωτιαίο μυελό σε αρουραίους, και επίσης 70 γονίδια κατά την διάρκεια ανάπτυξης του ιππόκαμπου. Πολλά από αυτά τα δεδομένα είναι διαθέσιμα.

3.1.4 Σειριακή ανάλυση των εκφράσεων γονιδίων

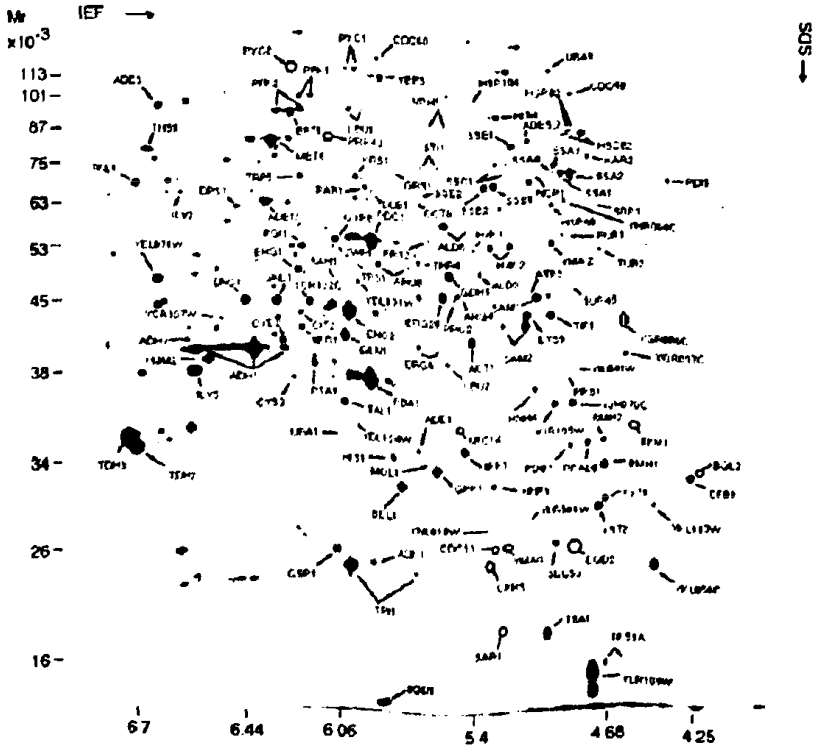
Η μέθοδος SAGE χρησιμοποιεί πολύ διαφορετικές τεχνικές μέτρησης των επιπέδων του mRNA. Αρχικά, διπλά stranded cDNA δημιουργούνται από mRNA. 10 base ζευγάρια "αλληλουχία tag" κόβονται από μια συγκεκριμένη τοποθεσία σε κάθε

cDNA. Οι σειρές συνδέονται σε ένα μεγάλο διπλό stranded DNA το οποίο μπορεί να μετασχηματιστεί. Αυτή η μέθοδος έχει δύο πλεονεκτήματα: οι σειρές mRNA δεν χρειάζεται να είναι εκ των προτέρων γνωστές -άρα αναγνωρίζει και άγνωστα γονίδια—και χρησιμοποιεί τεχνολογία η οποία μπορεί να είναι διαθέσιμη από πολλά εργαστήρια.

Η μέθοδος SAGE έχει χρησιμοποιηθεί για ανάλυση συνόλου γονιδίων τα οποία εκφράζονται κατά την διάρκεια τριών διαφορετικών φάσεων του κύκλου το κυττάρου του μύκητα (yeast). Η SAGE έχει επίσης χρησιμοποιηθεί για την παρακολούθηση της έκφρασης τουλάχιστον 45,000 ανθρώπινων γονιδίων.

3.2 Protein levels

Τα επίπεδα πρωτεϊνών (Εικόνα 3.4) είναι δυσκολότερο να ποσοτικοποιηθούν σε σχέση με τα mRNA επίπεδα. Η 2D-PAGE διαχωρίζει τις πρωτεΐνες σε ένα δύο διαστάσεων φύλλο από gel. Πρώτα γίνεται ο διαχωρισμός στη μια διάσταση βασισμένο στα ισοηλεκτρικά σημεία, και στην συνέχεια στην άλλη διάσταση βασισμένη στο μοριακό βάρος. Τα αποτελέσματα είναι δύο διαστάσεων εικόνα με μεγάλο αριθμό πρωτεϊνικών "spots". Η ένταση του κάθε spot είναι ανάλογη του πλήθους συγκεκριμένων πρωτεϊνών οι οποίες παρίστανται.



Εικόνα 3-4 Επίπεδα πρωτεϊνών

Αυτό δεν είναι εκ των προτέρων γνώση για το ποία πρωτεΐνη αντιστοιχεί σε πιο spot, παρόλο που η θέση των γνωστών πρωτεϊνών μπορεί να εκτιμηθεί. Επίσης νέες μικροακολουθίες και πολλές φασματομετρικές τεχνικές επιτρέπουν στα spots να αναγνωρίσουν πρωτεΐνες.

Υπάρχουν πολλά 2D gel βάσεις για το E. coli, το yeast, τη Drosophila, τον αρουραίο, το ποντίκι, τον άνθρωπο, κτλ. Ένα από τα πιο σημαντικά είναι η SWISS-2DPAGE βάση δεδομένων, η οποία περιλαμβάνει συνολικά 518 καταγραφές από τον άνθρωπο, το yeast, το E. coli και το Dictyostelium.

3.3 Ορισμός της ανάλυσης γενετικών δικτύων

Τα περισσότερα βιολογικά φαινόμενα προκαλούνται από ένα σύνολο συνεργαζόμενων βιοχημικών οντοτήτων συμπεριλαμβανομένου του mRNA, των πρωτεϊνών, των μικρών μορίων (όπως οι ορμόνες) ή των ιόντων. Η ανάλυση γενετικών δικτύων εκμεταλλεύεται μαζικά τις παράλληλες μετρήσεις αυτών των στοιχείων προκειμένου να καθοριστούν οι ρυθμιστικές αλληλεπιδράσεις μεταξύ των γονιδίων και των παραγώγων τους. Με βάση αυτές τις ρυθμιστικές αλληλεπιδράσεις, η ανάλυση σκοπεύει επίσης να παρέχει τη δυνατότητα πρόβλεψης για τη συμπεριφορά των μεμονωμένων γονιδίων κάτω από συγκεκριμένες καταστάσεις αλλά και τη γενική συμπεριφορά του συστήματος.

3.4 Dataset μικροσυστοιχιών

Για τις ανάγκες της εργασίας αποφασίστηκε να χρησιμοποιηθούν τα δεδομένα από πειράματα σε κύτταρα κατά την διάρκεια του κύκλου του κυττάρου. Σε αυτά τα πειράματα οι τιμές έκφρασης των στοιχείων μετρήθηκαν κατά τη διάρκεια διαφόρων περιβαλλοντικών μεταβολών. Τα DNA microarrays χρησιμοποιήθηκαν για να μετρηθούν οι αλλαγές στα επίπεδα μεταβολών. Τα κυτταρικά στοιχεία αποκρίθηκαν σε μεταβολές θερμοκρασίας, σε προσβολή τους με υπεροξείδιο του υδρογόνου, menadione φαρμάκων, sulfhudryl -οξειδωτικό diamide, δισουλφιδίων-μείωσης, υπέρ και υπό-οσμωτικό κλονισμό, προσβολή με amino-acid, μείωση πηγής αζώτου. Ένα μεγάλο σύνολο γονιδίων, που αποτελείται από περίπου 6500 γονίδια, παρουσίασε μεταβολές στις διαταραχές που προκλήθηκαν. Τα δεδομένα που χρησιμοποιήθηκαν περιγράφονται λεπτομερέστερα στο κεφάλαιο της μεθοδολογίας

Κεφάλαιο 4: Ανακατασκευή γενετικών δικτύων: βιβλιογραφία

Σε αυτό το κεφάλαιο περιγράφονται περιληπτικά μέθοδοι και τεχνικές που έχουν προταθεί για την ανακατασκευή των γενετικών δικτύων. Υπάρχουν δύο ειδών αναλύσεις που γίνονται σε δεδομένα έκφρασης γονιδίων. Ο τύπος ανάλυσης που επιλέγεται εξαρτάται από τη μέθοδο με την οποία τα δεδομένα παράχθηκαν, το οποίο εξαρτάται στη συνέχεια από τους πειραματικούς στόχους.

Ο πρώτος τύπος ανάλυσης εφαρμόζεται όταν παράγεται ένα σύνολο δεδομένων από γονίδια διαφόρων οργανισμών. Αυτά τα δεδομένα μπορούν να χρησιμοποιηθούν με σκοπό την ταξινόμηση, όπου τα άτομα πρέπει να ομαδοποιηθούν σε κατηγορίες σύμφωνα με κάποια γνωστή παρατήρηση της κατάστασής τους. Σε αυτόν τον τύπο ανάλυσης, ο υπολογιστικός στόχος είναι να ταξινομηθεί σωστά και να προβλεφθεί η κατάσταση (π.χ. "ασθενές" ή "μη ασθενές") του γονιδίου, λαμβάνοντας υπόψη το γενετικό του profile. Έρευνες αυτού του τύπου μπορούν να βρεθούν σε εργασίες όπως οι [Golub et al. 1999] και [Su et al. 2002] όπου με υπολογιστικές τεχνικές γίνεται ταξινόμηση στα πραγματικά microarray δεδομένα.

Ο δεύτερος τύπος ανάλυσης μπορεί να γίνει όταν παράγονται διαδοχικά microarrays κατά τη διάρκεια ενός πειράματος σε έναν οργανισμό. Η καταγραφή αυτών των γενετικών πληροφοριών επιτρέπει να ερευνηθούν οι αλληλεπιδράσεις μεταξύ των γονιδίων στον οργανισμό κατά τη διάρκεια κάποιου χρονικού διαστήματος. Η διαδοχή των microarrays μπορεί να θεωρηθεί ως ένα κλειστό σύστημα στο οποίο τα επίπεδα έκφρασης γονιδίων σε ένα χρονικό βήμα έχουν άμεση επίδραση στις τιμές έκφρασης των γονιδίων στο επόμενο. Η εύρεση αυτών των αλληλεπιδράσεων μεταξύ των γονιδίων με υπολογιστικές μεθόδους θα μπορούσε να έχει πολλά οφέλη για τη βιολογία και την ιατρική. Το αποτέλεσμα μιας τέτοιας μελέτης είναι η δημιουργία ενός δικτύου το οποίο περιγράφει τις αλληλεπιδράσεις των γονιδίων, τα δίκτυα αυτά είναι γνωστά γενικά ως γενετικά δίκτυα, ή ρυθμιστικά δίκτυα, μεταβολικά δίκτυα ανάλογα με τον τύπο της λειτουργίας του κυττάρου που μελετούν.

Ορισμένες επιτυχείς προσπάθειες έχουν συνδέσει στενά τη μοντελοποίηση και τα πειραματικά δεδομένα, τέτοιες μελέτες μπορούμε να δούμε στις εργασίες [Yuh, et al. 1998],[Yuh, et al. 2001]. Ο τομέας των γενετικών δικτύων είναι ένας τομέας μεγάλου ενδιαφέροντος, κάτι το οποίο αποδεικνύεται από την αύξηση των ερευνητικών εργασιών που δημοσιεύτηκαν. Στις αρχές του 2000 μια αναζήτηση στον όρο "δίκτυα γονιδίων" επέστρεφε λιγότερες από δέκα αναφορές, και καμία από τις εργασίες δεν περιέλαβε τη μοντελοποίηση. Τους πρώτους μήνες του 2003 εντούτοις, η ίδια αναζήτηση επέστρεφε αρκετά περισσότερα αποτελέσματα, και πολλά από αυτά συμπεριλάμβαναν εργασίες για την μοντελοποίηση διαφόρων ειδών γενετικών δικτύων. Διάφορες μέθοδοι που έχουν χρησιμοποιηθεί για την μοντελοποίηση βιολογικών δικτύων είναι τα Μπεϋζιανά δίκτυα [Friedman 200], τα φορμαλιστικά δίκτυα βασισμένα σε κανόνες [Meyers & Friedland, 1984], τα Boolean δίκτυα και τα υβριδικά Boolean/συνεχή [Yuh, et al. 1998],[Yuh, et al. 2001].

4.1 Θεωρία γράφων

Τα γενετικά δίκτυα βρίσκονται αυτήν την περίοδο στην αιχμή της βιολογικής έρευνας. Οι βιολόγοι σήμερα για να εξακριβώσουν το ρόλο ενός γονιδίου σε ένα κύτταρο, διαταράσσουν συστηματικά τα γονίδια του οργανισμού και μετρούν τις αλλαγές στην έκφρασή τους. Με την ανάλυση των αποτελεσμάτων του οργανισμού στο σύνολο των διαταραχών, οι βιολόγοι ελπίζουν να μάθουν τις αιτιώδεις σχέσεις που κυβερνούν τα γενετικά δίκτυα. Σε αυτό το πρόβλημα το 2001, ο Wagner [Wagner et al. 2001] παρουσιάζει έναν αλγόριθμο για τα γενετικά δίκτυα. Ο αλγόριθμος του Wagner χρησιμοποιεί τη θεωρία γράφων για να αναλύσει τα microarray δεδομένα που παρήχθησαν σε πειράματα διαταραχής γονιδίων. Δεδομένου ότι τρέχει σχετικά γρήγορα, ο αλγόριθμος επιχειρεί την ανακατασκευή γενετικών δικτύων που περιέχουν πολλούς κόμβους.

4.2 Boolean δίκτυα

Τα Boolean δίκτυα χρησιμοποιήθηκαν σε μια προσπάθεια να μειωθεί η πολυπλοκότητα της εξέτασης των πραγματικών δεδομένων, ο D'Haesseleer [D'Haeseleer et al. 1999] (και παλαιοτέρα [Kauffman et al. 1996]) περιγράφει μια Boolean αναπαράσταση των δεδομένων έκφρασης γονιδίων. Τα Boolean δεδομένα, αποτελούνται από μηδέν και ένα τα οποία περιγράφουν "τα επίπεδα ενεργοποίησης"

τού γονιδίου. Αυτό είναι συμφέρον για διάφορους λόγους, αρχικά επειδή περιορίζει την πολυπλοκότητα των δεδομένων και επομένως του παραγόμενου μοντέλου, δεύτερον επειδή τα Boolean δίκτυα περιγράφονται από εκφράσεις που μπορούν να αναλυθούν με μαθηματικά [Wuensche et al. 1998], και τρίτον επειδή τα δεδομένα μπορούν σχετικά εύκολα να δημιουργηθούν και επομένως να δοκιμαστούν.

4.3 Μπεϋζιανά δίκτυα

Για το σκοπό της ανακατασκευής των γενετικών δικτύων έχουν χρησιμοποιηθεί και Μπεϋζιανά δίκτυα. Η πιο γνωστή εργασία στην περιοχή των Μπεϋζιανών δικτύων και δεδομένων έκφρασης γονιδίων [Spirtes et al. 2000], περιγράφει την κατασκευή των Μπεϋζιανών δικτύων από τα δεδομένα με τον τρόπο που δημιουργούνται και τα δίκτυα γονιδίων. Τα Μπεϋζιανά δίκτυα προσφέρονται για αυτόν τον τύπο ανάλυσης επειδή είναι κατευθυνόμενα άκυκλα γραφήματα που μπορούν να αντιπροσωπεύσουν τις βασικές σχέσεις μεταξύ των μεταβλητών (όπως τα επίπεδα έκφρασης γονιδίων). Η εργασία περιγράφει τη γενική προσέγγιση των Μπεϋζιανών δικτύων όσον αφορά τα δεδομένα έκφρασης γονιδίων, αλλά δεν περιγράφει οποιαδήποτε πειράματα με αυτά τα δεδομένα.

4.4 Γενετικοί αλγόριθμοι

Στις εργασίες [Ando et al. 2000], [Ando et al. 2001a], [Ando et al. 2001b] περιγράφεται η εφαρμογή των γενετικών αλγορίθμων για την ανακατασκευή ρυθμιστικών δικτύων από δεδομένα έκφρασης γονιδίων. Οι γενετικοί αλγόριθμοι εφαρμόζονται στο πρόβλημα με διάφορους τρόπους, π.χ. προσθέτοντας θόρυβο στην reverse engineering διαδικασία, αλλά τα αποτελέσματα που εξάγονται συνήθως αναφέρονται σε μια χρωμοσωματική αναπαράσταση του δικτύου. Το χρωμόσωμα ενός GA είναι ένας κωδικοποιημένος πίνακας τιμών κινητής υποδιαστολής που αντιστοιχούν σε πίνακα βαρών μεταξύ των χρονικών βημάτων των γονιδίων. Τα πλεονεκτήματα αυτής της διαδικασίας είναι ότι ο γενετικός αλγόριθμος [Weaver et al. 1999] ορίζει μια μέθοδο μοντελοποίησης των ρυθμιστικών δικτύων γονιδίων, ενώ χρησιμοποιεί μια reverse engineering τεχνική. Ωστόσο η εφαρμογή γενετικών αλγορίθμων σε αυτό το πρόβλημα είναι νέα κάτι που σημαίνει ότι ορισμένα στοιχεία του προβλήματος πολυπλοκότητας εξετάζονται ακόμα σε αυτήν την περιοχή. Υπάρχουν εντούτοις διάφορα μειονεκτήματα στην εφαρμογή γενετικών αλγορίθμων

σε αυτήν την περιοχή. Σε δοκιμές, έχει βρεθεί πως οι γενετικοί αλγόριθμοι είναι ιδιαίτερα φτωχοί στην εύρεση των βέλτιστων πεδίων τιμών υπό όρους. Οι γενετικοί αλγόριθμοι τείνουν να είναι ασυμπτωτικοί σε προβλήματα που ζητάμε βέλτιστη λύση κάτι που τους κάνει πολύ αργούς.

4.5 Gaussian μοντέλα

Μια ακόμα μέθοδος που προτείνεται στην εργασία [Toh et al. 2000] χρησιμοποιεί γραφικά Gaussian μοντέλα -μια προσέγγιση παρόμοια με τα Μπεϋζιανά δίκτυα- για να εξαγάγουν πληροφορίες συσχετισμού γονιδίων. Η μέθοδος είναι στατιστική και περιλαμβάνει τη χρησιμοποίηση της συμεταβλητότητας των γονιδίων ως αρχική τεχνική αξιολόγησης. Τα αποτελέσματα απεικονίζονται και ως κατευθυνόμενες γραφικές παραστάσεις και ως πίνακας των βαρών μεταξύ των γονιδίων. Τα προβλήματα με το σύστημα όπως περιγράφονται από τους συγγραφείς είναι ότι οι σχέσεις αντιστοιχούν σε συσχετισμούς μεταξύ των γονιδίων και όχι σε κάποια αιτιώδη σχέση. Χωρίς αιτιώδεις σχέσεις, τα μοντέλα μπορούν μόνο να αντιμετωπισθούν ως πληροφοριακά και όχι ως αιτιοκρατικά. Αυτό που κάνει αυτήν την προσέγγιση ενδιαφέρουσα είναι ότι τα παραγόμενα δίκτυα συγκρίνονται με τα Μπεϋζιανά αποτελέσματα δικτύων όπως [Friedman et al. 2000] στο ίδιο σύνολο δεδομένων. Και οι δύο τεχνικές χρησιμοποιούν διαφορετικές μεθόδους ομαδοποίησης που δίνει διαφορετικές ομάδες γονιδίων που χρησιμοποιούνται ως κόμβοι στη γραφική παράσταση, αλλά μερικές σχέσεις που βρίσκονται από την Μπεϋζιανή μέθοδο είναι όμοιες με αυτές της Gaussian προσέγγισης.

4.6 Διαφορικές εξισώσεις

Άλλά μαθηματικά μοντέλα που έχουν επίσης χρησιμοποιηθεί είναι μοντέλα διαφορικών εξισώσεων ODE (ordinary differential equations), με χρήση της αρχής δράσης της μάζας (Mass Action), που δηλώνει ότι το ποσοστό της αντίδρασης είναι ανάλογο προς τη συγκέντρωση των αντιδραστηρίων. Η χρονική εξέλιξη περιγράφεται από ένα σύστημα διαφορικών εξισώσεων. Οι λόγοι για την επικράτηση της θεωρίας δράσης της μάζας είναι πολλοί, ο σημαντικότερος είναι ότι οι προσεγγίσεις βασισμένες στα αποτελέσματα διαφορικών εξισώσεων είναι γενικά σε συμφωνία με πολλά δεδομένα [Hynne, et al. 2001], [Poolman, et al. 2001]. Παρόλα αυτά, οι διαφορικές εξισώσεις μπορούν να μην είναι το καλύτερο εργαλείο για τη

μοντελοποίηση των βιολογικών διαδικασιών που λαμβάνουν χώρα στα ζωντανά κύτταρα. Συγκρινόμενες με τις διαφορικές εξισώσεις, (όπου έχουν επικρατήσει στη μοντελοποίηση χημικών διαδικασιών), οι στοχαστικές προσεγγίσεις στη βιολογία είναι ακόμα σε μια σχετικά πρώιμη φάση. Αυτό αλλάζει τελευταία αφού αρχίζουν να εμφανίζονται γενικευμένα εργαλεία με χρήση στοχαστικών μοντέλων [Bray, et al. 2001] [Kierzek, et al. 2002].

Μια στοχαστική διαδικασία είναι μια διαδικασία που εξαρτάται από μια τυχαία πράξη, και σε ένα βιολογικό περιβάλλον αυτό σημαίνει ότι το σύστημα υπόκειται σε διακυμάνσεις. Αυτές οι διακυμάνσεις μπορούν να εξαρτώνται από τον αριθμό των παρόντων μορίων, το χρόνο που παίρνει για να γίνει κάποια μοριακή αλληλεπίδραση, κτλ. Ενδιαφέρον παρουσιάζουν ορισμένα υβριδικά μοντέλα όπου τα αποτελέσματα πιθανοτικών μοντέλων χρησιμοποιούνται παράλληλα με ντετερμινιστικές μεθόδους. Σύμφωνα με αυτήν την ιδέα ορισμένοι παράγοντες προσεγγίζονται πιθανολογικά με την προσθήκη ενός όρου "θορύβου" στην ειδάλλως αιτιοκρατική επεξεργασία τους [Meinhardt, et al. 2001]. Η προκύπτουσα διαφορική εξίσωση που προκύπτει καλείται εξίσωση Langevin και είναι της μορφής

$$\frac{dx(t)}{dt} = -aX(t) + f(t) \qquad (4.1)$$

όπου η συνάρτηση παραγωγής του θορύβου $f(t)$ θεωρείται πως είναι η Gaussian.

Κεφάλαιο 5: Αναπαράσταση Βιολογικής Γνώσης

Ένα από τα προβλήματα ου δημιουργήθηκαν σχεδόν παράλληλα με την ανακάλυψη των τεχνικών microarray είναι η διαχείριση και η εύκολη εκμετάλλευση των δεδομένων που παράγονται από αυτά τα πειράματα. Πολύς χρόνος και προσπάθεια σπαταλάται στην έρευνα για όλες τις διαθέσιμες πληροφορίες. Αυτό εμποδίζεται περαιτέρω από τις πολλές παραλλαγές στην ορολογία, και που εμποδίζουν την αποτελεσματική έρευνα. Για το λόγο αυτό είχε καταστεί απαραίτητη η δημιουργία όχι απλά βάσεων δεδομένων που θα αποθηκεύεται αυτή η πληροφορία αλλά τεχνικών αναπαράστασης γνώσης με την οποία θα καλυπτόταν το ευρύ αυτό πεδίο. Για το λόγο αυτό χρησιμοποιήθηκαν οντολογίες, με πιο σημαντικό δείγμα της δουλειάς σε αυτόν τον τομέα το GeneOntology [GO].

5.1 Οντολογίες στην Βιοπληροφορική

Οι οντολογίες ορίζονται με σκοπό τον προσδιορισμό των εννοιών σε ένα συγκεκριμένο πεδίο καθώς επίσης και των σχέσεων μεταξύ των εννοιών. Παρέχουν γνώση της φύσης της πληροφορίας η οποία παράγεται από αυτό το πεδίο και αποτελεί ένα βασικό συστατικό σε κάθε προσπάθεια κατανόησης των όρων σε ένα πεδίο. Κατά συνέπεια η υλοποίηση οντολογιών για τη μοριακή βιολογία και η χρήση των οντολογιών στην βιοπληροφορική είναι ένα από τα βασικά προβλήματα της σύγχρονης βιοπληροφορικής.

Οι οντολογίες στην βιοπληροφορική πρέπει να παράγονται με ένα τυποποιημένο τρόπο π.χ., δημιουργώντας ένα αρχείο με καλά ορισμένη σύνταξη και σημασιολογία. Η ανταλλαγή οντολογιών στην βιοπληροφορική μπορεί να απλοποιηθεί αν συμφωνηθεί ένας μικρός αριθμός από πρότυπα και να βελτιστοποιηθεί αν συμφωνηθεί μόνο ένα.

Κάθε βιολογική βάση δεδομένων χρησιμοποιεί μία οντολογία, είτε συγκεκριμένη είτε ασαφή, με σκοπό την μοντελοποίηση των δεδομένων. Όσο πιο λεπτομερώς ορίζουμε μια οντολογία τόσο πιο μεγάλη ακρίβεια θα μπορέσουμε να αποδώσουμε στα μοντέλα των οποίων η πληροφορία προέρχεται από τις βάσεις δεδομένων. Μια γενικευμένη οντολογία μοντελοποιεί μόνο επιφανειακές όψεις των δεδομένων, και

κατά συνέπεια δεν μπορεί να δεσμεύσει στοιχεία τα οποία είναι σημαντικά για την επίλυση κάποιου προβλήματος. Για παράδειγμα, μια βάση δεδομένων γονιδίων η οποία αποτυγχάνει να καταγράψει ποιοι γενετικοί κώδικες χρησιμοποιούνται για να περιγράψουν μια ακολουθία DNA, δεν παρέχει την πληροφορία την οποία οι χρήστες της βάσης χρειάζονται για να μεταφράσουν αξιόπιστα την σειρά DNA στην αντίστοιχη πρωτεϊνική ακολουθία. Μια σημασιολογικά δύσμορφη οντολογία είναι αυτή που περιγράφει και μοντελοποιεί λανθασμένα το πεδίο εφαρμογών της, και συνεπώς αποδίδει μια βάση δεδομένων της οποίας η δομή παραποιεί ή περιορίζει την πληροφορία την οποία κρατάει. Για παράδειγμα, μια βάση δεδομένων μεταβολισμών η οποία ορίζει μια ένα προς ένα σχέση μεταξύ των ενζύμων και τις αντιδράσεις που καταλύουν δεν μπορεί να μοντελοποιήσει αξιόπιστα το γεγονός ότι ένα δι-λειτουργικό (bifunctional) ένζυμο καταλύει δύο ξεχωριστές αντιδράσεις.

5.2 GeneOntology - GO

Ο στόχος του Gene Ontology Consortium είναι να παράγει ένα δυναμικό Controlled Vocabulary, (CV), το οποίο μπορεί να εφαρμοστεί σε όλους τους οργανισμούς ακόμα και αν η γνώση των γονιδίων και ο ρόλος των πρωτεϊνών στα κύτταρα αυξάνεται και συνεχώς μεταβάλλεται με άλλα λόγια ακόμα και αν η έρευνα παράγει τεράστιο πλήθος από νέα δεδομένα. Το παραγόμενο CV μπορεί να εφαρμοστεί σε όλα τα γονίδια όλων των ευκαριωτικών κυττάρων. Υπάρχουν πολλοί οργανισμοί οι οποίοι συνεισφέρουν στο GO, οι τρεις πρώτοι ήταν η Fly Base (βάση δεδομένων για την μύγα Drosophila melanogaster) Saccharomyces Genome Database (SGD – βάση δεδομένων για budding yeast), Mouse Genome Database και Gene Expression Database (MGD και GXD – βάση δεδομένων για το mouse musculus).

Η GO είναι στην πραγματικότητα τρεις οντολογίες. Αυτές αποτελούν τα controlled Vocabularies για την περιγραφή των Μοριακών Λειτουργιών (Molecular function), των Βιολογικών διαδικασιών (Biological process), και των Κυτταρικών Συνιστωσών (Cellular component) των παράγωγων των γονιδίων. Οι όροι χρησιμοποιούνται σαν γνωρίσματα και χρησιμοποιούνται για να περιγράψουν και να σχολιάσουν συνεργαζόμενες βάσεις δεδομένων.

Οι οντολογίες Molecular function, Biological process και Cellular component αναπαρίστανται σαν κατευθυνόμενοι άκυκλοι γράφοι (Directed Acyclic Graphs

DAGs) ή δίκτυα. Υπάρχουν πολλοί browsers διαθέσιμοι για την αναζήτηση στην GO (π.χ. MGI GO, AmiGo).

"Η GO είναι ένα τρόπος για να επιτευχθεί ενοποίηση των βιολογικών βάσεων δεδομένων." Παρέχει κοινά λεξικά ενώ τομείς όπως 3D δομή, εξέλιξη, κλπ δεν καλύπτονται από την GO.

5.3 Ορισμοί και Σύνταξη των Οντολογιών της GO

5.3.1 Ορισμοί

Το προϊόν ενός γονιδίου είναι ένα φυσικό αντικείμενο. Μπορεί να είναι μια πρωτεΐνη ή ένα RNA. Παραδείγματα παράγωγων γονιδίων είναι τα: alpha-globin και ribosomal RNA.

- Η Μοριακές λειτουργία (**Molecular function**) είναι κάτι το οποίο συμβαίνει. Είναι μια δυνατότητα που έχει ένα φυσικό προϊόν (ή σύνολο προϊόντων). Περιγράφει μόνο τι μπορεί να κάνει χωρίς να ορίζει το πού ή το πότε μπορεί να συμβεί. Παραδείγματα γενικών όρων είναι: "enzyme," "transporter," ή "ligand." Παραδείγματα πιο εξειδικευμένων όρων είναι : "adenylate cyclase," ή "Toll receptor ligand."

- Βιολογικές Διαδικασίες (**Biological process**) είναι ένα βιολογικό συμβάν. Μια Βιολογική διαδικασία επιτυγχάνεται μέσω μιας ή περισσότερων ακολουθιών από μοριακές λειτουργίες (molecular functions). Μια βιολογική διαδικασία δεν ταυτίζεται με ένα pathway.

- Ένα κυτταρικό στοιχείο (**Cellular component**) είναι ακριβώς ένα συστατικό του κυττάρου, αλλά με την προϋπόθεση ότι το στοιχείο αυτό είναι μέρος κάποιου μεγαλύτερου αντικειμένου, το οποίο μπορεί να είναι μια ανατομική δομή π.χ. "rough endoplasmic reticulum" ή "nucleus" ή ένα σύνολο γονιδιακών παραγώγων, π.χ. "ribosome," "proteasome" ή heterodimeric protein.

Τα Molecular function, Biological process και Cellular component είναι γνωρίσματα των γονιδιακών παραγώγων. Κατά κανόνα μπορεί να είναι ανεξάρτητα ή να κληρονομούν χαρακτηριστικά από άλλα; στην πράξη είναι ανεξάρτητα, με άλλα

λόγια οι τρεις γράφοι function, process και cellular component, είναι ανεξάρτητοι/μη συνδεδεμένοι.

Οι σχέσεις μεταξύ των γονιδιακών παραγώγων και των μοριακών λειτουργιών, βιολογικών διαδικασιών και κυτταρικών στοιχείων είναι πολλά προς πολλά.

5.3.2 Συντακτικό Οντολογιών

Γραμμές σχολίων: Οι γραμμές που ξεκινούν με ! θεωρούνται σχόλια.

Γραμμές οι οποίες ξεκινούν με $ αναπαριστούν είτε το πεδίο είτε ιδιότητες των οντολογιών (π.χ. $ Versioning)

Οι πρώτες γραμμές σε κάθε αρχείο συνήθως κρατούν πληροφορίες σχετικά με την έκδοση, την ημερομηνία τελευταίας ενημέρωσης, την πηγή των αρχείων, το πεδίο του αρχείου και τον κειμενογράφο στον οποίο δημιουργήθηκε, (βλ. παράδειγμα 1)

Σχέσεις parent-child μεταξύ των όρων αναπαρίστανται με εσοχές:

parent_term
    child_term

Σχέση στιγμιότυπο (isa):

%term0
    %term1 % term2

Διαβάζετε ως: ο term1 είναι στιγμιότυπο του term0 και επίσης στιγμιότυπο του term2.]

Σχέση Part of:

%term0
    %term1 < term2 < term3

Διαβάζετε ως: ο term1 είναι στιγμιότυπο του term0 και **part-of** του term2 και term3.
Σύνταξη γραμμής:

```
< | % term [; db cross ref]* [; synonym:text]* [ < | % term]*
!autogenerated-by:   DAG-Edit version 1.311
!saved-by:       gwg
!date:         Thu Nov 07 14:59:21 GMT 2002
!version: $Revision: 2.619 $
!note:   file automatically generated by GO-Editor
$Gene_Ontology ; GO:0003673
 <biological_process ; GO:0008150
 %behavior ; GO:0007610
  %adult behavior ; GO:0030534
  %adult behavior (sensu Insecta) ; GO:0008044
   %response to cocaine (sensu Insecta) ; GO:0008341 % response to
cocaine ; GO:0042220
   %response to ethanol (sensu Insecta) ; GO:0045473 % response to
ethanol ; GO:0045471
   %response to ether (sensu Insecta) ; GO:0045474 % response to
ether ; GO:0045472
  %adult feeding behavior ; GO:0008343 % feeding behavior ;
GO:0007631
   %adult feeding behavior (sensu Insecta) ; GO:0030535
  %behavioral fear response ; GO:0001662
  %chemosensory behavior ; GO:0007635
  %chemosensory jump behavior ; GO:0007636
  %olfactory behavior ; GO:0042048
  %proboscis extension reflex ; GO:0007637
  %feeding behavior ; GO:0007631 ; synonym:drinking ; synonym:eating
```

Παράδειγμα 1 Μέρος της οντολογίας Biological Process

Κεφάλαιο 6: Γραφικά μοντέλα

Τα γραφικά μοντέλα είναι ένας συνδυασμός της θεωρίας πιθανοτήτων και της θεωρίας γράφων. Τα γραφικά μοντέλα παρέχουν ένα φυσικό εργαλείο το οποίο μπορεί να χρησιμοποιηθούν σε πολλά προβλήματα που εμφανίζονται στα εφαρμοσμένα μαθηματικά και στην εφαρμοσμένη μηχανική, και έχουν να κάνουν με την αβεβαιότητα και πολυπλοκότητα συστημάτων. Ειδικότερα χρησιμοποιούνται όλο και περισσότερο στον σχεδιασμό και την ανάλυση αλγορίθμων μηχανικής μάθησης. Η ιδέα πίσω από ένα γραφικό μοντέλο είναι η έννοια του "διαμορφώσιμου" όπου ένα σύνθετο σύστημα χτίζεται με συνδυασμό απλούστερων συστημάτων. Η θεωρία πιθανοτήτων παρέχει το υπόβαθρο για τον συνδυασμό των επιμέρους συστημάτων, εξασφαλίζοντας ότι το σύστημα είναι συνολικά συνεπές, και παρέχει τους τρόπους για να διασυνδεθεί το μοντέλο στα δεδομένα. Η θεωρητική πλευρά των γραφικών μοντέλων, που αναφέρεται στην θεωρία γράφων, παρέχει και μια διαισθητικά ελκυστική διεπαφή με βάση την οποία μπορούν να μοντελοποιήσουν τα αλληλεπιδρώντα σύνολα μεταβλητών, καθώς επίσης και τη δομή δεδομένων που παρέχει το πλαίσιο για τον σχεδιασμό αποδοτικών αλγορίθμων. Πολλά από τα στοχαστικά συστήματα που μελετώνται σε τομείς όπως οι στατιστική, η εφαρμοσμένη μηχανική, η θεωρία της πληροφορίας, η αναγνώριση προτύπων κλπ είναι ειδικές περιπτώσεις των γενικών φορμαλισμών που παράγονται από την θεωρία των γραφικών μοντέλων. Σε αυτά περιλαμβάνονται τα mixture models, η factor analysis, τα hidden Markov models, τα Kalman filters και τα Ising μοντέλα. Στο Κεφάλαιο αυτό γίνεται μια συνοπτική επισκόπηση των Μπεϋζιανών δικτύων και ακολουθεί μια βιβλιογραφική έρευνα στους αλγορίθμους συμπερασματολογίας και εκμάθησης.

6.1 Boolean δίκτυα

Μερικά από τα πιο απλά γραφικά μοντέλα είναι τα Boolean δίκτυα, τα οποία έχουν χρησιμοποιηθεί ήδη σε πολλές εφαρμογές μια εκ των οποίων και στην μοντελοποίηση απλών γενετικών δικτύων και κυρίως των pathways.

Τα Boolean δίκτυα μπορούν να μοντελοποιήσουν τις αλληλεπιδράσεις μεταξύ των γονιδίων ως ένα διάγραμμα καταστάσεων. Το επίπεδο γονιδιακής έκφρασης μπορεί

να θεωρηθεί είτε ως 0 είτε ως 1, κάτι το οποίο διαμορφώνει την κατάσταση των γονιδίων. Ένα Boolean δίκτυο στο οποίο τα γονίδια είναι οι κόμβοι και οι αλληλεπιδράσεις μεταξύ των γονιδίων είναι οι ακμές μεταξύ των κόμβων μπορεί να αποτελέσει ένα καλό μοντέλο. Το δίκτυο αλληλεπιδράσεων στη συνέχεια μετατρέπεται σε Boolean κανόνες, κάνοντας χρήση γνωστών αλγορίθμων εκπαίδευσης.

Ένα Boolean δίκτυο $G\ (V, F)$ αποτελείται από ένα σύνολο κόμβων B που αντιπροσωπεύουν τα γονίδια και ένα σύνολο F Boolean συναρτήσεων, όπου η κάθε Boolean συνάρτηση παίρνει ως είσοδο συγκεκριμένους κόμβους και αναθέτει τιμή σε ένα συγκεκριμένο κόμβο. Το πρότυπο έκφρασης ενός γονιδίου είναι μια συνάρτηση στο {0, 1}, (οι καταστάσεις του γονιδίου). Σε ένα Boolean δίκτυο το πρότυπο έκφρασης στον χρόνο $t+1$ καθορίζεται από την συνάρτηση $f \in F$ και από την έκφραση του γονιδίου στον χρόνο t.

6.2 Εισαγωγή στα Μπεϋζιανά δίκτυα

Τα Μπεϋζιανά δίκτυα είναι μια ειδική περίπτωση των γραφικών μοντέλων στα οποία οι κόμβοι αντιπροσωπεύουν τυχαίες μεταβλητές, και η έλλειψη τόξων αντιπροσωπεύει υπό όρους ανεξαρτησία.

Ένα Μπεϋζιανό δίκτυο αποτελείται από δύο μέρη

- ένα DAG (Directed Acyclic Graph)
- και ένα σύνολο παραμέτρων

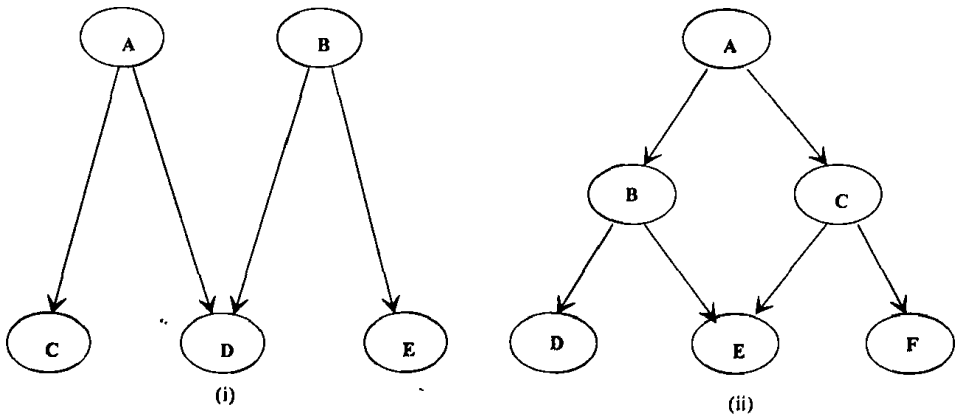
Το DAG, (κατευθυνόμενο άκυκλο γράφημα), αποτελείται από ένα σύνολο κόμβων που αντιπροσωπεύουν τις τυχαίες μεταβλητές και τα τόξα που αντιπροσωπεύουν τη σχέση μεταξύ των τυχαίων μεταβλητών. Το σύνολο παραμέτρων για κάθε κόμβο στο DAG ορίζει μια από κοινού κατανομή (joint distribution) για τις τυχαίες μεταβλητές. Η από κοινού κατανομή παραγοντοποιείται κατά μήκος της γραφικής παράστασης χρησιμοποιώντας τις υποθέσεις της υπό όρους ανεξαρτησίας [Pearl et al. 1998], [Heckerman et al. 1995].

Τα Μπεϋζιανά δίκτυα είναι πολύ καλά μοντέλα για σχέσεις αιτιότητας-αιτιατού, όπου τα κατευθυνόμενα τόξα αντιπροσωπεύουν την αιτία. Δεδομένου ότι μαρτυρίες

(evidences) μπορούν να οριστούν σε καθέναν από τους κόμβους ενός τόξου, αυτά τα δίκτυα χρησιμοποιούνται και για causal reasoning (γνωστές αιτίες με άγνωστα αποτελέσματα) και για diagnostic reasoning (γνωστά αποτελέσματα με άγνωστες αιτίες).

Ενώ τα Boolean δίκτυα, όπως προαναφέρθηκε πιο πάνω, είναι πιο διαισθητικά μοντέλα, αποτυγχάνουν να μοντελοποιήσουν time-based σχέσεις γονιδιακών δικτύων. Τα Bayesian δίκτυα παρέχουν ένα πιο στοχαστικό τρόπο αναπαράστασης των παραπάνω δικτύων, ενώ τα δυναμικά Μπεϋζιανά δίκτυα είναι ίσως τα αποτελεσματικότερα μαθηματικά μοντέλα για τα γενετικά δίκτυα επειδή όχι μόνο μοντελοποιούν τις υπό όρους εξαρτήσεις αλλά επίσης αντιπροσωπεύουν τις αλλαγές στις εκφράσεις γονιδίων κατά τη διάρκεια του χρόνου [Friedman et al. 1998]. Σε αυτό το κεφάλαιο θα εξετάσουμε τις θεμελιώδεις ιδιότητες των Μπεϋζιανών δικτύων. Επίσης γίνεται μια βιβλιογραφική έρευνα για τους αλγορίθμους συμπεράσματος και εκμάθησης για των Μπεϋζιανών δικτύων.

Τα Μπεϋζιανά δίκτυα μπορούν να είναι είτε απλά-συνδεμένα *singly-connected* ή πολλαπλά-συνδεμένα *multiply-connected*. Σε ένα απλά-συνδεμένο δίκτυο, υπάρχει το πολύ μια κατευθυνόμενη ακμή στο γράφο από έναν κόμβο σε οποιονδήποτε άλλο κόμβο, ενώ σε ένα multiply-connected δίκτυο, υπάρχει τουλάχιστον μια πορεία στον γράφο που περιλαμβάνει τον ίδιο κόμβο περισσότερες από μια φορές (*loop*). (Σχήμα 6-1)



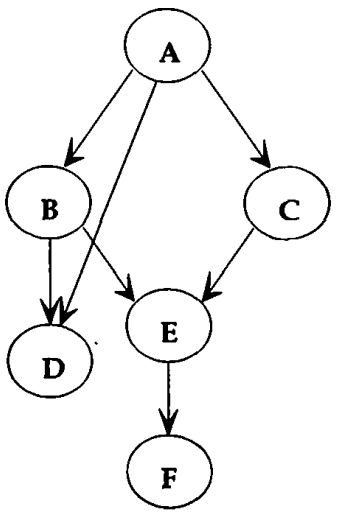
Σχήμα 6-1 Ένα απλά συνδεδεμένο δίκτυο. ii) Ένα πολλαπλά-συνδεδεμένο δίκτυο.

6.3 Ορισμοί, Έννοιες, Θεωρήματα

Directed Acyclic Graphs: Ένας κατευθυνόμενος άκυκλος γράφος (DAG) είναι ένας κατευθυνόμενος γράφος που δεν περιέχει κανένα κύκλο. Ένας κατευθυνόμενος γράφος είναι άκυκλος, εάν και μόνο εάν υπάρχει $(I-A)^{-1}$ και $(I-A)^{-1} \geq 0$, όπου I είναι ο μοναδιαίος πίνακας και A είναι ο πίνακας γειτνίασης [Ravi et al. 1985]

Causal Networks: Ο Neapolitan [Neapolitan et al. 1990] ορίζει ένα αιτιολογικό δίκτυο (causal network) ως εξής: "Έστω V είναι ένα πεπερασμένο σύνολο από μεταβλητές ορισμένες στον ίδιο χώρο πιθανοτήτων, έστω (V,G,P) είναι η από κοινού κατανομή πιθανοτήτων (joint probability distribution), και έστω G=(V,E) ένα DAG. Για κάθε $v \in V$, έστω $c(v) \subseteq V$ ένα σύνολο από όλους τους γονείς του v και $d(v) \subseteq V$ το σύνολο όλων των απογόνων του v. Επιπλέον, για $v \in V$, έστω $a(v) \subseteq V$ είναι V-(d(v)∪{v}), δηλαδή, το σύνολο των απογόνων μεταβλητών στο V εξαιρώντας τους απογόνους των v και v'. Υποθέτουμε ότι για κάθε υποσύνολο W του a(v), τα W και v είναι υπό συνθήκη ανεξάρτητα δοθέντος του c(v); δηλαδή, εάν $P(c(v))>0$ τότε $P(v \mid c(v))=0$ ή $P(W \mid c(v))=0$ ή $P(v \mid W \cup c(v))=P(v \mid c(v))$. Τότε το C=(V,E,P) ονομάζεται causal network. Το σύνολο c(v) ονομάζεται σύνολο αιτιών (γονείς) του v."

Για παράδειγμα, για το DAG στο Σχήμα 6-2 c(E)={B,C,}, d(B)={D,E,F} και a(c)={A,B,D}.



Σχήμα 6-2 Ένα κατευθυνόμενο άκυκλο γράφημα

- **Υπό συνθήκη ανεξαρτησία (Conditional Independence)**: Έστω U = {A, B...} είναι ένα πεπερασμένο σύνολο από μεταβλητές με διακριτές τιμές. Έστω *P*(.) η από κοινού πιθανότητα στις μεταβλητές *U*, και έστω *X*, *Y*, και *Z* τρία υποσύνολα του *U*. Τα *X* και *Y* λέμε ότι είναι υπό συνθήκη ανεξάρτητα (**conditionally independent**) δοθέντος του *Z* εάν P(x | y,z)=P(x | z) όπου P(y,z)>0. Τα *X* και *Y* λέγονται υπό συνθήκη ανεξάρτητα στο Z [Neapolitan et al. 1990].

Chain Rule: Έστω U = $\{A_1, A_2, A_3, \ldots, A_n\}$ ένα σύνολο από μεταβλητές. Ο «κανόνας αλυσίδας» (chain rule) παρέχει μια ενιαία αναπαράσταση της από κοινού πιθανότητας P(U) = $P(A_1, A_2, A_3 \ldots, A_n)$ για να είναι ο υπολογισμός της πιθανότητας ευκολότερος. Εάν ο πίνακας των από κοινού πιθανοτήτων *P*(*U*) είναι γνωστός, τότε οι πιθανότητες $P(A_i)$ μπορούν να υπολογιστούν τόσο εύκολα όσο και οι πιθανότητες $P(A_i \mid e)$, όπου το *e* είναι η μαρτυρία-evidence. Απ' την άλλη, εάν ο αριθμός των μεταβλητών στο δίκτυο αυξηθεί, το *P*(*U*) μεγαλώνει εκθετικά. Κατά συνέπεια μια πιο συμπαγής αναπαράσταση του *P*(*U*) είναι απαραίτητη: κάποιου είδους δεσμευμένη πληροφορία από την οποία το *P*(*U*) μπορεί να υπολογισθεί αν είναι απαραίτητο [Neapolitan et al. 1990]. Μια τέτοια αναπαράσταση βρίσκεται σε ένα Bayesian δίκτυο πάνω στο *U*. Το *P*(*U*) μπορεί να υπολογιστεί από τις υπό συνθήκη πιθανότητες σε ένα Bayesian δίκτυο εάν γνωρίζουμε τις υπό συνθήκη ανεξαρτησίες για το *U*. Το παρακάτω θεώρημα περιγράφει αυτή την αναπαράσταση.

Θεώρημα Chain Rule: Έστω ότι BN είναι ένα Bayesian δίκτυο στο U = $\{A_1, A_2, A_3, \ldots, A_n\}$, η από κοινού κατανομή πιθανότητας P(U) είναι το γινόμενο όλων των υπό συνθήκη πιθανοτήτων οι οποίες ορίζονται από το BN:

$$P(U) = \prod_i P(A_i / pa(A_i)) \qquad (6.1)$$

όπου pa(A_i) είναι οι γονείς του A_i.

Θεώρημα του Bayes: Το θεώρημα του Bayes [Cowell et al. 1999] είναι η βάση για τον υπολογισμό και την ενημέρωση των πεποιθήσεων σε ένα δίκτυο: "Έστω (V, G, P) είναι ένας χώρος πιθανοτήτων και {E1, E2, ..., En} είναι ένα σύνολο από αμοιβαία γεγονότα στο G τέτοια ώστε για κάθε $1 \le i \le n$, P(E_i)>0. Για κάθε E ∈ G τέτοιο ώστε P(E) > 0, έχουμε:

$$P(E_j \mid E) = \frac{P(E \mid E_j)P(E_j)}{\sum_{j=1}^{n} P(E \mid E_i)P(E_i)}, \quad 1 \le j \le n \tag{6.2}$$

Εάν E και E' είναι δύο οποιαδήποτε γεγονότα τέτοια ώστε $P(E)$ και $P(E')$ είναι και τα δύο θετικά, τότε η παρακάτω ισότητα προκύπτει από το θεώρημα

$$P(E / E') = \frac{P(E'|E)P(E)}{P(E')}. \tag{6.3}$$

Το θεώρημα του Bayes είναι πολύ σημαντικό επειδή διαμορφώνει την βάση ανάλυσης της συνολικής από κοινού πιθανότητας ενός δικτύου σε ένα γινόμενο από υπό συνθήκη και marginal πιθανότητες. Οι τρόποι για την ανάλυση της μαρτυρίας (evidence), περιγράφεται παρακάτω και βασίζεται στο ίδιο θεώρημα.

6.4 Μπεϋζιανοί αλγόριθμοι συμπερασματολογίας

Ένας κόμβος σε ένα Μπεϋζιανό δίκτυο λέγεται ότι είναι «ορισμένος άμεσα» (instantiated) εάν προσδιορίζεται σε αυτό κάποια μαρτυρία. Όταν ένας κόμβος ορίζεται άμεσα, οι πεποιθήσεις των άλλων κόμβων πρέπει να ενημερώνονται. Αλγόριθμοι συμπερασματολογίας χρησιμοποιούνται για ενημέρωση και αναπαραγωγή πεποιθήσεων μέσα στο δίκτυο. Η φύση των αλγορίθμων συμπερασματολογίας εξαρτάται από τον τύπο του δικτύου (απλά-συνδεμένος ή πολλαπλά-συνδεμένος).

6.4.1 Message-Passing Αλγόριθμος για Single Connected δίκτυα

Η μέθοδος του Pearl "Message-Passing" μεταξύ των κόμβων χρησιμοποιείται για την ενημέρωση και τη διάδοση των πεποιθήσεων στα single connected Μπεϋζιανά δίκτυα [Pearl et al. 1998] [Neapolitan et al. 1990]. Ο κόμβος που λαμβάνει μια μαρτυρία υπολογίζει μια νέα διανυσματική πιθανότητα χρησιμοποιώντας τον τύπο: $Bel(n)=\alpha*\lambda(n)*\pi(n)$, όπου n χρησιμοποιείται για τον κόμβο, α είναι μια σταθερά, λ είναι η μαρτυρία που έρχεται στον κόμβο από τους κόμβους παιδιά του, και π είναι η μαρτυρία που έρχεται στον κόμβο από τους γονείς του.

Μόλις υπολογιστεί το διάνυσμα πεποίθησης του κόμβου n, ο κόμβος στέλνει λ μηνύματα στους γονείς του και π μηνύματα στα παιδιά του. Οι κόμβοι που

λαμβάνουν τα λ και π μηνύματα αρχίζουν την δική τους διαδικασία ενημέρωσης και αναπαραγωγής. Αυτή η διαδικασία συνεχίζεται έως ότου αναλυθεί η επίδραση της μαρτυρίας του κόμβου *n* από κάθε άλλο κόμβο. Σε αυτήν την μέθοδο, εκτός από το διάνυσμα πεποίθησης, σε κάθε κόμβο ορίζεται ένα $p_i(\pi)$ διάνυσμα και ένα λ διάνυσμα.

6.4.2 Message-Passing αλγόριθμος - για multiply connected δίκτυα

Η μέθοδος των Lauritzen και Spiegelhalter "message-passing" μεταξύ κλίκων (cliques) χρησιμοποιείται για να διαδώσει και να ενημερώσει τις πεποιθήσεις σε ένα multiply connected δίκτυο [Cowell et al. 1999]. Μια κλίκα είναι ένα υποσύνολο ενός μη κατευθυνόμενου γράφου που είναι *complete* και *maximal*. Οι βρόχοι στα multiply connected γραφήματα αφαιρούνται με τη μετατροπή του γράφου σε ένα *δέντρο κλίκων*.

Αρχικά στον γράφο εφαρμόζεται μια διαδικασία γνωστή ως *"moralized"*, αυτό γίνεται με το να σχεδιάζονται ακμές μεταξύ των κόμβων γονέων οι οποίοι έχουν κοινά παιδιά και καμία διασύνδεση μεταξύ των γονέων. Ο προκύπτων γράφος είναι τριγωνικός από αυτόν παίρνουμε ένα σύνολο από κλίκες. Η συνολική διαδικασία, που καταλήγει σε ένα δέντρο κλικών όπως περιγράφεται λεπτομερώς στο [Cowell et al. 1999].

Κάθε κλίκα σε ένα multiply-connected γράφημα περιέχει ένα προϊόν το οποίο ονομάζεται *"clique potential"*. Αυτό σχετίζει τις υπό συνθήκη πιθανότητες των κόμβων που ανήκουν σε εκείνη τη κλίκα. Όταν μια μαρτυρία για ένα σύνολο κόμβων φθάσει, σε όλα τα "clique potentials" που περιέχουν "instantiated" (άμεσα ορισμένους) κόμβους ανατίθεται μια τιμή. Η διαδικασία για ενημέρωση και αναπαραγωγή πιθανοτήτων στη συνέχεια παράγει λ μηνύματα από όλα τα φύλλα του δέντρου των κλίκων προς την ρίζα (*Propagate - evidence*). Μόλις φθάσουν στη ρίζα, π μηνύματα διαδίδονται από τη ρίζα σε όλους τους κόμβους απογόνων της προς τα φύλλα (*Distribute-evidence*). Κατά τη διάρκεια αυτής της διαδικασίας, κάθε κλίκα υπολογίζει το δυναμικό της κάθε φορά που λαμβάνει μηνύματα. Μόλις τελειώσει το μήνυμα ο μηχανισμός τελειώνει, οι τιμές πεποίθησης για τους μεμονωμένους κόμβους μπορούν να υπολογιστούν από τα αντίστοιχα δυναμικά των κλίκων.

6.5 Εκπαίδευση Μπεϋζιανών δικτύων

Τα Μπεϋζιανά δίκτυα μπορούν να δημιουργηθούν από ειδικούς κάνοντας χρήση της γνώσης που υπάρχει σε ένα πεδίο ή δημιουργώντας σχέσεις από δεδομένα του συγκεκριμένου πεδίου χρησιμοποιώντας αυτόματες τεχνικές μάθησης, ή με έναν συνδυασμό αυτών των τεχνικών [Heckerman et al. 1995]. Η αυτοματοποιημένη κατασκευή δικτύων από δεδομένα περιλαμβάνει:

- Προσδιορισμό των σχετικών μεταβλητών και υπολογισμός της σχέσης αιτιότητας - αποτελέσματος μεταξύ αυτών των μεταβλητών
- Ποσοτικοποίηση αυτών των εξαρτήσεων ως υπό όρους εξαρτήσεις και υπολογισμό αυτών των τιμών.

Με άλλα λόγια, η εκπαίδευση των Μπεϋζιανών δικτύων μπορεί να θεωρηθεί ως συνδυασμός της μάθησης παραμέτρων και μάθησης της δομής [Cowell et al. 1999], [Heckerman et al. 1995], [Krause et al. 1996]. Η μάθηση παραμέτρων περιλαμβάνει την εκτίμηση των υπό όρους πιθανοτήτων στο δίκτυο, ενώ η εκπαίδευση της δομής περιλαμβάνει τον προσδιορισμό της τοπολογίας (συνδέσεις) του δικτύου. Η διαδικασία εκπαίδευσης των Μπεϋζιανών δικτύων παίρνει διάφορες μορφές εάν η δομή του δικτύου είναι γνωστή και εάν οι μεταβλητές έχουν παρατηρηθεί. Η δομή του δικτύου μπορεί να είναι *γνωστή* ή *άγνωστη*, και οι μεταβλητές μπορεί να έχουν παρατηρηθεί ή να είναι *κρυφές* σε όλα ή σε μερικά από τα δεδομένα. Η τελευταία διάκριση μπορεί επίσης να εκφραστεί και ως *πλήρη* ή *ελλιπή* δεδομένα. Συνεπώς, υπάρχουν τέσσερις περιπτώσεις για την εκπαίδευση των Μπεϋζιανών δικτύων από δεδομένα, γνωστή δομή και μη-κρυμμένες μεταβλητές, άγνωστη δομή και μη-κρυμμένες μεταβλητές, γνωστή δομή και κρυμμένες μεταβλητές, και άγνωστη δομή και κρυμμένες μεταβλητές [Cowell et al. 1999].

Οι τέσσερις περιπτώσεις εκπαίδευσης των Μπεϋζιανών δικτύων αναλύονται στις ακόλουθες παραγράφους.

6.5.1 Γνωστή δομή δικτύων και γνωστές μεταβλητές (πλήρη δεδομένα)

Σε αυτήν την περίπτωση η δομή των δικτύων είναι γνωστή και οι παράμετροι πρέπει να υπολογιστούν. Για την εκτίμηση των παραμέτρων χρειάζεται ο υπολογισμός των πινάκων των υπό όρους (CPT) πιθανοτήτων. Αυτό μπορεί να γίνει με χρήση

προσεγγιστικών μεθόδων όπως είναι οι Maximum Likelihood Estimation (MLE) και Bayesian Estimation [Spiegelhalter et al. 1993] [Cowell et al. 1999].

Το θεώρημα μέγιστης πιθανοφάνειας *Maximum Likelihood Estimate (MLE),* περιλαμβάνει την μεγιστοποίηση της συνάρτησης της πιθανοφάνειας για τις παραμέτρους Θ. Η MLE είναι ένας από τους πιο χρησιμοποιούμενους εκτιμητές στην στατιστική.

Οι εκτιμώμενες παράμετροι υπολογίζονται από τον τύπο:

$$\hat{\theta}_{x|u} = \frac{N(x,u)}{N(u)} \tag{6.4}$$

όπου N(x,u) είναι μια κατάλληλη στατιστική για το γεγονός X=x και τους γονείς U=u. Η κατάλληλη στατιστική είναι μια συνάρτηση των δεδομένων η οποία συνοψίζει την σχετική πληροφορία για τον υπολογισμό της συνάρτησης πιθανοφάνειας.

Μια άλλη τεχνική εκτίμησης παραμέτρου που ακολουθεί την Μπεϋζιανή φιλοσοφία, υποθέτει ότι υπάρχει μια άγνωστη αλλά σταθερή παράμετρος θ, και υπολογίζει την παράμετρο με κάποια βεβαιότητα, π.χ. η παράμετρος θ αντιμετωπίζεται σαν τυχαία μεταβλητή και ορίζεται μια κατανομή P(θ).Από το θεώρημα του Bayes έχουμε,

$$P(\theta \mid D) = \frac{P(D \mid \theta) P(\theta)}{P(D)} \tag{6.5}$$

όπου ο πρώτος όρος στον αριθμητή είναι η πιθανοφάνεια, η δεύτερη είναι η prior για την παράμετρο και ο τρίτος είναι ένας παράγοντας κανονικοποίησης - η marginal πιθανότητα των δεδομένων. Οι συνηθέστερες priors που χρησιμοποιούνται είναι οι Dirichlet priors [Heckerman et al. 1995],[Ramoni et al. 1997],[Ramoni et al. 1997] επειδή οι περισσότεροι από τους υπολογισμούς μπορούν να προκύψουν σε κλειστή μορφή. Ως εκ τούτου με βάση το θεώρημα του Bayes, η εκτίμηση των παραμέτρων υπολογίζεται ως:

$$P(\theta / D) = \hat{\theta}_{x|u} = Dir(\alpha_1 + N(x_1,u),...,\alpha_k + N(x_k,u) \tag{6.6}$$

όπου Dir είναι η Dirichlet κατανομή πάνω στις παραμέτρους.

Εάν τα δεδομένα παρήχθησαν πραγματικά από τη συγκεκριμένη δομή, τότε και οι δύο μέθοδοι συγκλίνουν ασυμπτωτικά στις σωστές παραμέτρους. Εάν όχι, τότε συγκλίνουν στη κατανομή με δεδομένη τη δομή που είναι πιο "κοντά" στη κατανομή από την οποία τα δεδομένα παρήχθησαν.

6.5.2 Άγνωστη δομή και γνωστές μεταβλητές (πλήρη δεδομένα)

Η βασική ιδέα εδώ είναι να αναδημιουργηθεί η τοπολογία του δικτύου από τις πλήρως παρατηρήσιμες μεταβλητές ακολουθούμενη από την εκτίμηση των παραμέτρων. Αυτή η μέθοδος είναι χρήσιμη σε εφαρμογές data-mining, όπου θα πρέπει να ερμηνεύσουμε μεγάλα ποσά δεδομένων.

Σε αυτήν την περίπτωση η Μπεϋζιανή κατασκευή δικτύων εμπίπτει σε δύο κατηγορίες:

- Τα Μπεϋζιανά δίκτυα μπορούν να αντιμετωπισθούν ως δομές που κωδικοποιούν την κοινή κατανομή ορισμένων γνωρισμάτων. Αυτό προτείνει ότι το καλύτερο Μπεϋζιανό δίκτυο (BN) είναι αυτό που καλύτερα προσαρμόζεται στα δεδομένα και οδηγεί σε Search & Score αλγορίθμους εκπαίδευσης οι οποίοι ψάχνουν μια δομή που μεγιστοποιεί μια score συνάρτηση.

- Εναλλακτικά, η δομή των BN μπορεί να αντιμετωπισθεί ως μια δομή που κωδικοποιεί μια ομάδα από υπό συνθήκη ανεξαρτησίες μεταξύ των κόμβων [Pearl et al. 1998]. Σε αυτήν την περίπτωση, ένα BN εκπαιδεύεται με τον προσδιορισμό των σχέσεων της υπό συνθήκη ανεξαρτησίας (CI) μεταξύ των κόμβων. Αυτές οι CI σχέσεις βρίσκονται από τις στατιστικές δοκιμές οι οποίες δρουν ως περιορισμοί στην κατασκευή του δικτύου. Αυτοί οι αλγόριθμοι αναφέρονται ως Dependency Analysis ή Conditional Independence ή Constraint αλγόριθμοι [Cheng et al. 1997] [Cowell et al. 1999].

Και οι δύο κατηγορίες αλγορίθμων έχουν τα πλεονεκτήματα και τα μειονεκτήματά τους. Οι Search & Score αλγόριθμοι έχουν λιγότερη χρονική πολυπλοκότητα στη χειρότερη περίπτωση (όταν το DAG συνδέεται πυκνά), αλλά μπορεί να μην βρουν την καλύτερη λύση λόγω της ευρετικής φύσης τους. Οι βασισμένοι σε περιορισμούς αλγόριθμοι είναι συνήθως ασυμπτωτικά σωστοί [Cooper et al. 1992], αλλά όπως ο

Cooper et.al. [Cooper et al. 1992] επισήμανε, έχοντας μεγάλα σύνολα συνθηκών οι αλγόριθμοι μπορεί να είναι αναξιόπιστοι εκτός και αν ο όγκος των στοιχείων είναι τεράστιος.

6.5.2.1 Score-based Μέθοδοι

K2 Αλγόριθμος - Cooper and Herskovits [Cooper et al. 1992]

- K2 είναι ένας *search and scoring* αλγόριθμος για εκπαίδευση BNs. Εφαρμόζει μια Μπεϋζιανή μέθοδο βαθμολογίας (score). Εκτός από το σύνολο δεδομένων, ο αλγόριθμος απαιτεί μια διάταξη κόμβων σαν είσοδο. Δοθέντος ενός συνόλου δεδομένων D, ο K2 στοχεύει στην εύρεση του Μπεϋζιανού δικτύου B που μεγιστοποιεί την πιθανότητα *P(B/D)*.

$$P(B \mid D) = \frac{P(D \mid B)P(B)}{P(D)} \tag{6.7}$$

Η πιθανότητα $P(D \mid B)$ υπολογίζεται:

$$P(D \mid B) = \int P(D/B,\theta_B)P(B \mid \theta_B)d\theta_B \tag{6.8}$$

Σε αυτήν την περίπτωση, όσο περισσότερες παραμέτρους έχουμε σε τόσο περισσότερες μεταβλητές ολοκληρώνουμε. Ως εκ τούτου αυτή η προσέγγιση ευνοεί τα πρότυπα με λίγες παραμέτρους.

HGC Αλγόριθμος -Heckerman, Geiger and Chickering [Heckerman et al. 1996]

Αυτός ο αλγόριθμος χρησιμοποιεί επίσης ένα Μπεϋζιανό score. Ο αλγόριθμος εστιάζεται στον συνδυασμό της γνώσης των χρηστών και διαφόρων στατιστικών στοιχείων. Οι συγγραφείς προσδιορίζουν δύο σημαντικές ιδιότητες της score-based μεθόδου - τα *event equivalence* και *parameter modularity*. Αυτές οι ιδιότητες, όταν συνδυάζονται με score προσεγγίσεις, απλοποιούν την κωδικοποίηση της εκ των προτέρον γνώσης του χρήστη.

BENEDICT Αλγόριθμος - Acid, Campos [Acid et al. 1996]

Αυτός ο αλγόριθμος απαιτεί διάταξη κόμβων και χρήσεις ένα διαφορετικό score εντροπίας. Αφού βρεθεί η δομή με χρήση μιας ευρετικής μεθόδου αναζήτησης, ο αλγόριθμος αναλύει τις υπό συνθήκη ανεξαρτησίες που υπονοούνται στη δομή με τη χρησιμοποίηση της έννοιας d-separation [Pearl et al. 1998]. Κατόπιν ο αλγόριθμος υπολογίζει τη διαφορά μεταξύ των υπονοούμενων υπό συνθήκη ανεξαρτησιών και των πραγματικών υπό συνθήκη ανεξαρτησιών στα δεδομένα. Το score εντροπίας είναι το άθροισμα των αποτελεσμάτων μιας ομάδας CI δοκιμών.

CB Αλγόριθμος -Singh, Valtorta [46]

Αυτός ο αλγόριθμος αντιμετωπίζει το ζήτημα δεδομένου ότι η διάταξη κόμβων πρέπει να υπάρχει. Είναι μια υβριδική μέθοδος που χρησιμοποιεί τόσο τη μέθοδο ανάλυσης εξαρτήσεων όσο και την ***Search & Score*** μέθοδο. Στη φάση I του αλγορίθμου χρησιμοποιείται ένα CI test για να παραχθεί ένας μη κατευθυνόμενος γράφος, ακολουθούμενος από διευθυνσιοδότηση των ακμών για να βρεθεί μια διάταξη κόμβων. Η φάση II του αλγορίθμου χρησιμοποιεί μια τροποποιημένη έκδοση του K2 αλγορίθμου για να εκπαιδεύσει το δίκτυο.

Lam and Bacchus Αλγόριθμος - Lam and Bacchus [Lam et al. 1994]

Ο Lam-Bacchus αλγόριθμος χρησιμοποιεί ένα score βασισμένο στο minimum description length (MDL), που προσπαθεί να προσφέρει ένα tradeoff μεταξύ της απλότητας του δικτύου και της συσχέτισης με τα δεδομένα. Η σημασία αυτής της εργασίας είναι ότι δεν απαιτεί μια διάταξη κόμβων, και ότι μπορεί να προσανατολίσει τις ακμές χρησιμοποιώντας μια απλή search & scoring τεχνική. Το MDL score μπορεί να αντιμετωπιστείτε ως προσέγγιση του Bayesian score. Το MDL score και ο αλγόριθμος εκπαίδευσης περιγράφεται στο επόμενο κεφάλαιο.

Αλγόριθμος Suzuki- Suzuki [Suzuki et al. 1996]

Ο αλγόριθμος Suzuki επίσης χρησιμοποιεί το MDL score για να εκπαιδεύσει ένα BN. Αυτή η μέθοδος, η οποία απαιτεί μια διάταξη κόμβων, χρησιμοποιεί μια "branch and bounds" τεχνική και εγγυάται μια βέλτιστη δομή. Η "branch and bound" τεχνική υπολογίζει έναν κατώτερο όριο αφότου προστεθεί μια ακμή στο δίκτυο και καθορίζει εάν είναι απαραίτητη η περαιτέρω αναζήτηση σε αυτόν τον κλάδο. Ο

αλγόριθμος αποδίδει καλά σε μικρά μεγέθη δεδομένων, και γίνεται λιγότερο αποδοτικός όσο τα δεδομένα αυξάνονται.

6.5.2.2 Μέθοδοι βασισμένες σε περιορισμούς

BN Power Constructor – Cheng et.al [Cheng et al. 1997]

Οι Cheng et.al [Cheng et al. 1997] πρότειναν έναν αλγόριθμο βασισμένο σε υπό συνθήκη ανεξαρτησίες που κάνει ανάλυση των εξαρτήσεων με χρήση αρχών από την θεωρία της πληροφορίας (information theory) για να εκπαιδεύσει την δομή των BN. Ο αλγόριθμός τους επεκτείνει το αλγόριθμο των Chow and Liu για κατασκευή δέντρων [Chow et al. 1968]. Στην πρώτη φάση, ο αλγόριθμος υπολογίζει την αμοιβαία πληροφορία (mutual information) για κάθε ζευγάρι κόμβων ως μέτρο της συνάφειας (δοκιμή CI) και δημιουργεί ένα σχέδιο του BN βασισμένο σε αυτές τις πληροφορίες. Η mutual information μεταξύ δύο μεταβλητών *Xi, Xj* ορίζεται ως

$$I(X_i, X_j) = \sum_{X_i, X_j} P(X_i, X_j) \cdot \log \frac{P(X_i, X_j)}{P(X_i) \cdot P(X_j)} \qquad (6.9)$$

Όπου $I(X_i, X_j)$ περιορίζεται από ένα κάτω όριο ε, X_i και X_j θεωρούνται ανεξάρτητα.

Στη δεύτερη φάση, ο αλγόριθμος προσθέτει ακμές όταν τα ζευγάρια των κόμβων δεν μπορούν να είναι d-separated [Pearl et al. 1998]. Στην τρίτη φάση, κάθε ακμή του γράφου εξετάζεται χρησιμοποιώντας CI test και αφαιρείται εάν οι δύο κόμβοι των ακρών μπορούν να είναι *d-separated*. Ενώ η πρώτη φάση είναι η ίδια όπως του αλγορίθμου κατασκευής δέντρων των Chow και Liu's, οι επόμενες δύο φάσεις επεκτείνονται ώστε να αναφέρονται στην κατασκευή BN.

Εάν η διάταξη των κόμβων δεν είναι γνωστή, πραγματοποιείται και μια διαδικασία προσανατολισμού των ακμών. Εάν η διάταξη των κόμβων είναι διαθέσιμη, ο αλγόριθμος απαιτεί $O(N^2)$ CI tests, διαφορετικά απαιτεί $O(N^4)$ CI tests. Άλλες μορφές γνώσης όπως η μερική διάταξη των κόμβων και άμεσες σχέσεις αιτίου και αιτιατού μπορούν επίσης να χρησιμοποιηθούν για να αυξήσουν την απόδοση του αλγορίθμου.

PC Αλγόριθμος

Αυτός ο αλγόριθμος χρησιμοποιεί επίσης τις CI test για να μάθει την δομή του BN. Μπορεί αυτόματα να προσανατολίσει τις ακμές. Είναι αποδοτικότερο κατά την κατασκευή των Μπεϋζιανών δικτύων από τα δεδομένα.

6.5.3 Γνωστή δομή και κρυμμένες μεταβλητές (ελλιπή δεδομένα)

Η εκπαίδευση των Μπεϋζιανών δικτύων με γνωστή δομή και κρυφές μεταβλητές έχει μελετηθεί από τους [Cowell et al. 1999], [Lauritzen et al. 1995]. Ένας από τους αλγορίθμους που χρησιμοποιούν είναι ο expectation maximization (EM) αλγόριθμος [Lauritzen et al. 1995] [Friedman et al. 1998]. Ο αλγόριθμος EM είναι μια επαναληπτική μέθοδος που υπολογίζει την εκτίμηση μέγιστης πιθανότητας (MLEs) και τη μέγιστο εκτίμηση της εκ των υστέρων πιθανότητας για τις παραμέτρους του δικτύου. Ο αλγόριθμος EM εναλλάσσει ένα βήμα εκτίμησης με ένα βήμα μεγιστοποίησης. Στο βήμα εκτίμησης, άγνωστες ποσότητες ανάλογα με τις ελλιπείς μεταβλητές αντικαθίστανται από τις αναμενόμενες τιμές τους με βάση την πιθανοφάνεια. Στο βήμα μεγιστοποίησης, η πιθανοφάνεια που υπολογίζεται στο βήμα εκτίμησης μεγιστοποιείται με βάση τις άγνωστες παραμέτρους, και οι προκύπτουσες εκτιμήσεις υιοθετούνται για να αντικαταστήσουν τις άγνωστες ποσότητες στο επόμενο βήμα εκτίμησης.

Ο αλγόριθμος συνεχίζεται έως ότου η διαφορά μεταξύ δύο διαδοχικών εκτιμήσεων είναι μικρότερη από ένα σταθερό κατώτατο όριο [Lauritzen et al. 1995]. Ο Lauritzen είχε αναφέρει μερικές δυσκολίες για τη χρήση του EM αλγορίθμου όπως η αργή σύγκλιση και η σύγκλιση σε τοπικά μέγιστα. Προτείνει να χρησιμοποιηθεί ο αλγόριθμος καθόδου κλίσης (gradient descent) ως πιθανή εναλλακτική λύση [Lauritzen et al. 1995].

Μια άλλη προσέγγιση, προτείνεται από τον Heckerman [Heckerman et al. 1995], και χρησιμοποιεί Gibbs sampling (GS). Ο αλγόριθμος GS παράγει τιμές για το ελλιπή δεδομένα από μερικές υπό συνθήκη κατανομές και παρέχει στοχαστικές εκτιμήσεις των μεταγενέστερων πιθανοτήτων [Ramoni et al. 1997].

Και οι δύο αλγόριθμοι GS και EM χρησιμοποιούν μια βασική στρατηγική αποκαλούμενη *missing information principle*: σύμφωνα με την οποία συμπληρώνονται οι ελλιπείς παρατηρήσεις βάσει των διαθέσιμων πληροφοριών. Δυστυχώς, αυτές οι κατά προσέγγιση μέθοδοι είναι επιρρεπείς σε λάθη όταν μικρές ή/και πολωμένες

πληροφορίες είναι διαθέσιμες για τα πρότυπα των ελλειπόντων δεδομένων[Ramoni et al. 1997].

Οι Sebastiani και Ramoni [Sebastiani et al. 1985] πρότειναν έναν άλλο αλγόριθμο που ονομάζεται Bound και Collapse (BC), που είναι μια αιτιοκρατική μέθοδος υπολογισμού των υπό συνθήκη πιθανοτήτων από ελλιπή δεδομένα. Η μέθοδος *περιορίζει* το σύνολο των πιθανών εκτιμήσεων σύμφωνων με τις διαθέσιμες πληροφορίες και τον υπολογισμό των ελάχιστων και μέγιστων εκτιμήσεων που συλλέγονται από όλες τις πιθανές συμπληρώσεις των δεδομένων.

6.5.4 Άγνωστη δομή και κρυφές μεταβλητές (ελλιπή δεδομένα)

Αυτή είναι η δυσκολότερη περίπτωση που επιλύεται επειδή η δομή των δικτύων είναι άγνωστη και οι μεταβλητές δεν είναι πλήρως παρατηρήσιμες. Υπάρχουν δύο πρόσφατα αναπτυγμένες μέθοδοι που ανακτούν την δομή των Μπεϋζιανών δικτύων με τις κρυμμένες μεταβλητές.

Ο πρώτος αλγόριθμος ονομάζεται *Structural* EM (SEM) [Russel et al. 1985] συνδυάζει τον αλγόριθμο EM, που βελτιστοποιεί τις παραμέτρους του δικτύου, με την αναζήτηση δομών για την επιλογή του μοντέλου. Η κύρια ιδέα είναι να μεγιστοποιηθεί το *expected score* των μοντέλων σε κάθε επανάληψη αντί για το πραγματικών score. Ο αλγόριθμος SEM προσπαθεί να βελτιστοποιήσει άμεσα το αληθινό Bayesian score μέσα στις επαναλήψεις του EM αντί να επιδιώξει μια ασυμπτωτική προσέγγιση.

Ο δεύτερος αλγόριθμος προτάθηκε από τους Sebastiani και Ramoni [Sebastiani et al. 1985]. Έδειξαν ότι με έναν τροποποιημένο BC αλγόριθμο, ήταν δυνατό να εκπαιδευτεί η δομή ενός δικτύου και να υπολογιστούν οι παράμετροί του.

Κεφάλαιο 7: Μεθοδολογία

7.1 Εισαγωγή

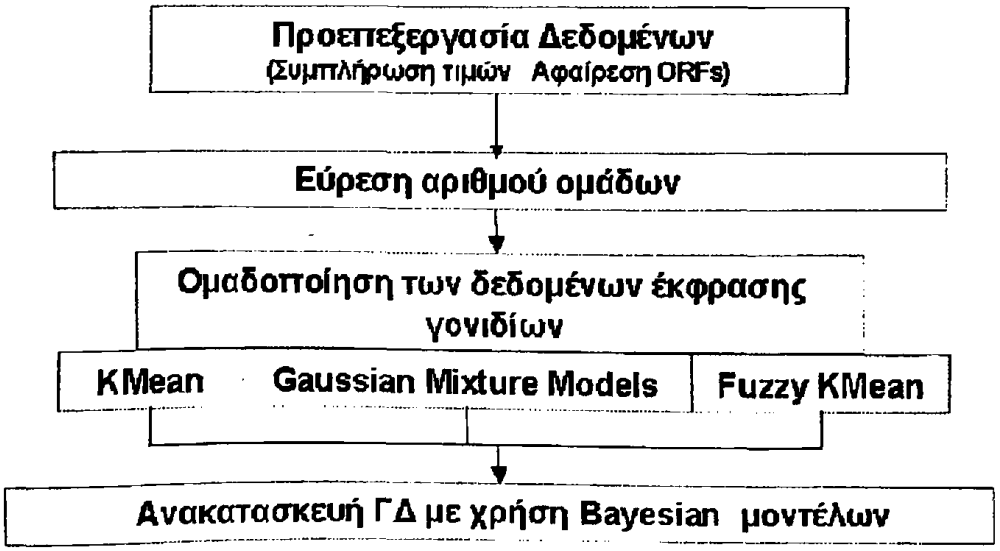
Η ανακατασκευή και η μοντελοποίηση των γενετικών δικτύων είναι ένα από τα πιο σημαντικά προβλήματα που ανήκουν στην κατηγορία των functional genomics. Ο έλεγχος της έκφρασης των γονιδίων θεωρείται μια από τις πιο ελπιδοφόρες τεχνικές για την απόκτηση γνώσης πάνω στα ρυθμιστικά δίκτυα γονιδίων. Μέχρι σήμερα έχουν αναφερθεί πολλές και διαφορετικές προσεγγίσεις για την περιγραφή και την μοντελοποίηση αυτών των δικτύων, όπως για παράδειγμα Boolean δίκτυα, μοντέλα βασισμένα σε διαφορικές εξισώσεις, και Μπεϋζιανά δίκτυα. Το κοινό μεταξύ αυτών των μεθόδων είναι η παραδοχή ότι η έκφραση κάθε γονιδίου στο δίκτυο εξαρτάται από την έκφραση μερικών άλλων γονιδίων [Kauffman et al. 1969], [D'haeseleer et al. 2000], [Pe'er et al. 2001], [Akutsu et al. 2000], [Friedman et al. 2000]. Για την αναδημιουργία ενός τέτοιου δικτύου πρέπει να απαντηθούν δύο βασικά ερωτήματα για κάθε γονίδιο στο δίκτυο: πρώτον ποια γονίδια επιδρούν σε ποία, και δεύτερον πώς επιδρούν, π.χ., θετικά, αρνητικά ή με έναν πιο σύνθετο τρόπο.

Τα περισσότερα μοντέλα γενετικών δικτύων μπορούν να περιγραφούν ως γραφικές παραστάσεις στις οποίες κάθε κόμβος αντιπροσωπεύει ένα γονίδιο και η παρουσία μιας άκρης μεταξύ δύο κόμβων δείχνει την ύπαρξη μιας αλληλεπίδρασης μεταξύ των συνδεδεμένων γονιδίων. Οι ακμές μπορούν να έχουν και διαφορετικές ερμηνείες μπορούν να σημάνουν είτε άμεσες αλληλεπιδράσεις είτε απλά παρατηρήσεις από δεδομένα, τα οποία μπορούν στη συνέχεια να είναι το αποτέλεσμα είτε των άμεσων είτε έμμεσων αλληλεπιδράσεων.

Στο κεφάλαιο αυτό περιγράφουμε μια διαφορετική προσέγγιση για την αναδημιουργία των δεδομένων των γενετικών δικτύων βασισμένων στην πρόβλεψη της έκφρασης (ή των αλλαγών της έκφρασης) ενός δεδομένου γονιδίου από την έκφραση (ή των αλλαγών της έκφρασης) άλλων γονιδίων. Η μεθοδολογία που ακολουθήθηκε παρουσιάζεται αναλυτικά σε αυτό το κεφάλαιο. Πιο συγκεκριμένα στα κεφάλαια που ακολουθούν περιγράφεται αναλυτικά κάθε βήμα της μεθοδολογίας. Αρχικά περιγράφεται το στάδιο της προεπεξεργασίας των δεδομένων, στη συνέχεια περιγράφεται το στάδιο της αυτόματης εύρεσης του αριθμού των

συνολικών ομάδων και του αλγορίθμου συμπλήρωσης των χαμένων τιμών με χρήση του αλγορίθμου KNN Impute. Στα επόμενα κεφάλαια περιγράφονται οι αλγόριθμοι ομαδοποίησης των γονιδίων σύμφωνα με τα προφίλ έκφρασής τους. Για την ομαδοποίηση χρησιμοποιήθηκαν 3 διαφορετικοί αλγόριθμοι, ο KMeans, ο Fuzzy KMeans, χρήση Gaussian Mixture Models (GMM) και του αλγορίθμου EM. Τέλος περιγράφεται η μέθοδος που χρησιμοποιείται για την ανακατασκευή του γενετικού δικτύου με χρήση γραφικών στοχαστικών μοντέλων . (βλ. Πίνακας 7.1)

Η μεθοδολογία εφαρμόστηκε σε microarray σύνολα δεδομένων [Spellman et al. 1998], [Cho et al. 1998] για τον κύκλο των κυττάρων ζύμης Yeast Saccharomyces cerevisiae. Η περιγραφή των δεδομένων και του τρόπου παραγωγής τους παρουσιάζεται στην επόμενη παράγραφο.



Σχήμα 7-1 Γενική μεθοδολογία

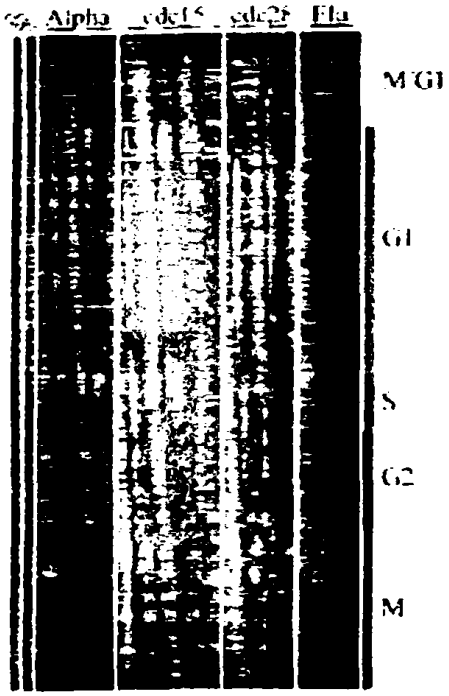
Βήμα	Περιγραφή
1ο	Κανονικοποίηση συνόλου δεδομένων (έχει εφαρμοστεί ήδη στα δεδομένα που χρησιμοποιήσαμε)
2ο	Αφαίρεση γονιδίων και πειραμάτων με μικρό αριθμό μετρήσεων
3ο	Συμπλήρωση χαμένων τιμών

4ο	Απαλοιφή μη σημαντικών γονιδίων
5ο	Εύρεση βέλτιστου αριθμού ομάδων
6ο	Ομαδοποίηση
7ο	Ανακατασκευή γενετικού δικτύου

Πίνακας 7.1 Περιγραφή βημάτων της μεθοδολογίας

7.2 Σύνολο Δεδομένων

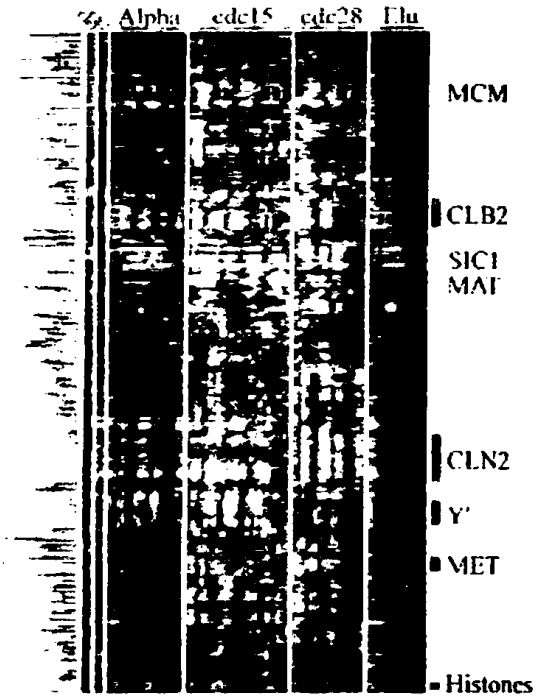
Τα microarray πειράματα παράγουν σύνολα δεδομένων που μπορούν να βοηθήσουν στην ανακατασκευή δικτύων γονιδίων. Τα δεδομένα τα οποία χρησιμοποιήσαμε στην μέθοδό μας έχουν παρθεί από μετρήσεις οι οποίες έγιναν στον οργανισμό Yeast Saccharomyces cerevisiae κατά την διάρκεια του κύκλου του κυττάρου [Spellman 98], [Cho et al. 1998]. Περιέχει τα επίπεδα έκφρασης 6.218 S. cerevisiae υποθετικών αντιγράφων γονιδίων (που προσδιορίζονται ως ORFs Open Reading Frames) τα οποία έχουν μετρηθεί ανά διαστήματα 10 λεπτών άνω των δύο κύκλων κυττάρων (160 λεπτά). Σκοπός της εργασίας [Spellman 98] ήταν να δημιουργηθεί ένας περιεκτικός κατάλογος γονιδίων της ζύμης Yeast Saccharomyces cerevisiae των οποίων τα επίπεδα έκφρασης θα μεταβάλλονται περιοδικά κατά την διάρκεια του κύκλου του κυττάρου.



Εικόνα 7-1: Εκφράσεις γονιδίων κατά την διάρκεια του κύκλου του κυττάρου της ζύμης. Τα γονίδια αντιστοιχούν στις σειρές, και τα χρονικά σημεία κάθε πειράματος είναι οι στήλες. Η αναλογία της επαγωγής/καταστολής παρουσιάζεται για κάθε γονίδιο έτσι ώστε το μέγεθος να υποδεικνύεται από την ένταση των χρωμάτων που φαίνονται. Εάν το χρώμα είναι μαύρο, η αναλογία του ελεγχόμενου (γνωστού) προς το πειραματικό (άγνωστο) cDNA είναι ίση με 1, τα φωτεινότερα χρώματα (κόκκινο και πράσινο) αντιπροσωπεύουν μια αναλογία 2.8:1. Αναλογίες > 2.8 παρουσιάζονται με φωτεινότερο χρώμα. Σε όλες τις περιπτώσεις το κόκκινο δείχνει μια αύξηση στην ποσότητα mRNA, ενώ το πράσινο δείχνει μια μείωση στην ποσότητα. Οι γκρίζες περιοχές (αν είναι ορατές) δείχνουν τα απόντα δεδομένα ή τα δεδομένα χαμηλής ποιότητας. Οι περιοχές χρώματος δεξιά δείχνουν την ομάδα φάσης στην οποία ένα γονίδιο ανήκει (M/G1, yellow; G1, green; S, purple; G2, red; M, orange).

Για τον σκοπό αυτό, χρησιμοποιήθηκαν DNA microarrays σε δείγματα από πληθυσμούς ζύμης στα οποία εφαρμόστηκαν τρεις ανεξάρτητοι παράγοντες συγχρονισμού: α) καθυστέρηση του παράγοντα alpha, β) elutriation, και γ) καθυστέρηση μιας cdc15 μετάλλαξης ευαίσθητης στην θερμοκρασία. Στην πορεία χρησιμοποιώντας απλούς αλγορίθμους περιοδικότητας και συσχετισμού, προσδιορίστηκαν 800 γονίδια που ικανοποιούν ένα αντικειμενικό ελάχιστο κριτήριο που υποδεικνύει την συμμετοχή τους στις ρυθμιστικές διαδικασίες του κύκλου του

κυττάρου. Μια πλήρης περιγραφή του συνόλου δεδομένων είναι διαθέσιμη στο http://cellcycle-www.stanford.edu



Εικόνα 7-2 Τα γονίδια που έχουν παρόμοια πρότυπα έκφρασης ομαδοποιούνται από έναν αλγόριθμο ομαδοποίησης. Η ιεραρχική διάταξη στο αριστερό μέρος παρουσιάζει την ιεραρχική δομή των ομάδων.

Παρακάτω περιγράφονται αναλυτικά οι τρόποι με τους οποίους έγινε ο συγχρονισμός σε δύο από τα πειράματα που εκτελέσθηκαν.

7.2.1 Συγχρονισμός Alpha Factor

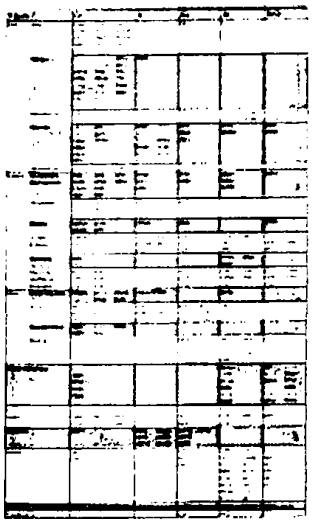
Αρχικά η πίεση της ζύμης DBY8724 αυξήθηκε σε ένα OD600 από 0.2 της YEP γλυκόζης, και λήφθηκε ένα ασύγχρονο δείγμα. Ο παράγοντας α προστέθηκε σε συγκέντρωση 12 ng/ml. Μετά από 120 λεπτά. ο παράγοντας α αφαιρέθηκε με την κοκκοποίηση των κυττάρων για 5 λεπτά σε ένα (Newtown, CT) S34 στροφέα Sorvall σε 3000 στροφές/λεπτό ενώ πραγματοποιήθηκε και μετάγγιση του υπερκείμενου νερού. Τα συλληφθέντα κύτταρα επαναρτήθηκαν σε νέα γλυκόζη YEP σε ένα OD600 0,18. Κάθε 7 λεπτά, για τα επόμενα 140 λεπτά, 25-ml δειγμάτων λήφθηκαν για RNA, και 5-ml FACS ανάλυση. Σε 91 λεπτά μετά την απελευθέρωση το OD600 του πληθυσμού μειώθηκε σε ~0.2 από ~0.4.

Εικόνα 7-3 Κύτταρα της ζύμης (yeast) Saccharomyces cerevisiae

7.2.2 Cdc15 συγχρονισμός

Η πίεση στο cdc15-2 (DBY8728) αυξήθηκε σε 2.5 × 106 cells/ml στην YEP γλυκόζη στους 23°C. Ο πληθυσμός στη συνέχεια μετατοπίστηκε σε έναν επωαστήρα αέρα 37°C και κρατήθηκε σε εκείνη την θερμοκρασία για 3.5 ώρες. Στο μεταξύ, η πυκνότητα των κυττάρων είχε φθάσει στα 6.6×106 cells/ml, και 96% των κυττάρων είχαν τα απαραίτητα χαρακτηριστικά για την cdc15 λήψη. Τα κύτταρα έπειτα απελευθερώθηκαν από τη CDC15 δέσμευση με τη μετατόπιση του πληθυσμού σε ένα λουτρό ύδατος 23°C. Τα δείγματα λήφθηκαν κάθε 10 λεπτά για 300 λεπτά, ξεκινώντας την στιγμή της μετατόπισης στους 23°C. Μέχρι και 300 λεπτά μετά από τη μετατόπιση, η πυκνότητα των κυττάρων είχε φθάσει στα 4 × 107 cells/ml. Μέρος του ίδιου αρχικού πληθυσμού αναπτύχθηκε στους 23°C σε 1 × 107 cells/ml, και τα κύτταρα συγκομίστηκαν για την εξαγωγή του mRNA ελέγχου (mRNA σημείο αναφοράς). Η πρόοδος του κύκλου των κυττάρων ελέγχθηκε από την εμφάνιση νέων γενέσεων.

Επίσης στην ίδια εργασία παράχθηκαν μετρήσεις από τα πειράματα Cln3 και Clb2 ενώ στο σύνολο δεδομένων περιλαμβάνονται και πειραματικές μετρήσεις που έγιναν κάνοντας χρήση του συγχρονισμού CDC28 και αναφέρονται στην εργασία [Cho et al. 1998]. Λεπτομερέστερες περιγραφές των διαδικασιών λήψης των μετρήσεων, του εξοπλισμού και του υλικού των μετρήσεων μπορεί να βρεθεί στις εργασίες [Cho et al. 1998], [Spellman 98] και στο http://cellcycle-www.stanford.edu



Εικόνα 7-4 Γονίδια που ρυθμίζονται κατά την διάρκεια του κύκλου του κυττάρου με χαρακτηρισμένες λειτουργίες. Διακόσια ενενήντα επτά από τα ρυθμισμένα γονίδια του κύκλου του κυττάρου ομαδοποιούνται και από τη λειτουργικότητά τους και από τις μετρήσεις έκφρασης. Με κόκκινο παρουσιάζονται τα γονίδια που είναι γνωστό ότι ρυθμίζονται κατά την διάρκεια του κύκλου του κυττάρου (ο πίνακας παρουσιάζεται σε κανονικό μέγεθος στο τέλος του κεφαλαίου)

Αποτέλεσμα της εργασίας [Spellman 98] ήταν η αναφορά 800 γονιδίων ζύμης (Εικόνα 7 4)των οποίων τα αντίγραφα ταλαντεύονται περιοδικά ανά τον κύκλο του κυττάρου. Τα 800 γονίδια καθορίστηκαν με τη χρήση ενός αντικειμενικού, εμπειρικού μοντέλου, στο οποίο κάποια κατώφλια είναι αυθαίρετα. Είναι γεγονός ότι κάτω από αυτό το κατώτατο όριο μπορούν να υπάρξουν γονίδια των οποίων η έκφραση να είναι περιοδική και των οποίων η περιοδικότητα δεν έχει βιολογική σημασία. Δυστυχώς δεν μπορούμε να ανιχνεύσουμε τέτοια γονίδια, αλλά είναι πιθανό ότι τα γονίδια αυτά είναι σχετικά λίγα σε αριθμό διότι εκφράζονται περιοδικά κατά τη διάρκεια του κύκλου του κυττάρου βρίσκονται κάτω από το εμπειρικό κατώφλι.

7.3 Κανονικοποίηση και μετασχηματισμός των δεδομένων μικροσυστοιχιών

Ο στόχος των περισσότερων πειραμάτων microarray είναι να ερευνηθούν τα πρότυπα της έκφρασης γονιδίων με το να αναλυθούν τα επίπεδα έκφρασης χιλιάδων γονιδίων σε μια ενιαία ανάλυση. Τυπικά, το RNA είναι το πρώτο που απομονώνεται από διάφορους ιστούς, αναπτυξιακά στάδια, καταστάσεις ασθενειών, κτλ. Το RNA

αφού απομονωθεί χαρακτηρίζεται και υβριδοποιείται στους πίνακες χρησιμοποιώντας μια μέθοδο που επιτρέπει στην κάθε έκφραση να αναλυθεί και να συγκριθεί με κατάλληλα ζεύγη δειγμάτων. Οι μέθοδοι που χρησιμοποιούνται για την παραγωγή των επιπέδων έκφρασης έχουν περιγραφεί στο κεφάλαιο 3.

7.3.1 Λόγος έκφρασης

Στα περισσότερα πειράματα microarray ερευνώνται οι σχέσεις μεταξύ σχετικών βιολογικών δειγμάτων βασισμένων σε πρότυπα έκφρασης. Έστω ότι έχουμε έναν πίνακα που έχει N_{array} στοιχεία, και έχει παραχθεί συγκρίνοντας ένα άγνωστο γονίδιο και ένα γονίδιο δείγμα αναφοράς, τα οποία για ευκολία θα χαρακτηρίζουμε R και G, αντίστοιχα (εξαιτίας του κόκκινου και πράσινου χρώματος που χρησιμοποιείται συνήθως για να αντιπροσωπεύσουν τα στοιχεία του πίνακα). Η αναλογία (T) για το i_{th} γονίδιο (όπου το i είναι ένας δείκτης που τρέχει σε όλα τα παραταγμένα γονίδια από 1 έως N_{array}) μπορεί να γραφτεί ως:

$$T_i = \frac{R_i}{G_i} \tag{7.1}$$

Ο ορισμός 7.1 δεν μας περιορίζει στην χρήση κάποιας συγκεκριμένης τεχνολογίας (για την παραγωγή των R_i, G_i): τα μέτρα R_i, G_i μπορούν να δημιουργούνται είτε σε έναν απλό πίνακα είτε σε δύο ρεπλίκες πινάκων. Επιπλέον, όλοι οι μετασχηματισμοί που περιγράφονται παρακάτω μπορούν να εφαρμοστούν σε στοιχεία παραγόμενα από οποιαδήποτε πλατφόρμα microarray.

Αν και ο λόγος παρέχει ένα διαισθητικό μέτρο των μεταβολών της έκφρασης, έχουν το μειονέκτημα ότι μεταχειρίζονται τα υπό- και υπέρ-εκφραζόμενα γονίδια διαφορετικά. Τα γονίδια που υπέρ-εκφράζονται κατά έναν παράγοντα 2 έχουν αναλογία έκφρασης 2, ενώ εκείνα τα οποία υπό-εκφράζονται κατά τον ίδιο παράγοντα έχουν αναλογία έκφρασης (-0,5). Ο πιο κοινά χρησιμοποιημένος εναλλακτικός μετασχηματισμός για τους λόγους είναι ο λογάριθμος με βάση 2, ο οποίος έχει το πλεονέκτημα να παράγει ένα συνεχές φάσμα τιμών για τα υπέρ- και υπό-εκφρασμένα γονίδια. Σημειώνουμε ότι οι λογάριθμοι μεταχειρίζονται τους αριθμούς και τα παράγωγά τους συμμετρικά: $\log_2(1) = 0$, $\log_2(2) = 1$, $\log_2(½) = -1$, $\log_2(4) = 2$, $\log_2(1/4) = -2$, κτλ. Οι λογάριθμοι των λόγων έκφρασης αντιμετωπίζονται επίσης συμμετρικά, έτσι ώστε ένα γονίδιο που υπέρ-εκφράζεται

κατά έναν παράγοντα 2 έχει $\log_2$(αναλογία)=1, ένα γονίδιο το οποίο υπό-εκφράζεται κατά έναν παράγοντα 2 έχει log2(αναλογία) = -1, και ένα γονίδιο που εκφράζεται κατά μια σταθερά (με ένα λόγο 1) έχει $\log_2$(αναλογία) ίσος με μηδέν. Κατά την προεπεξεργασία θα χρησιμοποιηθεί η $\log_2$(αναλογία) τιμή για να αντιπροσωπεύσει τα επίπεδα έκφρασης.

7.3.2 Κανονικοποίηση των δεδομένων

Τυπικά, ο πρώτος μετασχηματισμός που εφαρμόζεται στα δεδομένα έκφρασης, είναι η κανονικοποίηση, σύμφωνα με την οποία ρυθμίζονται ανεξάρτητες εντάσεις υβριδοποίησης ώστε να ισορροπηθούν κατάλληλα και να μπορούν να γίνουν βιολογικές συγκρίσεις. Υπάρχουν διάφοροι λόγοι για τους οποίους τα δεδομένα πρέπει να κανονικοποιηθούν, μερικοί από αυτούς αναφέρονται στην διαφορετικότητα των αρχικών ποσοτήτων του RNA, στις διαφορές των χαρακτηρισμών ή στην ανίχνευση του φθορισμού των χρησιμοποιούμενων χρωστικών ουσιών. Εννοιολογικά, η κανονικοποίηση είναι παρόμοια με τη ρύθμιση των επιπέδων έκφρασης που μετριούνται από το ποσοτικό αντίστροφο της μεταγραφής PCR (RT-PCR) σχετικά με την έκφραση ενός ή περισσότερων γονιδίων αναφοράς των οποίων τα επίπεδα υποτίθεται ότι ήταν σταθερά κατά την διάρκεια της δειγματοληψίας. Υπάρχουν πολλές προσεγγίσεις στην κανονικοποίηση των επιπέδων έκφρασης. Μερικές μέθοδοι όπως η συνολική κανονικοποίηση της έντασης, είναι βασισμένες σε απλές υποθέσεις, μια από αυτές είναι ότι αρχικά οι ποσότητες RNA για τα δύο δείγματα που πρόκειται να συγκρίνουμε είναι ίσες. Δεδομένου ότι υπάρχουν εκατομμύρια ανεξάρτητα μόρια RNA σε κάθε δείγμα, θα υποθέσουμε ότι η μέση μάζα κάθε μορίου είναι περίπου η ίδια, και ότι, κατά συνέπεια, ο αριθμός μορίων σε κάθε δείγμα είναι επίσης ο ίδιος. Επίσης, υποθέτουμε ότι τα παραταγμένα στοιχεία (στον πίνακα) αντιπροσωπεύουν μια τυχαία δειγματοληψία των γονιδίων στον οργανισμό. Εάν τα παραταγμένα γονίδια επιλέγονται για να αντιπροσωπεύσουν μόνο εκείνα τα οποία γνωρίζουμε ότι μεταβάλλονται, τότε πιθανότατα να υπό-δειγματολειπτήσουμε ή να υπέρ-δειγματολειπτήσουμε τα γονίδια κατά την σύγκριση των βιολογικών δειγμάτων. Εάν ο πίνακας περιέχει μια αρκετά μεγάλη διάταξη τυχαίων γονιδίων, δεν αναμένουμε να δούμε κάποια τέτοια πόλωση. Αυτό συμβαίνει επειδή για ένα πεπερασμένο δείγμα RNA, όταν η αναπαράσταση ενός μειωθεί η αναπαράσταση των άλλων πρέπει να αυξηθεί. Συνεπώς, πρέπει να υβριδοποιείται περίπου ο ίδιος αριθμός χαρακτηριζόμενων

μορίων από κάθε δείγμα στους πίνακες και, επομένως, οι συνολικές εντάσεις υβριδοποίησης που αθροίζονται σε όλα τα στοιχεία στις σειρές πρέπει να είναι οι ίδιες για κάθε δείγμα. Χρησιμοποιώντας αυτήν την προσέγγιση, ένας παράγοντας κανονικοποίησης υπολογίζεται με το άθροισμα των μετρημένων εντάσεων και στα δύο κανάλια

$$N_{total} = \frac{\sum_{i=1}^{N_{array}} R_i}{\sum_{i=1}^{N_{array}} G_i} \tag{7.2}$$

όπου οι G_i και R_i είναι οι μετρημένες εντάσεις για το i_{th} στοιχείο του πίνακα (οι πράσινες και κόκκινες εντάσεις σε ένα microarray πείραμα δύο-χρωμάτων) και N_{array} είναι ο συνολικός αριθμός στοιχείων που αντιπροσωπεύονται στον microarray. Η μια ή και οι δύο εντάσεις διαβαθμίζονται κατάλληλα, π.χ.,

$$G_k' = N_{total} G_k \text{ and } R_k' = R_k \tag{7.3}$$

άρα ο κανονικοποιημένος λόγος έκφρασης για κάθε στοιχείο είναι:

$$T_i = \frac{R_i}{G_i} = \frac{1}{N_{total}} \frac{R_i}{G_i} \tag{7.4}$$

ο λόγος αυτός είναι γνωστός ως log-ratio και ρυθμίζει κάθε λόγο ώστε ο μέσος λόγος να είναι ίσος με 1. Αυτή η διαδικασία είναι ισοδύναμη με την αφαίρεση μιας σταθεράς από το λογάριθμο του λόγου έκφρασης,

$$\log_2(T_i') = \log_2(T_i) - \log_2(N_{total}) \tag{7.5}$$

Στα δεδομένα τα οποία χρησιμοποιήσαμε οι τιμές εκφράζουν το log(ratio) με βάση τον ορισμό που δόθηκε παραπάνω.

7.3.3 Ολική και τοπική κανονικοποίηση

Πολλές φορές οι αλγόριθμοι κανονικοποίησης, εφαρμόζονται είτε συνολικά (σε ολόκληρο το σύνολο δεδομένων) είτε τοπικά (σε κάποιο φυσικό υποσύνολο των

δεδομένων). Για τους spotted πίνακες (αρχικές εικόνες των microarrays), η τοπική κανονικοποίηση εφαρμόζεται συχνά σε κάθε ομάδα στοιχείων του πίνακα που παράγεται από μια κοινή γραφίδα (μερικές φορές τα σύνολα καλούνται "pen group" ή "subgrid"). Η τοπική κανονικοποίηση έχει το πλεονέκτημα το ότι μπορεί να βοηθήσει την διόρθωση της χωρικής μεταβλητότητας που μπορεί να εμφανιστεί στον πίνακα, συμπεριλαμβανομένων των ασυνεπειών μεταξύ γραφίδων που χρησιμοποιούνται για να δημιουργήσουν τον πίνακα, τη μεταβλητότητα στην επιφάνεια των διαφανειών (slides), και τις μικρές τοπικές διαφορές στους όρους υβριδοποίησης που εφαρμόζονται πάνω στον πίνακα.

Εκτιμώντας ότι η κανονικοποίηση ρυθμίζει το μέσο όρο των λογαρίθμων των μετρήσεων, οι στοχαστικές διαδικασίες μπορούν να αναγκάσουν τη μεταβλητότητα των μετρημένων $\log_2$ τιμών (λόγου) να διαφέρουν από μια περιοχή του πίνακα σε μια άλλη. Μια προσέγγιση στην εξέταση αυτού του προβλήματος είναι να ρυθμιστούν τα $\log_2$ μέτρα (λόγου) έτσι ώστε η διαφορά να είναι η ίδια. Εάν εξετάζουμε ένα πίνακα με διακριτά subgrids για τα οποία έχουμε κάνει τοπική κανονικοποίηση, τότε αυτό που επιδιώκουμε είναι να βρούμε ένα παράγοντα για κάθε subgrid που μπορούμε να χρησιμοποιήσουμε για να κλιμακώσουμε (scaling) όλες τις μετρήσεις μέσα σε αυτό το subgrid.

Ένας κατάλληλος παράγοντας κλιμάκωσης είναι η μεταβλητότητα ενός subgrid διαιρεμένο με το γεωμετρικό μέσο όρο των διαφορών για όλα τα subgrids. Εάν υποθέτουμε ότι κάθε subgrid έχει M στοιχεία, επειδή έχουμε ρυθμίσει ήδη το μέσο όρο των $\log_2$ τιμών (λόγου) σε κάθε subgrid να είναι μηδέν, η διαφορά τους στον n_{th} subgrid είναι

$$\sigma_n^2 = \sum_{j=1}^{M} \left[\log_2(T_j)\right]^2 \tag{7.6}$$

όπου το άθροισμα διατρέχει όλα τα στοιχεία σε αυτό το subgrid. Εάν ο αριθμός των subgrids στον πίνακα είναι N_{grids}, τότε ο κατάλληλος παράγοντας scaling για τα στοιχεία του k_{th} subgrid στον πίνακα είναι

$$a_k = \frac{\sigma_k^2}{\left[\prod_{n=1}^{N_{grids}} \sigma_n^2\right]^{1/N_{grids}}} \tag{7.7}$$

Στη συνέχεια εφαρμόζουμε το scaling σε όλα τα στοιχεία στο k_{th} subgrid διαιρώντας με την ίδια αξία a_k που υπολογίζεται για αυτό το subgrid,

$$\log_2(T_i) = \frac{\log_2(T_i)}{a_k} \tag{7.8}$$

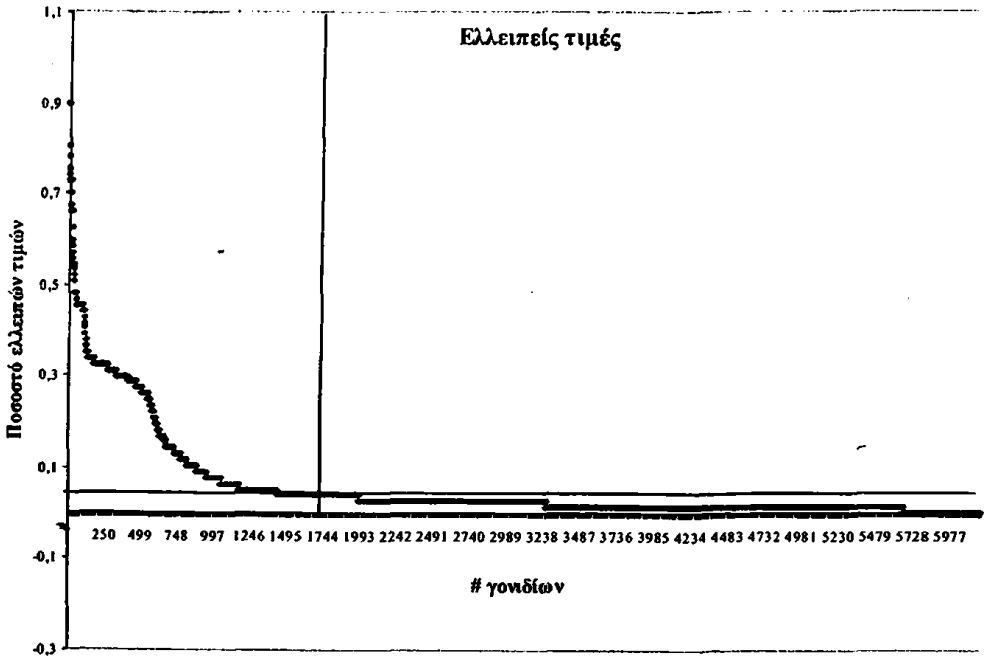
Αυτό είναι ισοδύναμο με το παίρνουμε την a_k ρίζα των μεμονωμένων εντάσεων στο k subgrid,

$$G_i^{'} = [G_i]^{1/a_k} \text{ and } R_i^{'} = [R_i]^{1/a_k} \tag{7.9}$$

Πρέπει να σημειωθεί ότι έχουν προταθεί και άλλοι παράγοντες κανονικοποίησης σε διάφορες ερευνητικές εργασίες.

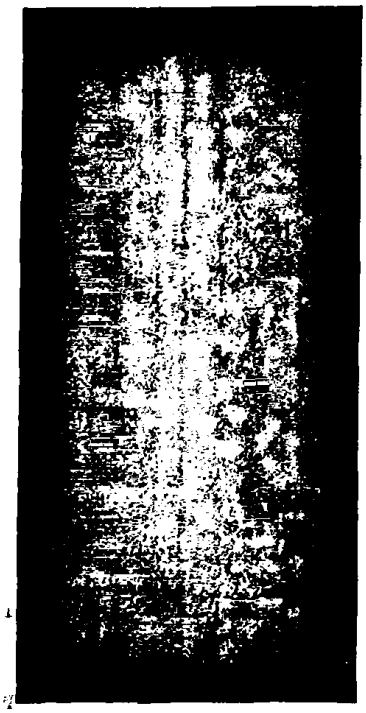
7.4 Αφαίρεση μη σημαντικών γονιδίων - "Significance Cuts I"

Τα δεδομένα έκφρασης γονιδίων που χρησιμοποιούνται στην εργασία όπως έχουμε αναφέρει αναπαρίστανται με την μορφή ενός πίνακα διαστάσεων $N \times M$, ο οποίος αποτελείται από σύνολο N γονιδίων των οποίων οι λόγοι έκφρασης μετριούνται σε M διαφορετικά πειράματα (καταστάσεις). Ο λογαριθμικός λόγος έκφρασης για το γονίδιο που μετρήθηκε στο πείραμα j είναι το x_{ij}. Στις περισσότερες περιπτώσεις χρησιμοποιείται, ένας πίνακας στοιχείων με λιγότερα δεδομένα και από τις δύο διαστάσεις n και m, ο λόγος είναι ότι σε πολλά από τα γονίδια λείπουν τιμές κάτι το οποίο οφείλεται είτε σε σφάλματα των μετρήσεων, είτε σε σφάλματα στην σάρωσης του slide από το ειδικό μηχάνημα. Για το λόγο αυτό είναι απαραίτητο κάποιου είδους φιλτράρισμα στο αρχικό σύνολο δεδομένων. Αυτό το φιλτράρισμα είναι γνωστό και με το όνομα "significance cut".



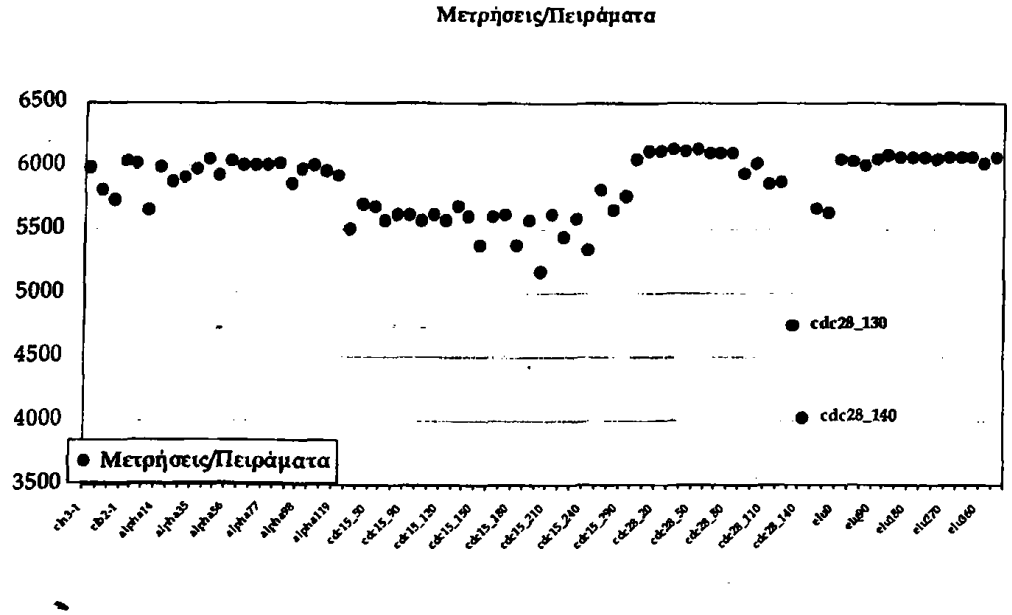
Εικόνα 7-5 Ποσοστά ελλιπών τιμών στις 77 μετρήσεις πειραμάτων. Ο άξονας x είναι τα γονίδια, ενώ ο άξονας y το ποσοστό των τιμών που λείπουν από κάθε γονίδιο.

Σε πρώτη φάση γίνεται αφαίρεση των γονιδίων στα οποία λείπουν περισσότερες από το 7% των μετρήσεων. Στην Εικόνα 7-5 φαίνεται το ποσοστό των τιμών που λείπουν από τα γονίδια για τις 77 μετρήσεις. Σκοπός αυτής της απαλοιφής είναι να παραμείνουν τα γονίδια για τα οποία υπάρχουν αρκετές πειραματικές μετρήσεις. Όπως είναι γνωστό πολλοί αλγόριθμοι ομαδοποίησης (όπως και αυτοί που έχουν εφαρμοστεί παρακάτω είναι ευαίσθητοι όταν δεν υπάρχουν τιμές σε μετρήσεις.



Εικόνα 7-6 Στον χάρτη παρουσιάζονται με έντονο μοβ οι τιμές των γονιδίων που λείπουν (έχουν επιλεχθεί 160 γονίδια)

Κάνοντας τον ίδιο έλεγχο παίρνοντας ως μεταβλητές τα πειράματα θα παρατηρούσαμε ότι υπάρχουν 2 πειράματα (cdc28_130 , cdc28_140) τα οποία δεν έχουν δώσει μετρήσεις στο ~40% των γονιδίων (Εικόνα 7.6 και 7.7). Αφαιρώντας εκείνα τα πειράματα καταφέρνουμε να ελαττώσουμε τον αριθμό των γονιδίων που θα αφαιρέσουμε οριστικά από το σύνολο δεδομένων.



Εικόνα 7-7 Ο αριθμός των γονιδίων που έχουν μετρήσεις για κάθε πείραμα

Συνολικά σε αυτό το βήμα αφαιρέθηκαν 1655 γονίδια.

7.5 Αλγόριθμος για συμπλήρωση χαμένων τιμών

Τα δεδομένα των πειραμάτων microarray που χρησιμοποιήσαμε έχουν μεγάλες ελλείψεις στις τιμές των πειραμάτων για κάθε γονίδιο, υπολογίστηκε ότι μόνο ένα 8.125% των γονιδίων (μεταβλητών) έχουν τιμές σε όλες τις μετρήσεις, είναι όμως αδύνατο να εφαρμόσουμε τις υπόλοιπες μεθόδους σε αυτό το ποσοστό των γονιδίων. Απ' την άλλη ένα 82% των γονιδίων έχει μικρές ελλείψεις (λείπουν 1 ή 2 τιμές) στις πειραματικές μετρήσεις, για τον λόγο αυτό η λύση που προτάθηκε είναι η συμπλήρωση των τιμών για εκείνα τα γονίδια.

Τα τυπικά microarray chips αποτελούνται από χιλιάδες σημεία και πολλά από αυτά παράγουν μη χρήσιμα δεδομένα. Οι περισσότερες από τις λάθος εκτιμήσεις προέρχονται από

1. ελαττωματικά σημεία στα τσιπ,
2. ελαττώματα στο απεικονιστικό μηχάνημα την ώρα που λαμβάνει την εικόνα, κ.τ.λ.

Το βέβαιο είναι πως η έλλειψη δεδομένων μπορεί να επηρεάσει την ανάλυση των γονιδιακών εκφράσεων και κατά συνέπεια να αποδώσει μια μεγάλη δόση θορύβου στα αποτελέσματα.

Ο αλγόριθμος που χρησιμοποιήθηκε σε αυτή την εργασία με σκοπό την συμπλήρωση των χαμένων δεδομένων ονομάζεται KNN impute. Σύμφωνα με την προτεινόμενη μεθοδολογία ο αλγόριθμος εφαρμόζεται σε ένα σύνολο δεδομένων στο οποίο οι χαμένες τιμές κατά την διάρκεια όλων των πειραμάτων δεν ξεπερνούν το 7% αυτών. Γενικά ο αλγόριθμος KNN impute χρησιμοποιείται για την προσθήκη χαμένων δεδομένων. Αυτό το πετυχαίνει ψάχνοντας ολόκληρο το σύνολο δεδομένων για παρόμοια πρότυπα γονιδιακής έκφρασης και έπειτα υπολογίζει τις χαμένες τιμές κάνοντας χρήση ενός κατάλληλου ποσοστού. Τα αποτελέσματα δείχνουν ότι ο αλγόριθμος KNN impute δουλεύει καλά όταν ο αριθμός των δεδομένων που λείπουν δεν είναι μεγάλος.

Πιο συγκεκριμένα ο KNN impute βρίσκει τους K πιο κοντινούς γείτονες του διανύσματος r από όλο τον πίνακα, το μέτρο ομοιότητας είναι η Euclidean απόσταση. Οποιοδήποτε διάνυσμα με ελλιπείς τιμές δεν χρησιμοποιείται στους υπολογισμούς [Troyanskaya et al. 2001].

Περιγραφή αλγορίθμου:

Έστω ότι εξετάζουμε το γονίδιο x^* που έχει μια χαμένη τιμή στο πείραμα 10, σκοπός μας είναι να βρούμε K άλλα γονίδια που έχουν παρόμοια έκφραση με το γονίδιο x^*.

Υπολογίζουμε την Ευκλείδεια απόσταση μεταξύ του x^* και των άλλων γονιδίων Xc, χρησιμοποιώντας μόνο εκείνες τις συντεταγμένες που δεν λείπει τιμή από το x*. Ορίζοδμε τα K κοντινότερα και αποδίδουμε ένα βάρος $w(i)$ σε καθένα από αυτά το οποίο ορίζεται από την σχέση:

$$dSum = \sum_{i=1}^{K} \frac{1}{dis(i)} \quad , w(i) = \frac{1}{dis(i)} dSum \tag{7.10}$$

Ένας μέσος όρος των τιμών με χρήση βαρών από τα K κοντινότερα γονίδια χρησιμοποιείται για τον υπολογισμό της χαμένης τιμής του γονιδίου A η οποία δίνεται από τον τύπο

$$value = \sum_{i=1}^{K} D(A,i) * w(i) \tag{7.11}$$

Αποτελέσματα:

Σύμφωνα με το προηγούμενο βήμα τα γονίδια που αποτελούν είσοδο στο τρέχον βήμα είναι εκείνα των οποίων οι χαμένες τιμές δεν ξεπερνούν το 7% των συνολικών τιμών για κάθε ολοκληρωμένη εκτέλεση του πειράματος.

-0.15	-0.15	-0.21	0.17	-0.42	-0.44	-0.15	0.24	-0.1	[redacted]	0.18	0.42	-0.25	-0.01	-0.13	0.77	-0.21	0.43
-0.14	-0.71	0.1	-0.32	-0.4	-0.58	0.11	0.21	0.09	0.57	-0.14	0.29	0.01	0.04	0.05	0.55	-0.08	0.33
-0.05	-0.53	-0.47	-0.06	0.11	-0.07	0.25	0.46	0.12	0.49	-0.42	0.28	-0.3	-0.18	-0.3	0.68	-0.24	0.22
-0.03	-0.27	0.17	-0.12	-0.27	0.06	0.23	0.11	0.03	-0.27	-0.24	0.36	-0.11	0	-0.27	0.46	0.23	-0.06

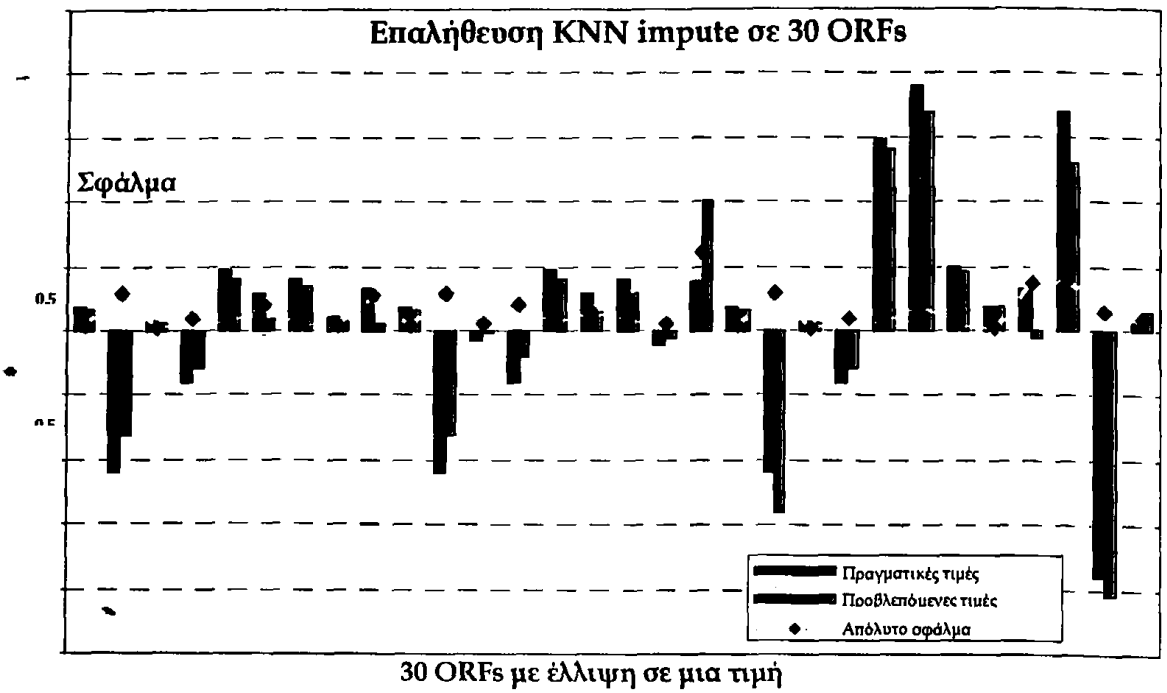
-0.05	0.13	0.13	-0.21	-0.45	-0.21	0.06	0.32	0	0.26	-0.5	0.31	0.1	-0.14	-0.71	0.51	0.31	0.13
0.02	-0.33	-0.49	-0.3	-0.15	-0.24	0.4	0.53	0.25	0.49	-0.3	0.13	-0.12	-0.35	-0.19	0.47	0.06	0.13
-0.1	-0.15	-0.01	-0.25	-0.16	-0.13	0.06	0.19	-0.06	0.05	-0.25	0.23	-0.05	-0.13	0.09	0.22	0.35	0.09
0	-0.01	0.12	-0.23	-0.13	0.25	0.3	-0.27	0.38	-0.06	-0.18	-0.16	-0.27	-0.16	0.09	0.09	0.22	0
0.06	0.01	0.17	-0.14	0.01	-0.24	0.15	-1.34	0.15	0.06	0.01	0.18	0.09	0.06	0.18	0.14	0.3	0.17
-0.24	-0.95	-0.23	0.12	-0.02	0.23	-0.11	0.11	0	0.16	0.19	0.35	0.02	0.14	-0.14	0.07	0.23	0.09
-0.02	-0.29	-0.07	-0.22	-0.06	-0.07	0.2	0.2	0.11	0.03	-0.33	0.27	-0.15	0.03	-0.83	1.09	0.11	0.01
-0.11	-0.17	-0.16	0.04	0.1	-0.02	0.08	0.13	-0.17	-0.12	-0.06	0.17	0.08	0.02	0.29	0.2	-0.21	-0.07
-0.36	-0.42	0.29	-0.14	-0.19	-0.52	0.04	0.04	0.37	0.24	0.13	0.22	0.04	-0.24	-0.22	-0.1	0.22	0.61
-0.57	-1.12	-0.21	0.1	0.63	0.41	0.35	0.02	-0.21	-0.3	-0.1	0.14	0.11	0.26	0.4	0.18	0.14	-0.26
-0.19	0.04	-0.03	0.56	0.19	-0.22	0.29	-0.08	0.09	0.03	-0.17	-0.19	-0.09	-0.19	0.32	0.01	-0.17	-0.21
0.17	-1.51	-0.6	-0.42	0.21	0.15	0.2	0.69	0.32	0.05	0.15	-0.12	0.09	0.03	-0.23	0.13	0.45	0.23
0.04	-0.24	-0.21	-0.23	-0.12	-0.31	-0.06	0.07	0.11	0.46	-0.62	0.29	-0.09	0	-0.28	0.63	0.19	0.38
-0.12	0.7	-0.5	-0.39	-0.15	-0.41	0.03	0.13	-0.08	0.18	-0.15	0.14	-0.1	0.1	-0.52	0.65	0.16	0.34
-0.54	1.04	0.05	-0.31	-0.27	-0.27	0.19	-0.38	0	0.3	-0.09	-0.02	0.58	-0.51	0.03	0.14	-0.27	0.33

Ο αλγόριθμος KNN impute ψάχνει τους K=16 πιο κοντινούς γείτονες του γονιδίου 1, υπολογίζοντας την Ευκλείδεια απόσταση χωρίς να παίρνει μέρος η συντεταγμένη που λείπει από το γονίδιο που εξετάζεται (στο παράδειγμα είναι η 10η τιμή)

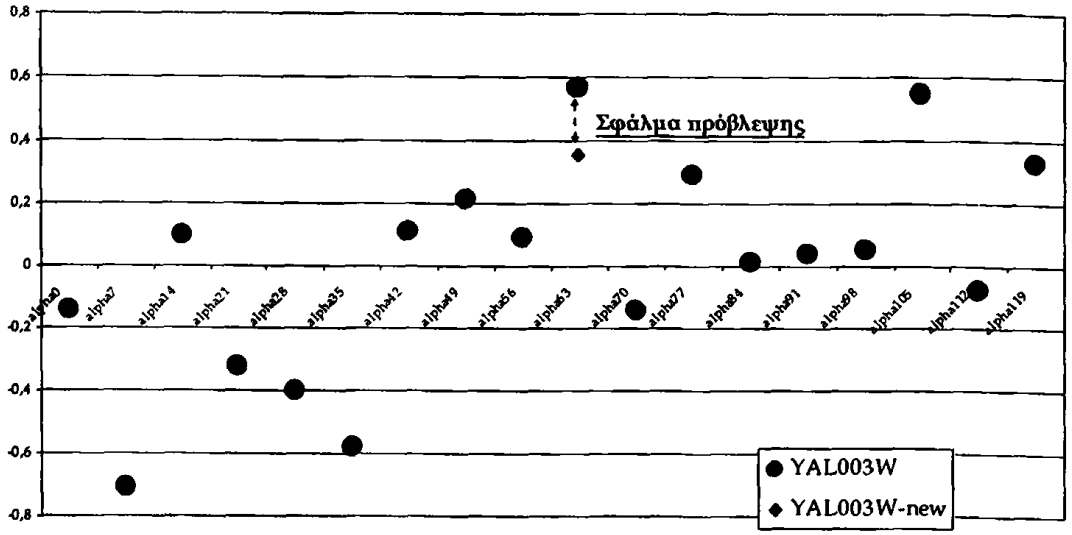
-0.150	-0.150	-0.210	0.170	-0.420	-0.440	-0.150	0.240	-0.100	0.341	0.180	0.420	-0.250	-0.010	-0.130	0.770	-0.210	0.430
-0.140	-0.710	0.100	-0.320	-0.400	-0.580	0.110	0.210	0.090	0.570	-0.140	0.290	0.010	0.040	0.050	0.550	-0.080	0.330
-0.050	-0.530	-0.470	-0.060	0.110	-0.070	0.250	0.460	0.120	0.490	-0.420	0.280	-0.300	-0.180	-0.300	0.680	-0.240	0.220
-0.030	-0.270	0.170	-0.120	-0.270	0.060	0.230	0.110	0.030	-0.270	-0.240	0.360	-0.110	0.000	-0.270	0.460	0.230	-0.060

-0.050	0.130	0.130	-0.210	-0.450	-0.210	0.060	0.320	0.000	0.260	-0.500	0.310	0.100	-0.140	-0.710	0.510	0.310	0.130
0.020	-0.330	-0.490	-0.300	-0.150	-0.240	0.400	0.530	0.250	0.490	-0.300	0.130	-0.120	-0.350	-0.190	0.470	0.060	0.130
-0.100	-0.150	-0.010	-0.250	-0.160	-0.130	0.060	0.190	-0.060	0.050	-0.250	0.230	-0.050	-0.130	0.090	0.220	0.350	0.090
0.000	-0.010	0.120	-0.230	-0.130	0.250	0.300	-0.270	0.380	-0.060	-0.180	-0.160	-0.270	-0.160	0.090	0.090	0.220	0.000
0.060	0.010	0.170	-0.140	0.010	-0.240	0.150	-1.340	0.150	0.060	0.010	0.180	0.090	0.060	0.180	0.140	0.300	0.170
-0.240	-0.950	-0.230	0.120	-0.020	0.230	-0.110	0.110	0.000	0.160	0.190	0.350	0.020	0.140	-0.140	0.070	0.230	0.090
-0.020	-0.290	-0.070	-0.220	-0.060	-0.070	0.200	0.200	0.110	0.030	-0.330	0.270	-0.150	0.030	-0.830	1.090	0.110	0.010
-0.110	-0.170	-0.160	0.040	0.100	-0.020	0.080	0.130	-0.170	-0.120	-0.060	0.170	0.080	0.020	0.290	0.200	-0.210	-0.070
-0.360	-0.420	0.290	-0.140	-0.190	-0.520	0.040	0.040	0.370	0.240	0.130	0.220	0.040	-0.240	-0.220	-0.100	0.220	0.610
-0.570	-1.120	-0.210	0.100	0.630	0.410	0.350	0.020	-0.210	-0.300	-0.100	0.140	0.110	0.260	0.400	0.180	0.140	-0.260
-0.190	0.040	-0.030	0.560	0.190	-0.220	0.290	-0.080	0.090	0.030	-0.170	-0.190	-0.090	-0.190	0.320	0.010	-0.170	-0.210
0.170	-1.510	-0.600	-0.420	0.210	0.150	0.200	0.690	0.320	0.050	0.150	-0.120	0.090	0.030	-0.230	0.130	0.450	0.230
0.040	-0.240	-0.210	-0.230	-0.120	-0.310	-0.060	0.070	0.110	0.460	-0.620	0.290	-0.090	0.000	-0.280	0.630	0.190	0.380
-0.120	0.700	-0.500	-0.390	-0.150	-0.410	0.030	0.130	-0.080	0.180	-0.150	0.140	-0.100	0.100	-0.520	0.650	0.160	0.340
-0.540	1.040	0.050	-0.310	-0.270	-0.270	0.190	-0.380	0.000	0.300	-0.090	-0.020	0.580	-0.510	0.030	0.140	-0.270	0.330

Η επαλήθευση του αλγορίθμου έγινε αφαιρώντας μια τιμή από 30 διανύσματα (γονίδια) και εύρεση με χρήση του αλγορίθμου. Στην Εικόνα 7-8 φαίνεται γραφικά ο έλεγχος ορθότητας του αλγορίθμου εύρεσης χαμένων τιμών.



Εικόνα 7-8 Αποτελέσματα επαλήθευσης του αλγορίθμου KNN-impute

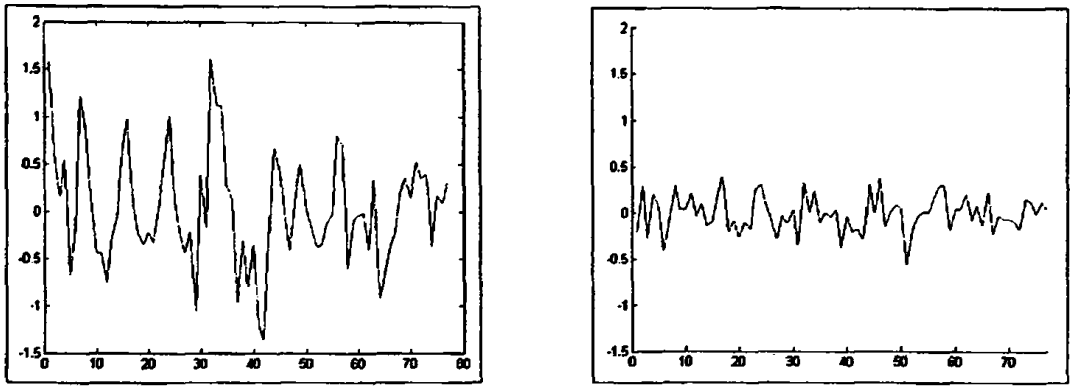


Εικόνα 7-9 πραγματική τιμή, προβλεπόμενη τιμή και σφάλμα για το γονίδιο YAL003W (κατά την διάρκεια επαλήθευσης του αλγορίθμου χρησιμοποιήθηκαν οι τιμές από το alpha πείραμα)

7.6 Αφαίρεση μη σημαντικών γονιδίων

Σκοπός μας σε αυτή τη φάση της μεθοδολογίας είναι η αφαίρεση των γονιδίων των οποίων οι μετρήσεις δεν παράγουν χρήσιμα δεδομένα. Οι μετρήσεις αυτές μπορούν να θεωρηθούν θόρυβος ο οποίος παράχθηκε κατά την διάρκεια της σάρωσης του microarray chip, είτε παράχθηκε κατά την διάρκεια των πειραμάτων. Απ' την άλλη μπορούμε απλά να θεωρήσουμε ότι στο σύστημα που εξετάζουμε υπάρχουν γονίδια τα οποία δεν επηρεάζονται και δεν επηρεάζουν το σύστημα. Τα γονίδια αυτά μπορούν να θεωρηθούν θόρυβος για τις επόμενες μεθόδους ομαδοποίησης και εκπαίδευσης του Μπεϋζιανού δικτύου και για το σκοπό αυτό επιθυμούμε να τα αφαιρέσουμε.

Τα γονίδια τα οποία δεν επηρεάζονται από το σύστημα και δεν επηρεάζουν αυτό έχουν γονιδιακές εκφράσεις οι οποίες δεν μεταβάλλονται σημαντικά κατά την διάρκεια των πειραμάτων (βλ. Εικόνα 7-10). Το γεγονός αυτό δείχνει ότι τα συγκεκριμένα γονίδια είτε δεν συμμετέχουν σε αλληλεπιδράσεις με άλλα γονίδια είτε ότι οι μετρήσεις που έχουμε για αυτά περιέχουν μεγάλες ποσότητες θορύβου.



Εικόνα 7-10 Εκφράσεις γονιδίων με τυπική απόκλιση 0.6 και 0,2 αντίστοιχα (άξονας y: επίπεδα έκφρασης γονιδίων, άξονας x: χρόνος)

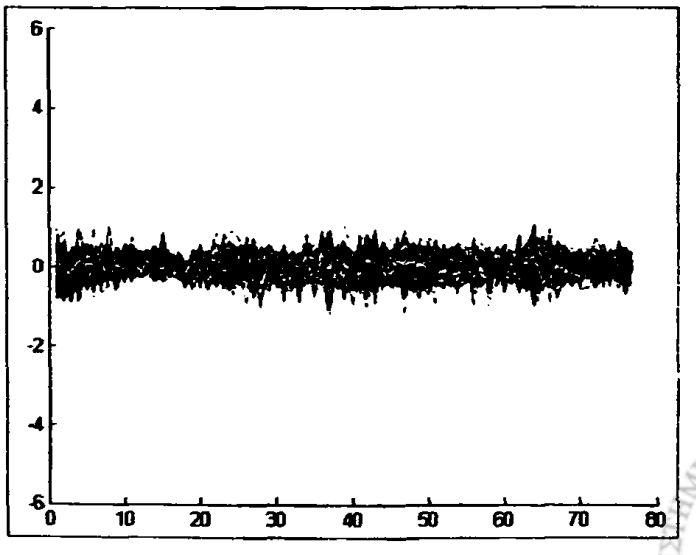
Για το σκοπό αυτό έχει χρησιμοποιηθεί η τυπική απόκλιση. Εφ' όσον τα δεδομένα έχουν προέλθει από *n* σειρές μετρήσεων, τότε το σφάλμα της μέσης τιμής, γνωστό ως τυπική απόκλιση (standard deviation) για το κάθε γονίδιο, προκύπτει από την παρακάτω σχέση:

$$s_{j,j=1,\ldots,m} = \left(1/(n-1) \sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2 \right)^{1/2} \tag{7.12}$$

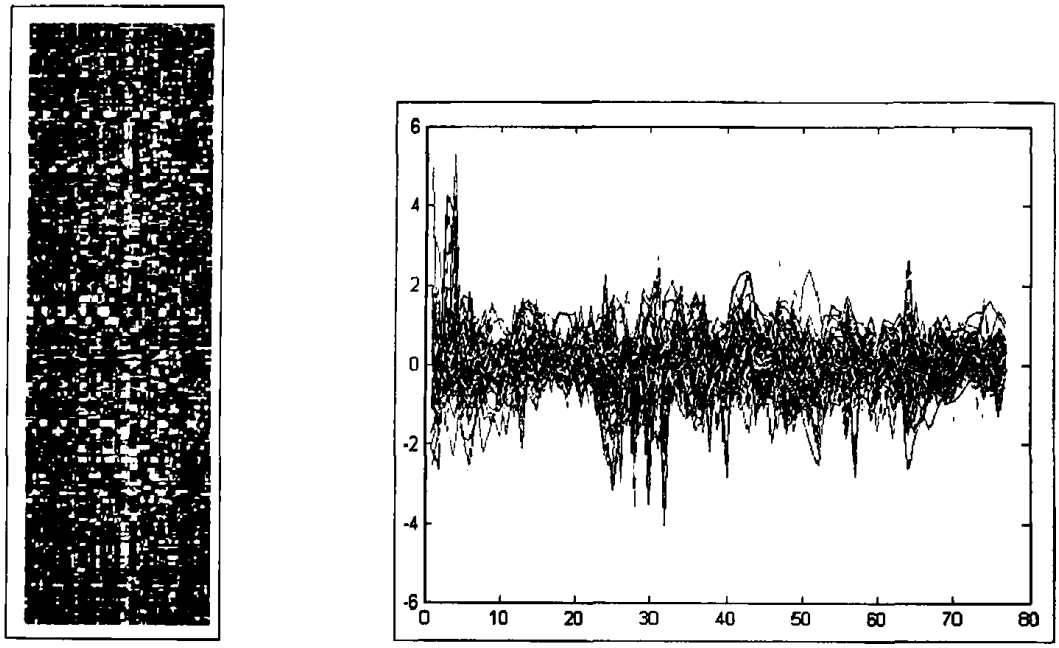
Το κριτήριο το οποίο εφαρμόστηκε είναι το εξής: αν η τυπική απόκλιση του γονιδίου x_i είναι μικρότερη από το 1/3 του μέσου της τυπικής απόκλισης όλων των γονιδίων, ή μεγαλύτερη από το πενταπλάσιο αυτής τότε το συγκεκριμένο γονιδίου το απορρίπτουμε. Εάν ένα διάνυσμα x_i ικανοποιεί το κριτήριο, το γονίδιο i λέγεται ότι είναι σημαντικό.

$$if \left(s_j \leq 0.33\bar{s} \text{ or } s_j \geq 5\bar{s} \right) then \quad remove(gene_j) \tag{7.13}$$

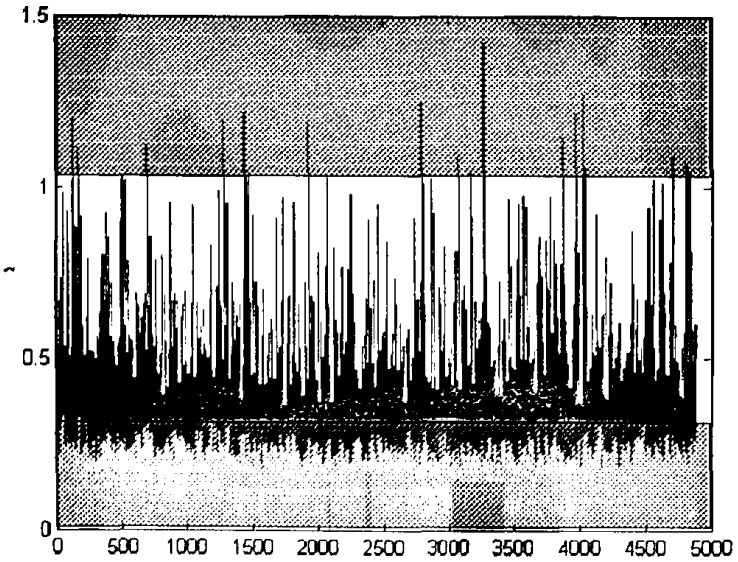
Στην Εικόνα 7-13 βλέπουμε την εφαρμογή των κατωφλίων στα δεδομένα έκφρασης γονιδίων. Εφαρμόζοντας το παραπάνω κριτήριο συνολικά απορρίφθηκαν 470 ORFs, στην Εικόνα 7-11 βλέπουμε τις εκφράσεις των γονιδίων που απορρίφθηκαν, ενώ στην εικόνα Εικόνα 7-12 βλέπουμε τις εκφράσεις μερικών από τα γονίδια που επιλέχθηκαν.



Εικόνα 7-11 Στην εικόνα βλέπουμε τις εκφράσεις των 407 ORFs, τα οποία απορρίφθηκαν κάνοντας χρήση της τυπικής απόκλισης, (άξονας y: επίπεδα έκφρασης γονιδίων, άξονας x: χρόνος)



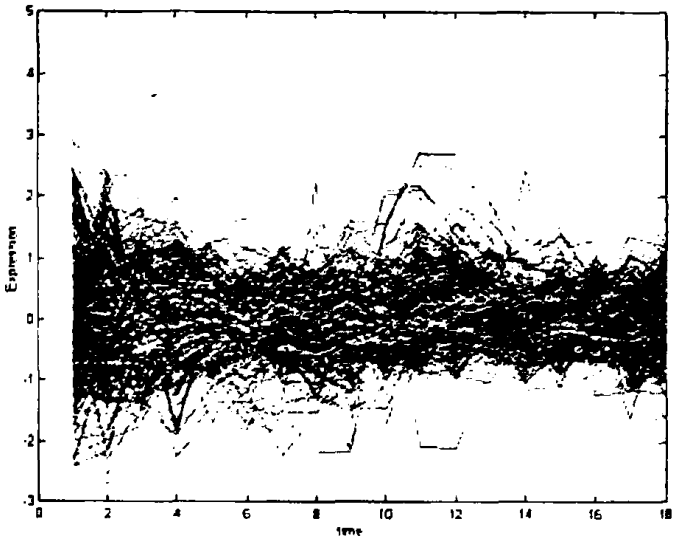
Εικόνα 7-12 Οι εκφράσεις 250 από τα γονίδια τα οποία επιλέχθηκαν μετά την εφαρμογή του κριτηρίου της τυπικής απόκλισης (άξονας y: επίπεδα έκφρασης γονιδίων, άξονας x: χρόνος)



Εικόνα 7-13 Το άνω και κάτω threshold μετά τον υπολογισμό της τυπικής απόκλισης του πίνακα των δεδομένων μας (άξονας y: τυπική απόκλιση, άξονας x: γονίδια)

7.7 Αλγόριθμος εύρεσης βέλτιστου αριθμού ομάδων

Σκοπός μας σε αυτό το βήμα είναι η εύρεση του αριθμού ομάδων στις οποίες μπορούν να κατηγοριοποιηθούν τα γονίδια που έχουν προκύψει από το προηγούμενο βήμα. Το κίνητρο για την εφαρμογή αυτού του βήματος είναι αφενός η παραγωγή μια πλήρους αυτοματοποιημένης διαδικασίας αφετέρου δε η εύρεση των λειτουργικών κατηγοριών στις οποίες ανήκουν τα γονίδια μιας και δεν υπάρχει σημαντική εκ των προτέρων γνώση για αυτές. Για τον καθορισμό του αριθμού των ομάδων θα ακολουθηθεί μία διαδικασία η οποία βασίζεται σε τροποποίηση η οποία έγινε σε μία εργασία [Kim et al. 2001] Από το βήμα αυτό και για τα επόμενα βήματα χρησιμοποιούμε το σύνολο δεδομένων που παράχθηκε από τις μετρήσεις που πραγματοποιήθηκαν κατά την διάρκεια του alpha πειράματος (βλ. Εικόνα).



Εικόνα 7-14 Δεδομένα από το alpha πείραμα (άξονας y: επίπεδα έκφρασης γονιδίων, άξονας x: χρόνος)

Στην βιβλιογραφία προτείνονται αρκετές συναρτήσεις για την εύρεση του βέλτιστου αριθμού των clusters οι οποίες καλούνται cluster validity ή validity criteria. Ο Bezdek [Bezdek et al. 1981] πρότεινε το partition coefficient υ_{PC} και το partition entropy υ_{PE} οι οποίες χρησιμοποιούν το πίνακα των βαθμών συμμετοχής. Οι Fukayama και Sugeno (υ_{FS}) [Fukayama et al. 1989] καθώς και οι Xie και Beni (υ_{XB}

) [Xie et al. 1991] πρότειναν συναρτήσεις οι οποίες λαμβάνουν υπόψη τους και γεωμετρικές ιδιότητες των δεδομένων εισόδου. Τα υ_{PE} και υ_{PC} είναι ευαίσθητα στην ύπαρξη θορύβου ενώ το υ_{FS} δεν δίνει καλά αποτελέσματα όταν το m παίρνει είτε πολύ μικρές είτε πολύ μεγάλες τιμές. Το m είναι ένας εκθέτης βάρους (weighting exponent) που εκφράζει το βαθμό ασάφειας του αποτελέσματος της ομαδοποίησης (clustering). Επιπλέον τα υ_{PC}, υ_{PE}, υ_{FS} κρίθηκαν ότι δεν είναι πλέον χρήσιμα για τον αλγόριθμο C-Means. Σύμφωνα με μελέτη που έγινε από τους Pal και Bezdek [Bezdek et al. 1981] αποδείχθηκε ότι το υ_{XB} δίνει καλά αποτελέσματα όταν το c παίρνει τιμές από 2 έως 10 και το m από 1.01 έως 7. Παρόλα αυτά το υ_{XB} φθίνει μονότονα καθώς το c παίρνει τιμές πολύ κοντά στο πλήθος των δεδομένων. Για την αποφυγή αυτής της φθίνουσας κλίσης προστέθηκε ένας *ad hoc* όρο ποινής δημιουργώντας ένα καινούργιο validity index το υ_K.

Στην εργασία [Kim et al. 2001] για την αντιμετώπιση όλων των παραπάνω προβλημάτων πρότειναν ένα νέο κριτήριο (validity criterion) το οποίο περιγράφεται στη συνέχεια. Προκείμενου να βρεθεί ο βέλτιστος αριθμός ομάδων θα πρέπει να βρεθεί για ποια τιμή του c (αριθμός ομάδων) ελαχιστοποιείται η ακόλουθη ποσότητα:

$$\upsilon_{SV}(c,V;X) = \upsilon_{uN}(c,V;X) + \upsilon_{oN}(c,V). \qquad (7.14)$$

Το c αντιπροσωπεύει τον αριθμό των ομάδων και παίρνει τιμές από 2 έως $c_{\max}$. Το $c_{\max}$ στη συγκεκριμένη εργασία παίρνει την τιμή 6.

Το $X = [x_1, x_2, \ldots, x_n]^T$ αντιπροσωπεύει το σύνολο των δεδομένων. Το πλήθος των δεδομένων n είναι όσο και τα γονίδια x πειράματα της εικόνας (~4400 * 18). Το x_i εκφράζει το διάνυσμα των χαρακτηριστικών ενός γονιδίου. Στην συγκεκριμένη περίπτωση πρόκειται για διάνυσμα διάστασης 18, όσες και οι μετρήσεις που πάρθηκαν από το πείραμα alpha.

Το $V = [v_1, v_2, \ldots, v_c]^T$ είναι ένας $c \times p$ πίνακας προτύπων, όπου p το πλήθος των χαρακτηριστικών, $p = 18$. Το v_i είναι ένα διάνυσμα διάστασης p το οποίο χαρακτηρίζει το cluster i.

Το $\upsilon_{uN}(\cdot)$ δίδεται από την Εξίσωση (7.15):

$$\upsilon_{uN}(c,V;X)=\frac{\upsilon_u(c,V;X)-\upsilon_{\min}}{\upsilon_{\max}-\upsilon_{\min}} \tag{7.15}$$

Το $\upsilon_{\max}$ και το $\upsilon_{\min}$ δίδονται από τις Εξισώσεις 3α και 3β αντίστοιχα.

$$\upsilon_{\max}=\max_c \upsilon_u(c,V;X), \tag{7.16}$$

$$\upsilon_{\min}=\min_c \upsilon_u(c,V;X) \tag{7.17}$$

Δηλαδή το $\upsilon_{\max}$ αντιστοιχεί στο μεγαλύτερο στοιχείο του διανύσματος

$$v_u=[\upsilon_u(2,V;X),\ldots.,\upsilon_u(c_{\max},V;X)],$$

και το $\upsilon_{\min}$ αντιστοιχεί στο ελάχιστο στοιχείο του ίδιου διανύσματος.

Το $\upsilon_u(c,V;X)$ δίδεται από την Εξίσωση 7.18:

$$\upsilon_u(c,V;X)=\frac{1}{c}\sum_{i=1}^{c} MD_i, \quad 2\le c\le c_{\max} \tag{7.18}$$

Το MD_i είναι η μέση απόσταση των στοιχείων της κάθε ομάδας i από το κέντρο της ομάδας (mean intra cluster distance).

Το $\upsilon_{oN}(\cdot)$ δίδεται από την Εξίσωση 7.19:

$$\upsilon_{oN}(c,V)=\frac{\upsilon_o(c,V)-\upsilon_{\min}}{\upsilon_{\max}-\upsilon_{\min}} \tag{7.19}$$

Το $\upsilon_{\max}$ και τό $\upsilon_{\min}$ ορίζονται αυτή τη φορά ως εξής:

$$\upsilon_{\max}=\max_c \upsilon_o(c,V) \tag{7.20}$$

$$\upsilon_{\min}=\min_c \upsilon_o(c,V) \tag{7.21}$$

Δηλαδή το $\upsilon_{\max}$ αντιστοιχεί στο μεγαλύτερο στοιχείο του διανύσματος

$$v_o = [\upsilon_o(2,V),\ldots,\upsilon_o(c_{\max},V)],$$

και το $\upsilon_{\min}$ αντιστοιχεί στο ελάχιστο στοιχείο του ίδιου διανύσματος.

Το $\upsilon_o(c,V)$ δίδεται από την Εξίσωση 7.22:

$$\upsilon_o = \frac{c}{d_{\min}} \tag{7.22}$$

Το $d_{\min}$ εκφράζει τη μικρότερη απόσταση μεταξύ των κέντρων των ομάδων (minimum inter cluster distance).

Η ποιοτική ερμηνεία όλων των εξισώσεων οι οποίες αναφέρονται στην παράγραφο αυτή περιγράφονται στην συνέχεια.

Το σύνολο των δεδομένων θεωρείται ότι έχει υπό κατατμηθεί (under partitioned) όταν τουλάχιστον μία ομάδα έχει μεγάλη μέση intra cluster απόσταση. Η μέση intra cluster απόσταση στην εργασία των [Kim et al. 2001] ορίζεται ως $MD_i = \sum_{x\in\chi_i} \| v_i - x \|^2 / \eta_i$ όπου η_i ο αριθμός των δεδομένων τα οποία ανήκουν στην ομάδα i και χ_i είναι ένα σύνολο δεδομένων της ομάδας i. Η αλλαγή η οποία έγινε είναι ότι η μέση απόσταση των στοιχείων της κάθε ομάδας από το κέντρο της ομάδας (mean intra cluster distance) ορίστηκε ως $\frac{1}{n_i}\sum_{k=1}^{n} u_{ik}^m \| x_k - v_i \|^2$ όπου u_{ik} ο βαθμός με τον οποίο το x_k ανήκει στο cluster i. Το n_i δίδεται από τη σχέση $n_i = \sum_{k=1}^{n} u_{ik}^m$. Σύμφωνα με τον ορισμό αυτό η Εξίσωση 7.15 μπορεί να πάρει την μορφή:

$$\upsilon_u(c,V;X) = \frac{1}{c}\sum_{i=1}^{c}\frac{1}{n_i}\sum_{k=1}^{n} u_{ik}^m \| x_k - v_i \|^2 \tag{7.23}$$

Όταν ο διαχωρισμός (partition) πλησιάζει την βέλτιστη ή την υπέρ κατετμημένη (over partitioned) κατάσταση τότε η μέση απόσταση των στοιχείων της ομάδας από το κέντρο της ομάδας (mean intra cluster distance) μειώνεται απότομα. Δεν ισχύει όμως το ίδιο και για την ελάχιστη απόσταση μεταξύ των κέντρων των ομάδων

(minimum inter cluster distance) (d_{min}) η οποία αυξάνεται όταν τα δεδομένα βρίσκονται σε βέλτιστη ή υπό κατετμημένη (under partitioned) κατάσταση ενώ παίρνει πολύ μικρή τιμή όταν τα δεδομένα μεταβαίνουν σε υπέρ κατετμημένη (over partitioned) κατάσταση. Για το λόγο αυτό δεν είναι δυνατόν να βρεθεί ένας βέλτιστος αριθμός ομάδων με χρήση των δύο αυτών αποστάσεων.

Για να ελέγξουμε αν τα δεδομένα βρίσκονται σε υπό κατετμημένη (under partitioned) κατάσταση υπολογίζουμε το $\upsilon_u(c,V;X)$. Η ποσότητα αυτή παίρνει μικρή τιμή όταν τα δεδομένα βρίσκονται σε βέλτιστη ή υπέρ κατετμημένη (over partitioned) κατάσταση, επιπλέον γίνεται μηδενική όταν ο αριθμός των ομάδων πλησιάζει τον αριθμό των δεδομένων. Όταν τα δεδομένα όμως μεταβαίνουν σε υπό κατετμημένη (under partitioned) κατάσταση η τιμή της ποσότητας αυτής γίνεται σχετικά μεγάλη. Το $\upsilon_u(c,V;X)$ έχει ένα σημείο διαχωρισμού (break point) όταν το c πλησιάζει τη βέλτιστη τιμή του.

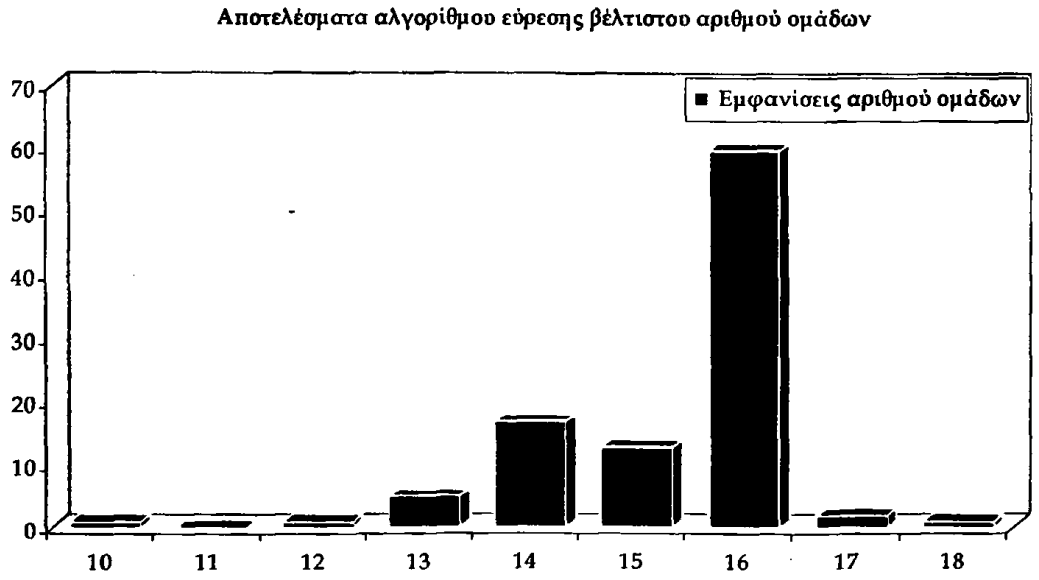
Για να ελέγξουμε αν τα δεδομένα βρίσκονται σε υπέρ κατετμημένη (over partitioned) κατάσταση υπολογίζουμε το $\upsilon_o = \frac{c}{d_{min}}$. Η ποσότητα αυτή παίρνει μικρή τιμή όταν το d_{min} παίρνει μεγάλη τιμή δηλαδή τα δεδομένα βρίσκονται σε βέλτιστη ή υπό κατετμημένη (under partitioned) κατάσταση. Όταν τα δεδομένα όμως μεταβαίνουν σε υπέρ κατετμημένη (over partitioned) κατάσταση δηλαδή το d_{min} έχει μικρή τιμή, η τιμή του $\upsilon_o = \frac{c}{d_{min}}$ γίνεται σχετικά μεγάλη. Επομένως και η ποσότητα αυτή παράγει ένα σημείο διαχωρισμού όταν ο αριθμός ομάδων πλησιάζει το βέλτιστο (c^*).

Εφόσον και οι δύο ποσότητες παίρνουν μικρή τιμή όταν το $c = c^*$ ένας κατάλληλος συνδυασμός των δύο αυτών ποσοτήτων παράγει το c^*. Ο συνδυασμός αυτών εκφράζεται μέσω της Εξίσωσης 7.14. Επειδή όμως οι συναρτήσεις αυτές ($\upsilon_u(c,V;X), \upsilon_o(c,V)$) παίρνουν τιμές σε διαφορετικά διαστήματα γίνεται κανονικοποίηση αυτών μέσω των Εξισώσεων 7.15 και 7.18 αντίστοιχα ($\upsilon_{uN}(\cdot), \upsilon_{oN}(\cdot)$).

7.7.1 Επαλήθευση του αλγορίθμου

Ο αλγόριθμος εύρεσης του αριθμού των ομάδων είναι πολύ ευαίσθητος στην αρχικοποίηση των διανυσμάτων που περιγράφουν τα αρχικά κέντρα. Για τον λόγο αυτό επιλέχθηκε η εφαρμογή του αλγορίθμου στο ίδιο σύνολο δεδομένων να γίνει 100 φορές για αριθμό ομάδων από 1 έως 32, και σαν αριθμός ομάδων να επιλεχθεί εκείνη η τιμή που εμφανίστηκε περισσότερες φορές. Ο αλγόριθμος εφαρμόστηκε στο σύνολο των δεδομένων του alpha factor (18 time points)

Στα αποτελέσματα στην παρακάτω εικόνα παρατηρούμε ότι στις περισσότερες εκτελέσεις και για τα τρία σύνολα δεδομένων ο αριθμός των ομάδων που είχαμε στην έξοδο ήταν 16.



Εικόνα 7-15 Αποτελέσματα επαλήθευσης αλγορίθμου εύρεσης αριθμού ομάδων

7.8 Ομαδοποίηση των προτύπων έκφρασης γονιδίων

Στόχος μας σε αυτό το βήμα είναι να προσδιορίσουμε τα γονίδια των οποίων οι εκφράσεις ακολουθούν κάποιο κοινό πρότυπο. Η προσπάθειά μας έχει εστιαστεί στον προσδιορισμό των προτύπων χρησιμοποιώντας τεχνικές ομαδοποίησης για σημεία στο πολυδιάστατο χώρο, όπου τα σημεία (διανύσματα) στο πολυδιάστατο χώρο αντιστοιχούν στα επίπεδα έκφρασης των γονιδίων. Είναι γνωστό από την βιολογία ότι τα γονίδια που εκφράζονται με όμοιο τρόπο σε μια βιολογική διαδικασία έχουν και παρόμοιες ιδιότητες-λειτουργίες μέσα στο κύτταρο. Την διαπίστωση αυτή θέλουμε να εκμεταλλευτούμε ώστε να δημιουργήσουμε ομάδες γονιδίων με βάση τις λειτουργίες τους, οι ομάδες αυτές θα χρησιμοποιηθούν ως είσοδοι σε στοχαστικά μοντέλα με σκοπό την εξαγωγή του γενετικού δικτύου. Υπάρχουν δύο λόγοι για τους οποίους ακολουθούμε την τεχνική ομαδοποίησης πριν την χρήση των στοχαστικών μοντέλων, ο πρώτος λόγος είναι ότι χρησιμοποιώντας όλα τα γονίδια (τα οποία έχουν προκύψει από τα προηγούμενα βήματα) ως είσοδο στον αλγόριθμο εκπαίδευσης του γραφικού μοντέλου αυξάνεται κατά πολύ η πολυπλοκότητα του αλγορίθμου με αποτέλεσμα η εύρεση του μοντέλου να μην είναι βέλτιστη. Ο δεύτερος λόγος έχει να κάνει με την δήλωση ότι "τα γονίδια που εκφράζονται με όμοιο τρόπο σε μια βιολογική διαδικασία ότι έχουν και παρόμοιες ιδιότητες-λειτουργίες μέσα στο κύτταρο", το οποίο μας οδηγεί στο συμπέρασμα ότι χρησιμοποιώντας ομάδες συν-εκφραζόμενων γονιδίων μπορούμε να πετύχουμε ένα πιο καλό μοντέλο για τα δεδομένα μας.

Στις μεθόδους ομαδοποίησης δεδομένων έκφρασης γονιδίων και πιο συγκεκριμένα του Yeast η μέθοδος η οποία έχει κυριαρχήσει είναι εκείνη της ιεραρχικής ομαδοποίησης (Εικόνα 1.2). Ο λόγος για τον οποίο δεν είναι δυνατόν να χρησιμοποιηθεί αυτή η μέθοδος στην μεθοδολογία που αναπτύσσεται στο κεφάλαιο αυτό είναι ότι η επιλογή κάποιου threshold με βάση το οποίο θα επιλεγόντουσαν οι ομάδες θα ήταν εμπειρικό εφόσον δεν υπάρχει κάποια εκ των προτέρων γνώση πάνω σε αυτό

Για την απεικόνιση των αποτελεσμάτων έγινε προβολή του πολυδιάστατου χώρου σε 2-διαστατο με χρήση της μεθόδου PCA από όπου και εξάχθηκαν τα 2 πιο ισχυρά components.

7.8.1 Εισαγωγή - ομαδοποίηση δεδομένων

- Η ομαδοποίηση είναι η οργάνωση μιας συλλογής από δεδομένα σε ομάδες με βάση κάποιο μέτρο ομοιότητας. Τα δεδομένα συνήθως περιγράφονται σαν διανύσματα τιμών κάποιων μέτρων ή αναπαριστώνται ως σημεία σε έναν πολυδιάστατο χώρο. Δεδομένα που ανήκουν στην ίδια ομάδα παρουσιάζουν μεγαλύτερη ομοιότητα από ότι δεδομένα που ανήκουν σε διαφορετικές ομάδες.

Η ομαδοποίηση είναι μια διαδικασία που εντάσσεται γενικότερα στην κατηγοριοποίηση χωρίς επίβλεψη. Στην κατηγοριοποίηση με επίβλεψη ένα σύνολο από προ-ομαδοποιημένα δεδομένα είναι διαθέσιμο, και αυτό που μας ζητείται είναι να εντάξουμε ένα νέο στοιχείο σε κάποια από τις υπάρχουσες ομάδες. Συνήθως τα προ-ομαδοποιημένα δεδομένα χρησιμοποιούνται για να περιγράψουν τις διαφορετικές ομάδες - κλάσεις στις οποίες θα εντάξουμε νέα δεδομένα. Αντίθετα στην ομαδοποίηση το πρόβλημα είναι να ομαδοποιήσουμε σε λογικές κλάσεις τα δεδομένα μας, χωρίς καμία γνώση για προϋπάρχουσες ομάδες. Έτσι μπορούμε να πούμε ότι η κατηγοριοποίηση είναι απόλυτα εξαρτώμενη από τα δεδομένα.

7.8.1.1 Μέτρα Ομοιότητας Κατηγοριών

Σε κάθε ομάδα τα δεδομένα που περιέχονται σε αυτό παρουσιάζουν ομοιότητα μεταξύ τους και αυτό είναι βασικό για να ορισθεί μια ξεχωριστή ομάδα. Έτσι για όλες τις τεχνικές ομαδοποίησης είναι σημαντικό να ορίζεται ένα μέτρο ομοιότητας μεταξύ δύο στοιχείων από το χώρο δεδομένων. Δεδομένης της μεγάλης ποικιλίας στα χαρακτηριστικά των στοιχείων η επιλογή του μέτρου ομοιότητας θα πρέπει να είναι πολύ προσεγμένη. Σε πολλές περιπτώσεις αυτό που συνήθως μετράται δεν είναι η ομοιότητα αλλά η διαφορετικότητα δυο στοιχείων. Υπάρχουν μέτρα ομοιότητας τα οποία είναι ευρέως διαδεδομένα, και χρησιμοποιούντα για την σύγκριση στοιχείων των οποίων τα χαρακτηριστικά περιγράφονται από συνεχείς τιμές. Το μέτρο

ομοιότητας καλείται και απόσταση και ικανοποιεί την τριγωνική ανισότητα για δύο στοιχεία x,y:

$$D(x,x) = 0 \tag{7.24}$$

$$D(x,y) = D(y,x) \tag{7.25}$$

$$D(x,y) = D(x,z) + D(z,y) \tag{7.26}$$

Το πιο γνωστό μέτρο ομοιότητας που χρησιμοποιείται είναι η Ευκλείδεια απόσταση η οποία ορίζεται ως εξής:

$$D(x,y) = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \tag{7.27}$$

Άλλοι τύποι που δίνουν την απόσταση μεταξύ δύο στοιχείων μπορεί να είναι η απόσταση Manhattan:

$$D(x,y) = \sum_{i=1}^{k} | x_i - y_i | \tag{7.28}$$

ή το μέγιστο της διαφοράς σε κάθε διάσταση:

$$D(x,y) = \max_{i=1}^{k} | x_i - y_i | \tag{7.29}$$

Η ευκλείδεια απόσταση χρησιμοποιείται ευρέως σε περιπτώσεις λίγων διαστάσεων και έχει καλά αποτελέσματα όταν δεδομένα κατηγοριοποιούνται σε συμπαγή και αρκετά απομονωμένα clusters. Ένα πρόβλημα που παρουσιάζει είναι ότι στις πολλές διαστάσεις το χαρακτηριστικό το οποίο παρουσιάζει την μεγαλύτερη διαφοροποίηση από τα άλλα κυριαρχεί και αποπροσανατολίζει το τελικό αποτέλεσμα. Εδώ πρόκειται για αυτό που συνήθως αναφέρεται ως *"κατάρα των πολλών διαστάσεων"* (curse of dimensionality).

Μερικοί αλγόριθμοι αντί να υπολογίζουν κάθε φορά την απόσταση μεταξύ δύο στοιχείων, χρησιμοποιούν ένα πίνακα στον οποίο τοποθετούν τις ομοιότητες των στοιχείων. Αυτό που γίνεται είναι ένας προ-υπολογισμός των n(n-1)/2 τιμών ομοιότητας για ένα σύνολο n στοιχείων.

7.8.1.2 Τεχνικές Ομαδοποίησης

Οι τεχνικές ομαδοποίησης μπορούν να διαχωριστούν με πολλούς τρόπους, όπως Ιεραρχική ομαδοποίηση σε αντίθεση με τη διαμεριστική, και καθένα από αυτά να χωριστεί σε άλλες υποκατηγορίες. Θα αναφερθούμε σε διαφορετικές προσεγγίσεις ομαδοποίησης παρακάτω αφού προηγουμένως δούμε κάποιους όρους και διαφοροποιήσεις που παρουσιάζουν οι διάφοροι αλγόριθμοι. Οι αλγόριθμοι για ομαδοποίηση μπορεί να είναι:

Συγκεντρωτικοί και Διαχωριστικοί (Agglomerative and Divisive). Η διαφοροποίηση των ειδών αυτών σχετίζεται με την λειτουργία και τις δομές του αλγορίθμου. Στην πρώτη περίπτωση ο αλγόριθμος ξεκινά θεωρώντας κάθε στοιχείο σαν ένα ξεχωριστό cluster, και προχωρά συγχωνεύοντας δεδομένα και clusters μέχρις ότου να ικανοποιηθεί μια συνθήκη. Στην περίπτωση ενός διαχωριστικού αλγορίθμου, όλα τα δεδομένα θεωρούνται ότι ανήκουν σε ένα cluster και ακολουθείται μια συνεχής διάσπαση του cluster αυτού σε υπό cluster μέχρις ότου να ικανοποιηθεί η συνθήκη τερματισμού.

Μονοθετικοί και Πολυθετικοί (Monothetic and Polythetic). Η διαφορά αυτών χαρακτηρίζει την σειριακή ή ταυτόχρονη χρησιμοποίηση των χαρακτηριστικών των στοιχείων κατά την διαδικασία της Ομαδοποίησης. Οι περισσότεροι αλγόριθμοι είναι πολυθετικοί, κάτι που σημαίνει ότι όλα τα χαρακτηριστικά των στοιχείων συμμετέχουν κάθε φορά στον καθορισμό της απόστασης του στοιχείου από κάποιο άλλο. Ένας μονοθετικός αλγόριθμος λαμβάνει υπόψη του μόνο ένα χαρακτηριστικό τη φορά και πραγματοποιεί ομαδοποιήσεις με βάση αυτό το χαρακτηριστικό. Σε επόμενη επανάληψη χρησιμοποιεί άλλο χαρακτηριστικό και διαχωρίζει τις ήδη υπάρχουσες ομάδες. Το πρόβλημα αυτών των αλγορίθμων είναι ότι τα δεδομένα χωρίζονται τελικά σε 2^d clusters όπου d είναι ο αριθμός των χαρακτηριστικών των στοιχείων. Αυτό συνήθως οδηγεί σε πολλά clusters εκ των οποίων τα περισσότερα είναι μικρά και ασήμαντα.

Σκληροί και ασαφείς (hard and fuzzy). Ένας σκληρός αλγόριθμος τοποθετεί κάθε στοιχείο σε ένα και μόνο cluster, σε αντίθεση με τους fuzzy αλγορίθμους οι οποίοι δίνουν σε κάθε στοιχείο για κάθε cluster έναν βαθμό που εκφράζει κατά πόσο το στοιχείο αυτό ανήκει στο cluster αυτό.

Ντετερμινιστικοί και Στοχαστικοί (Deterministic and Stochastic): Αυτοί οι αλγόριθμοι είναι κυρίως διαιρετικοί και σχετίζονται με την βελτιστοποίηση της ομαδοποίησης.

Αυξυντικοί και μη αυξιντικοί (incremental and non-incremental): Η διαφορά αυτών των αλγορίθμων εμφανίζεται όταν το σύνολο των δεδομένων προς ομαδοποίηση είναι πολύ μεγάλο και περιορισμοί που υπάρχουν στον χρόνο εκτέλεσης και τον διαθέσιμο χώρο μνήμης επηρεάζουν την αρχιτεκτονική του αλγορίθμου. Στα πρώτα βήματα της θεωρίας περί ομαδοποίησης τα δεδομένα δεν ήταν ιδιαίτερα πολλά και προβλήματα με το μέγεθος της πληροφορίας δεν υπήρχαν. Με την αύξηση όμως της πληροφορίας υπήρξε η ανάγκη για εύρεση αλγορίθμων οι οποίοι ελαχιστοποιούν τον αριθμό σαρώσεων των δεδομένων, μειώνουν τον αριθμό των στοιχείων που εξετάζονται η μειώνουν το μέγεθος των δομών που χρησιμοποιούνται κατά την εκτέλεση του αλγορίθμου.

7.8.2 Αναπαράσταση αποτελεσμάτων ομαδοποίησης

Η διάσταση του χώρου στον οποίο εφαρμόστηκαν όλοι οι αλγόριθμοι ομαδοποίησης (που θα δούμε παρακάτω) είναι 18 (όσα και τα διαστήματα στα οποία πάρθηκαν μετρήσεις κατά την διάρκεια των πειραμάτων). Για την αναπαράσταση των αποτελεσμάτων της ομαδοποίησης, και την γραφική αναπαράσταση των ομάδων έγινε ελάττωση του χώρου από τον $\Re^{18}$ στον $\Re^2$ κάνοντας χρήση μεθόδων γνωστές ως μέθοδοι "Ανάλυσης Κυριοτέρων Συνιστωσών" (PCA) και πιο συγκεκριμένα της μεθόδου SVD. Παρακάτω περιγράφεται εν συντομία η μέθοδος αυτή.

Η SVD μέθοδος είναι βασισμένη στο παρακάτω θεώρημα της Γραμμικής Άλγεβρας:

"Οποιοσδήποτε M x N πίνακας A του οποίου ο αριθμός των γραμμών είναι μεγαλύτερος ή ίσος από τον αριθμό των στηλών N μπορεί να γραφτεί σαν ένα γινόμενο ενός M x N ορθογώνιου πίνακα U , και N x N διαγώνιου πίνακα D ιδιοτιμών και του ανάστροφου ενός N x N ορθογώνιου πίνακα V:

$$\begin{pmatrix} & A & \end{pmatrix} = \begin{pmatrix} & U & \end{pmatrix} * \begin{pmatrix} d_1 & & \\ & d_2 & \\ & & d_3 \end{pmatrix} * \begin{pmatrix} & V & \end{pmatrix} \text{ή } A = UDV^T \quad (7.30)$$

όπου $UU^T = VV^T = \mathrm{I}$

και ο W είναι διαγώνιος πίνακας του οποίου τα δεδομένα είναι οι ιδιοτιμές του αρχικού πίνακα.

Για την προβολή των αποτελεσμάτων ομαδοποίησης χρησιμοποιήθηκαν οι δύο πιο σημαντικές ιδιοτιμές οι οποίες υπολογίζονται από την σχέση:

$$\{d_1, d_2\} = \max(D) \tag{7.31}$$

7.8.3 Εφαρμογή αλγορίθμων ομαδοποίησης

Στις παραγράφους που ακολουθούν εφαρμόσαμε τρεις αλγορίθμους ομαδοποίησης, τον KMeans που ανήκει στην κατηγορία "hard clustering" και τους Fuzzy KMeans και Mixture Models που ανήκουν στην κατηγορία "soft clustering". Σκοπός μας σε αυτό το βήμα είναι αφενός να παράγουμε τα δεδομένα τα οποία θα αποτελέσουν την είσοδο στο βήμα για την ανακατασκευή του γενετικού δικτύου, αφετέρου να συγκρίνουμε τα αποτελέσματά τους γνωρίζοντας τις ιδιαιτερότητες και τα χαρακτηριστικά του κάθε αλγορίθμου. Στις τρεις παραγράφους που ακολουθούν περιγράφεται με λεπτομέρεια ο τρόπος με τον οποίο χρησιμοποιήθηκαν οι τρεις αλγόριθμοι καθώς και τα αποτελέσματα που πήραμε από τον καθένα. Το σύνολο δεδομένων μας αποτελείτε πλέον από ~4400 ORFs στα οποία έγιναν μετρήσεις σε 18 διαφορετικές χρονικές στιγμές. Στην Εικόνα 7-14 φαίνεται το πλήρες σύνολο δεδομένων.

7.8.4 Εφαρμογή του αλγορίθμου KMeans

Στον αλγόριθμο KMeans χρησιμοποιήσαμε τη συνάρτηση τετραγωνικού λάθους (squared error function). Αυτή η συνάρτηση ορίζεται ως εξής:

$$e^2(K) = \sum_{j=1}^{K} \sum_{i=1}^{n} \| x_i^{(j)} - v_j \|^2 \tag{7.32}$$

όπου S το σύνολο δεδομένων και K ο αριθμός των ομάδων, $x_i^{(j)}$ είναι το i γονίδιο της j ομάδας, και το v_j είναι το κεντροειδές της j ομάδας. Ο αλγόριθμος ξεκινά με τυχαία

αρχικοποίηση των κεντροειδών και τοποθετεί τα δεδομένα στις ομάδες με βάση την απόσταση των στοιχείων από το κεντροειδές της ομάδας. Αυτή η διαδικασία σταματάει μέχρι η διαφοροποίηση των ομάδων (που εκπροσωπούνται από τα κεντροειδές τους) να είναι μικρή ή αν ξεπεραστεί ο μέγιστος αριθμός επαναλήψεων.

Ο αλγόριθμος αυτός είναι ιδανικός εξαιτίας της χαμηλής πολυπλοκότητάς του η οποία είναι της τάξης $n^*(O(n))$, όπου n είναι ο αριθμός των δεδομένων. Το πρόβλημα που έχει αυτός ο αλγόριθμος είναι στην αρχική επιλογή των ομάδων. Αν η επιλογή αυτή δεν γίνει σωστά το κριτήριο τετραγωνικού λάθους συγκλίνει σε τοπικά ελάχιστο κάνοντας την τελική επιλογή cluster ανεπιτυχή.

KMeans μέθοδος ομαδοποίησης

1. Επιλογή k=16 κεντροειδών ομάδων τα οποία αποτελούν και τα μόνα δεδομένα των k επιλεγμένων ομάδων.

2. Τοποθέτησε κάθε στοιχείο στο πιο κοντινό cluster μετά από υπολογισμό της απόστασης του σημείου από το κεντροειδές της κάθε ομάδας.

3. Υπολόγισε το νέο κεντροειδές από τον τύπο:

   $v_i^{l+1} = \frac{1}{N_i^l}\sum_{i=1}^{N_i^l} x_i$, η απόσταση υπολογίζεται από τον τύπο $d_E = \| x_j - v_i \|^2$

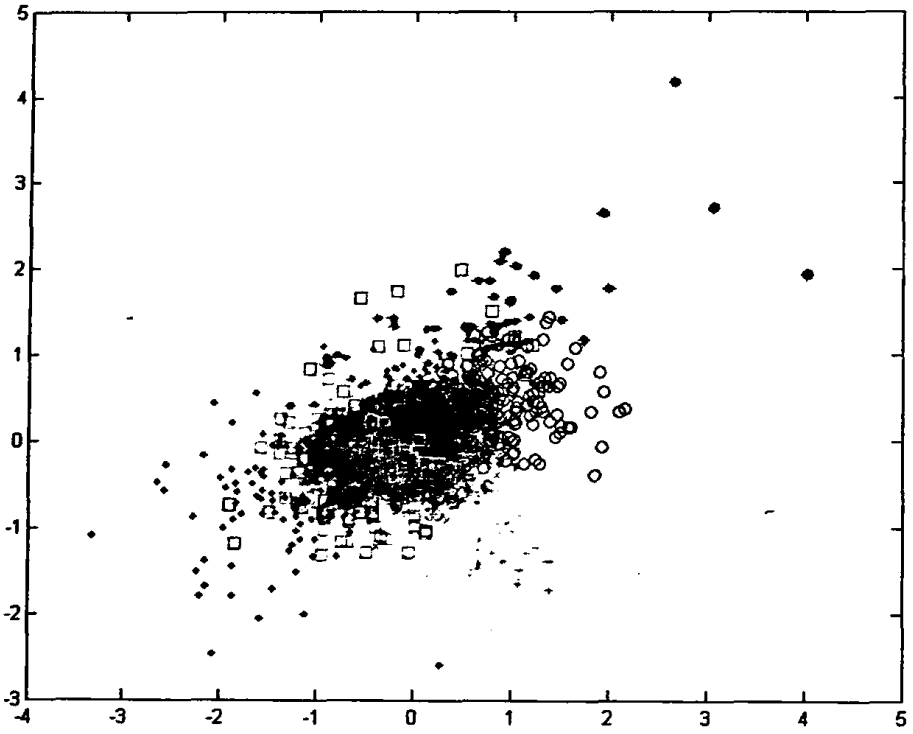
4. Αν το κριτήριο τερματισμού

$$| \vec{v}_i^l - \vec{v}_i^{l+1} | < e, \forall i \qquad (7.33)$$

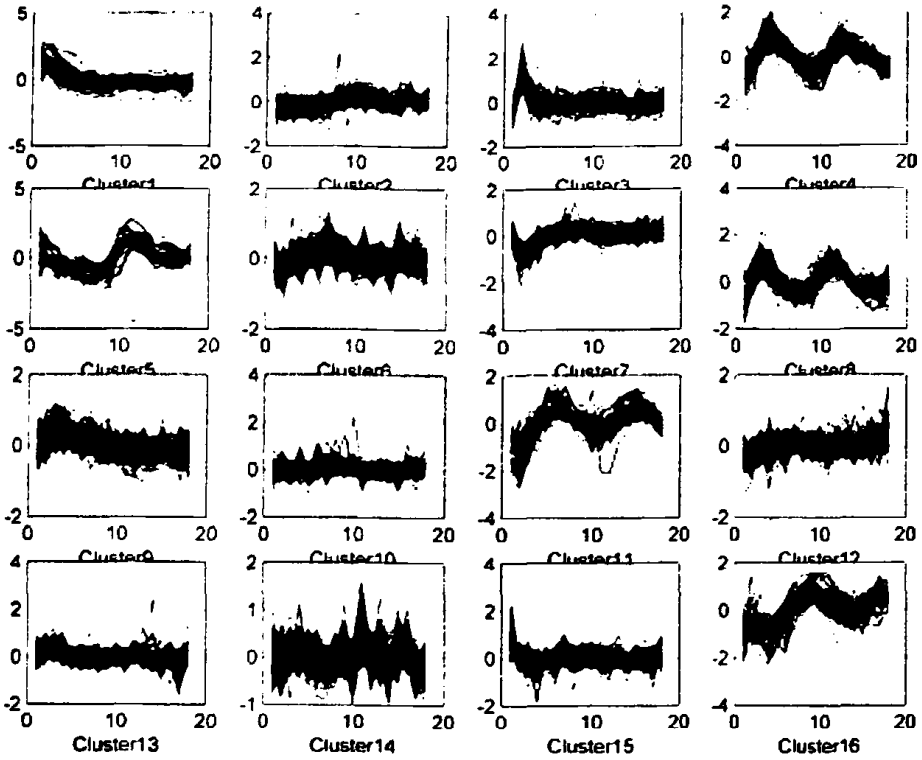
   δεν ικανοποιείται πήγαινε στο βήμα 2.

7.8.4.1 Αποτελέσματα του KMeans

Για την προβολή των αποτελεσμάτων έγινε προβολή του πολυδιάστατου χώρου σε 2-διαστατο με χρήση της PCA όπου χρησιμοποιήθηκαν τα 2 πιο ισχυρά components.



Εικόνα 7-16 Ομαδοποίηση γονίδιων χρησιμοποιώντας γονιδιακες εκφράσεις παραγόμενες από το alpha factor, η προβολή έγινε με χρήση της PCA όπου ο χώρος των δεδομένων ελαττώθηκε από τον $\Re^{18} \rightarrow \Re^{2}$



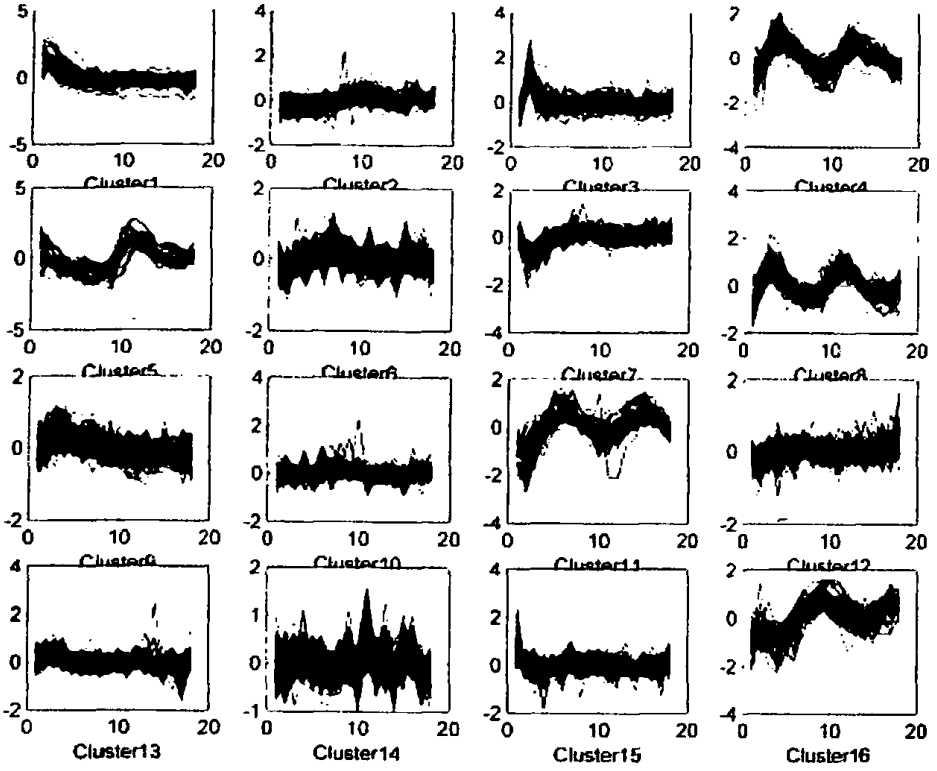
Εικόνα 7-17 Στο παραπάνω γράφημα φαίνονται οι γονιδιακές εκφράσεις των γονιδίων ανά cluster όπως αυτά έχουν διαμορφωθεί εκτελώντας τον KMeans

7.8.5 Εφαρμογή του Fuzzy KMeans

Ο Fuzzy KMeans σε αντίθεση με τον KMeans παράγει ένα περισσότερο ασαφή ορισμό των ομάδων δίνοντας ένα βαθμό συμμετοχής (πιθανότητα) σε κάθε γονίδιο για κάθε ομάδα. Οι αλγόριθμοι ομαδοποίησης που τοποθετούν ένα στοιχείο σε μια και μόνο ομάδα, όπως ο KMeans, ονομάζονται σκληροί αλγόριθμοι και αυτό συνεπάγεται ότι οι ομάδες σε αυτή τη περίπτωση είναι ξένα σύνολα μεταξύ τους. Η διαφορά του Fuzzy KMeans με τον KMean αναφέρεται στην συνάρτηση ελαχιστοποίησης όπου σε αυτή την περίπτωση είναι η εξής:

$$j = \sum_{i=1}^{C} \sum_{k=1}^{N} u_{ik}^{q} \| x_k - v_i \|^2 \qquad (7.34)$$

όπου u_{ik} είναι ο βαθμός συμμετοχής και q είναι ο βαθμός ασάφειας (εάν q=0 τότε έχουμε τον KMeans ενώ εάν q=1 τότε έχουμε αρκετά μεγάλη ασάφεια)



Εικόνα 7-17 Στο παραπάνω γράφημα φαίνονται οι γονιδιακές εκφράσεις των γονιδίων ανά cluster όπως αυτά έχουν διαμορφωθεί εκτελώντας τον KMeans

7.8.5 Εφαρμογή του Fuzzy KMeans

Ο Fuzzy KMeans σε αντίθεση με τον KMeans παράγει ένα περισσότερο ασαφή ορισμό των ομάδων δίνοντας ένα βαθμό συμμετοχής (πιθανότητα) σε κάθε γονίδιο για κάθε ομάδα. Οι αλγόριθμοι ομαδοποίησης που τοποθετούν ένα στοιχείο σε μια και μόνο ομάδα, όπως ο KMeans, ονομάζονται σκληροί αλγόριθμοι και αυτό συνεπάγεται ότι οι ομάδες σε αυτή τη περίπτωση είναι ξένα σύνολα μεταξύ τους. Η διαφορά του Fuzzy KMeans με τον KMean αναφέρεται στην συνάρτηση ελαχιστοποίησης όπου σε αυτή την περίπτωση είναι η εξής:

$$j = \sum_{i=1}^{C} \sum_{k=1}^{N} u_{ik}^{q} \| x_k - v_i \|^2 \qquad (7.34)$$

όπου u_{ik} είναι ο βαθμός συμμετοχής και q είναι ο βαθμός ασάφειας (εάν q=0 τότε έχουμε τον KMeans ενώ εάν q=1 τότε έχουμε αρκετά μεγάλη ασάφεια)

Ο αλγόριθμος fuzzy clustering εφαρμόστηκε ως εξής:

1. Επιλογή μιας fuzzy διαμέρισης των N γονιδίων σε K=16 clusters. Καθορισμός του πίνακα U=NxK του οποίου κάθε στοιχείο u_{ij} δηλώνει τον βαθμό συμμετοχής του στοιχείου i στο cluster j. Η τιμές των u είναι μεταξύ 0 και 1.

2. Χρησιμοποιώντας τον πίνακα U βρίσκεται η τιμή της συνάρτησης που αποτελεί και το κριτήριο τερματισμού, και η οποία πρέπει να βελτιστοποιηθεί. Συνεχώς επανατοποθετούμε δεδομένα στις ομάδες με νέες τιμές συμμετοχής και επαναπροσδιορίζουμε τον πίνακα U και την τιμή της συνάρτησης.

$$j = \sum_{i=1}^{C} \sum_{k=1}^{N} u_{ik}^{q} \| x_k - v_i \|^2 \tag{7.35}$$

όπου το u_{ik} υπολογίζεται από την εξίσωση:

$$u_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{d_{ik}}{d_{jk}} \right)^{2/(q-1)}} \tag{7.36}$$

η απόσταση υπολογίζεται με βάση την εξίσωση (Ευκλείδεια απόσταση):

$$d_{ik} = \| x_k - v_i \| \tag{7.37}$$

ενώ υπάρχει ο περιορισμός:

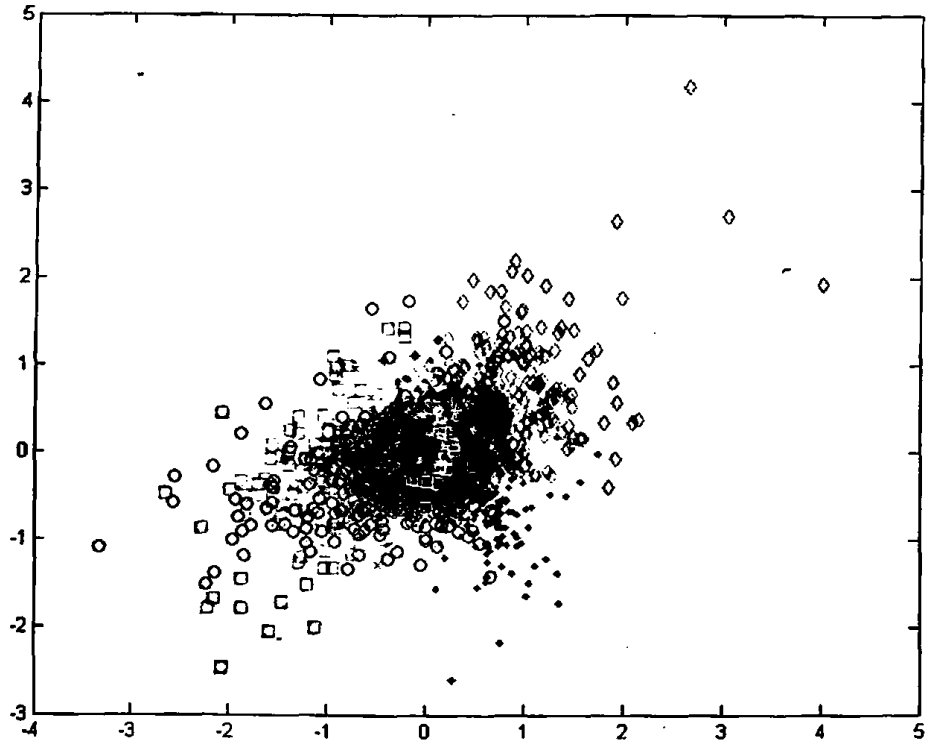
$$\sum_{i=1}^{C} u_{ik} = 1 \tag{7.38}$$

(u_{ik} είναι ο βαθμός συμμετοχής και q ο βαθμός ασάφειας)

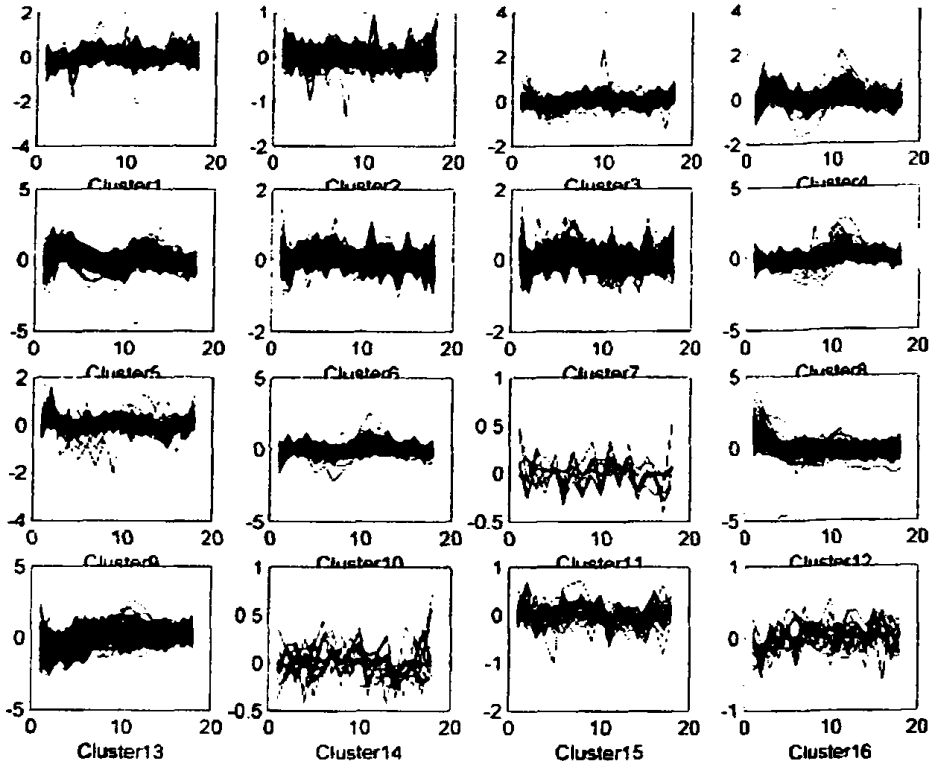
3. Επαναλαμβάνουμε το βήμα 2 μέχρι να μην επέρχονται σημαντικές αλλαγές στον πίνακα U και την τιμή της συνάρτησης.

7.8.5.1 Αποτελέσματα του Fuzzy KMeans

Για την προβολή των αποτελεσμάτων έγινε προβολή του πολυδιάστατου χώρου σε 2-διαστατο με χρήση της PCA όπου χρησιμοποιήθηκαν τα 2 πιο ισχυρά components.



Εικόνα 7-18 Ομαδοποίηση γονίδιων χρησιμοποιώντας γονιδιακες εκφράσεις παραγόμενες από το alpha factor, η προβολή έγινε με χρήση της PCA όπου ο χώρος των δεδομένων ελαττώθηκε από τον $\Re^{18} \rightarrow \Re^{2}$



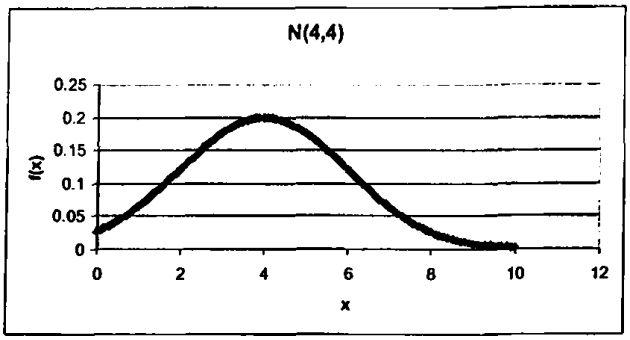
Εικόνα 7-19 Στο παραπάνω γράφημα φαίνονται οι γονιδιακές εκφράσεις των γονιδίων ανά cluster όπως αυτά έχουν διαμορφωθεί εκτελώντας τον Fuzzy KMeans

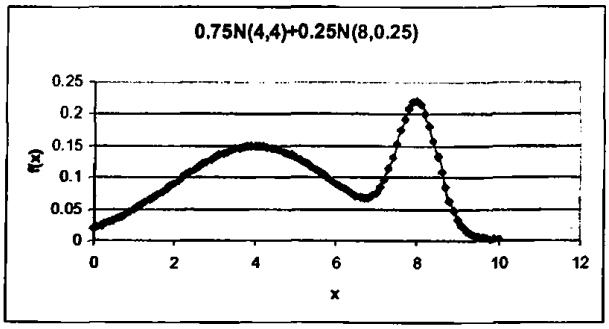
7.8.6 Μοντελοποίηση με χρήση Gaussian Mixture Models και του αλγορίθμου EM

Αυτή η μέθοδος κατηγοριοποίηση βασίζεται στην ύπαρξη ή μη κατάλληλων μοντέλων τα οποία εκφράζουν την πηγή των δεδομένων. Ας θεωρήσουμε μια συνεχή μεταβλητή, x. Μια κανονική κατανομή στην μεταβλητή x δίνεται από τον ακόλουθο τύπου συνάρτησης κατανομής πιθανότητας

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}, \qquad (7.39)$$

Αυτή είναι μια κανονική με δύο παραμέτρους, τον μέσο μ της κατανομής και την διακύμανση σ^2. Η γραφική παράσταση φαίνεται στο παρακάτω σχήμα





Εικόνα 7-20 N(4,4)

Εικόνα 7-21 Μια μεικτή κατανομή σε μια μεταβλητή με 4 παραμέτρους

Μια μεικτή κατανομή από k κανονικές κατανομές θα δίνεται από τον τύπο

$$f(x|\mu,\Sigma_i)=\sum_{i=1}^{k}\pi_i N(x|\mu_i,\Sigma_i)$$

$$\sum_{i=1}^{k}\pi_i=1 \qquad (7.40)$$

$$N(x|\mu,\Sigma)=\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}e^{\frac{-(x-\mu)\Sigma^{-1}(x-\mu)^T}{2}}$$

όπου οι συντελεστές π_i συχνά καλούνται ***βάρη***. Η τιμή του **k** συνήθως αντιπροσωπεύει τον αριθμό των ομάδων. Το μεικτό μοντέλο στο τελευταίο σχήμα έχει k=2 και 4 παραμέτρους.

Τέτοια μοντέλα μπορούν να γενικευτούν και σε πολλές διαστάσεις/μεταβλητές, οι παράμετροι υπολογίζονται μέσο των μεθόδων μέγιστης πιθανοφάνειας, εφαρμόζοντας τον αλγόριθμο EM. Εντούτοις, εάν οι παράμετροι συσχετισμού/διακύμανσης (correlation/covariance) συμπεριλαμβάνονται στο γενικό μοντέλο μεικτών κατανομών ο αριθμός των παραμέτρων αυξάνεται σημαντικά ανάλογα με τον αριθμό των μεταβλητών. Γενικά, εάν εργαζόμαστε στον

n-διάστατο χώρο θα υπάρξουν n(n+3)/2 παράμετροι για κάθε κεντροειδές. π.χ. Εάν n=10, και k=5, χρειαζόμαστε 325 παραμέτρους. Ως εκ τούτου η εκπαίδευση μπορεί να είναι αναποτελεσματική εκτός και αν τα δεδομένα είναι πάρα πολλά, οπότε σ' αυτή την περίπτωση η εκπαίδευση θα είναι πολύ αργή.

Παραπάνω θεωρήσαμε μόνο μια διάσταση. Υπάρχουν 2 προσεγγίσεις για να δουλέψουμε με παραπάνω διαστάσεις

1. Μπορούμε να κάνουμε την υπόθεση ότι κάθε διάσταση είναι ανεξάρτητη. Έτσι έχουμε:

$$p(c_i \mid x) = \frac{p(x \mid c_i)p(c_i)}{p(x)} = \frac{p(c_i)\prod_{l=1}^{d} p(x^l \mid c_i)}{p(x)}$$, όπου το d είναι ο αριθμός διαστάσεων.

Αυτό απαιτεί να υπολογίσουμε τις πιθανότητες σε όλες τις διαστάσεις d χωριστά και να τις πολλαπλασιάσουμε στο Ε-βήμα. Και όμοια μπορούμε να υπολογίσουμε τα $\mu_{c_i}^l, \sigma_{c_i}^l, p^l(c_i)$ $(l = 1...d)$ανά-διάστασης στο Μ-Βήμα.

2. Επίσης μπορούμε να δοκιμάσουμε να μοντελοποιήσουμε τα d-διαστάσεων δεδομένα απευθείας θεωρώντας μια κανονική κατανομή. Σε αυτήν την περίπτωση όλοι οι τύποι παραμένουν ουσιαστικά οι ίδιοι, εκτός του ότι πρέπει να υπολογίσουμε τον πίνακα συμεταβλητότητας (covariance matrix) Σ αντί της απόκλισης, και ένα d-διάστατο μέσο, και κάνοντας χρήση της πολυωνυμικής κανονικής κατανομής να υπολογίσουμε τις πιθανότητες όπως δίνονται παρακάτω:

$$p(x \mid c_i) = f\left(x \mid \mu_{c_i}, \Sigma_{c_i}\right) = \frac{1}{(2\pi)^{d/2} \mid \Sigma_{c_i} \mid^{1/2}} e^{\frac{-(x-\mu_{c_i})\Sigma_{c_i}^{-1}(x-\mu_{c_i})^T}{2}} \qquad (7.41)$$

7.8.6.1 Εφαρμογή των Gaussian Mixture Models και χρήση του αλγορίθμου ΕΜ στα δεδομένα έκφρασης γονιδίων

Η εφαρμογή της μεθόδου κατηγοριοποίησης έγινε στα δεδομένα που παράγονται κατά την διάρκεια του alpha συγχρονισμού. Σύμφωνα με την παραπάνω θεωρία,

υποθέσαμε ότι η πιθανότητα το $x_i = \left(x_i^1, \ldots x_i^n\right)$ γονίδιο όπου n ο αριθμός των μετρήσεων να έχει προκύψει από το μοντέλο $H(\theta)$ δίνεται από τον τύπο:

$$p(x|\theta) = \sum_{i=1}^{M} \pi_i \frac{1}{(2\pi)^{d/2} |\sum_i|^{1/2}} * e^{\left(\frac{-(x-\mu_i)^T \sum_i^{-1} (x-\mu_i)^T}{2}\right)} \tag{7.42}$$

ή για συντομία γράφουμε

$$p(x|\theta) = \sum_{i=1}^{M} \pi_i N(x|\mu_i, \Sigma_i) \tag{7.43}$$

θεωρούμε κρυμμένη πληροφορία Z η οποία φανερώνει την κατανόμή από την οποία προέκυψε το γονίδιο x, σκοπός μας είναι να βρούμε τις παραμέτρους Θ για τις οποίες μεγιστοποιείται η πιθανοφάνεια

$$\Theta^* = \arg\max_{\Theta}\left(LL(\theta)\right) \tag{7.44}$$

Όπου αρχικά η πιθανοφάνεια ορίζεται ως:

$$LL(\theta|x) = \sum_n \log p(x_n|\theta) = \sum_n \log \sum_i \pi_i N(x_n|\mu_i, \Sigma_i) \tag{7.45}$$

ενώ με την εισαγωγή της κρυμμένης πληροφορίας η πιθανοφάνεια του πλήρους συνόλου ορίζεται ως:

$$LL_c(\theta|x) = \sum_{n=1}^{N}\sum_{j=1}^{M} z_j^n \log\left(\pi_j N\left(x_n|\mu_j, \Sigma_j\right)\right) \tag{7.46}$$

Η εφαρμογή του EM αλγορίθμου έχει ως εξής:

Στο E-step υπολογίζεται η αναμενόμενη τιμή της πλήρους πιθανοφάνειας

$$E\left[L_c\left(\theta^{(t)}\right)\right] = \sum_{n=1}^{N}\sum_{j=1}^{M} E\left(z_j^n\right)\log\left(\pi_j N\left(x^n|\theta_j\right)\right) \tag{7.47}$$

Έστω τ^i η υπό συνθήκη πιθανότητα το i στοιχείο του Z να είναι 1, από τον κανόνα του Bayes έχουμε:

$$\tau^i \triangleq p(Z^i = 1 \mid x, \theta) = \frac{p(x \mid Z^i = 1, \theta_i)\, p(Z^i = 1 \mid \pi_i)}{p(x \mid \theta)} = \frac{\pi_i N(x \mid \mu_i, \Sigma_i)}{\sum_j \pi_j N(x \mid \mu_j, \Sigma_j)} \tag{7.48}$$

M-step: Μεγιστοποίηση του $E\left[L_c\left(\theta^{(t)}\right)\right]$ ως προς π_j και $\theta_j = (\mu_j, \Sigma_j)$

Όπου προκύπτουν:

$$\pi_i^{t+1} = \frac{1}{N}\sum_{n=1}^{N} \tau_n^i(t),\ \mu_i^{t+1} = \frac{\sum_{n=1}^{N} \tau_n^i(t)\, x_n}{\sum_{n=1}^{N} \tau_n^i(t)},\ \Sigma_i^{t+1} = \frac{\sum_{n=1}^{N} \tau_n^i \left(x_n - \mu_i^{t+1}\right)\left(x_n - \mu_i^{t+1}\right)^T}{\sum_{n=1}^{N} \tau_n^i(t)} \tag{7.49}$$

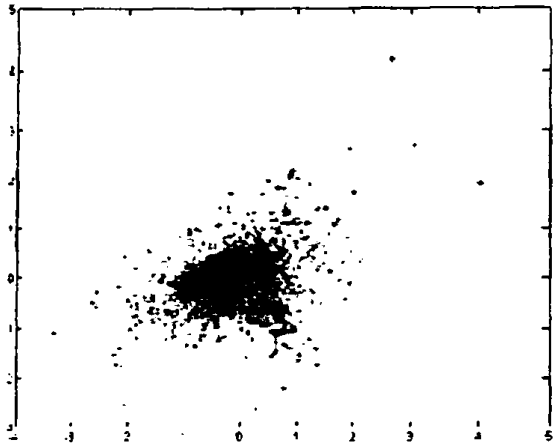
Πιο συγκεκριμένα για την κατηγοριοποίηση με χρήση GMM χρησιμοποιήθηκε ένα μεικτό Gaussian μοντέλο αποτελούμενο από 16 κανονικές κατανομές, ο πίνακας συμμεταβλητότητας (covariance) έχει την μορφή:

$$\Sigma = \begin{pmatrix} \sigma^2_1 & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{12} & \sigma^2_2 & \cdots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1m} & \sigma_{2m} & \cdots & \sigma^2_m \end{pmatrix} \tag{7.50}$$

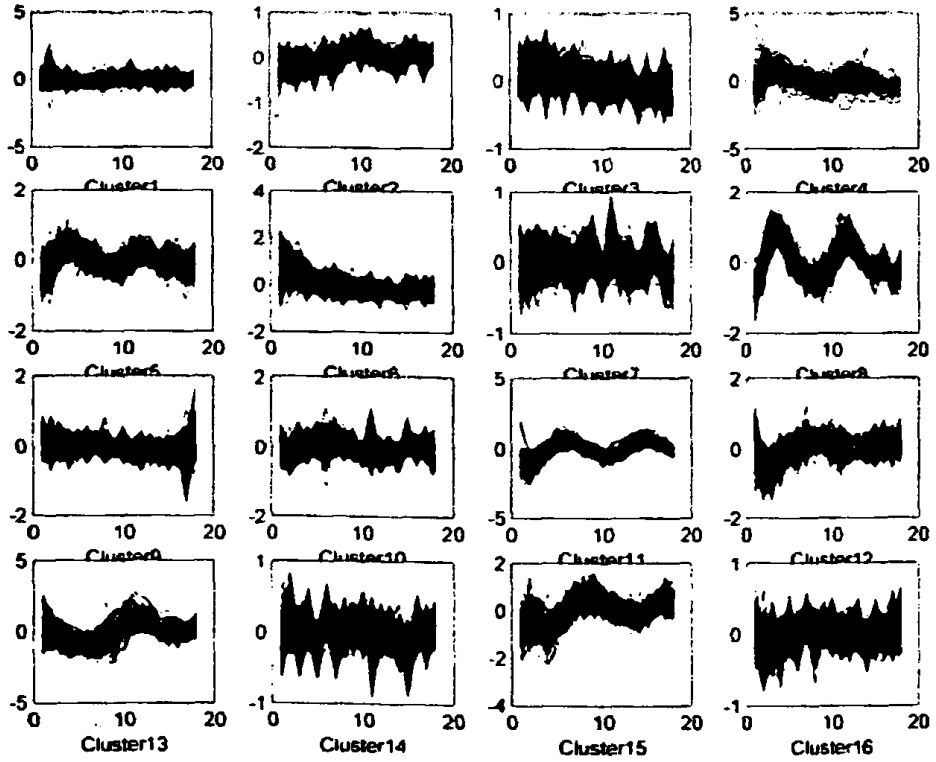
Σκοπός είναι να δημιουργηθούν ομάδες με όσο το δυνατό πιο ευέλικτο σχήμα, ένας πίνακας συμεταβλητότητας της παραπάνω μορφής δημιουργεί ομάδες με σχήμα έλλειψης σχηματίζοντας γωνία με τους κατακόρυφους άξονες.

7.8.6.2 Αποτελέσματα εφαρμογής Gaussian Mixture Models στα δεδομένα έκφρασης γονιδίων

Για την προβολή των αποτελεσμάτων έγινε προβολή του πολυδιάστατου χώρου σε 2-διαστατο με χρήση της PCA όπου χρησιμοποιήθηκαν τα 2 πιο ισχυρά components.



Εικόνα 7-22 Ομαδοποίηση γονιδίων χρησιμοποιώντας γονιδιακες εκφράσεις παραγόμενες από το alpha factor, η προβολή έγινε με χρήση της PCA όπου ο χώρος των δεδομένων ελαττώθηκε από τον $\Re^{18} \rightarrow \Re^{2}$



Εικόνα 7-23 Στο παραπάνω γράφημα φαίνονται οι γονιδιακές εκφράσεις των γονιδίων ανά cluster όπως αυτά έχουν διαμορφωθεί μετά την εκτέλεση του αλγορίθμου

7.9 Μοντελοποίηση με χρήση Bayesian δικτύων

Όλα τα κύτταρα σε έναν οργανισμό φέρνουν τα ίδια genomic δεδομένα, όμως ο πρωτεϊνικός τους χαρακτήρας μπορεί να είναι διαφορετικός χρονικά και χωρικά, λόγω της διαδικασίας ρύθμισης (regulation). Η πρωτεϊνική σύνθεση ρυθμίζεται από πολλούς μηχανισμούς στα διαφορετικά στάδιά της. Αυτοί περιλαμβάνουν τους μηχανισμούς για την έναρξη μεταγραφής, RNA σύνθεση, mRNA μεταφορά, έναρξη μεταφράσεων, μεταφραστικές τροποποιήσεις, και την υποβάθμιση mRNA/πρωτεΐνης. Μια από τις κύριες συνδέσεις στις οποίες οι διαδικασίες ρύθμισης εμφανίζονται είναι στην μεταγραφή του mRNA. Ένας σημαντικός ρόλος σε αυτούς τους μηχανισμούς διαδραματίζεται από τις ίδιες τις πρωτεΐνες, που δεσμεύονται στις ρυθμιστικές περιοχές κατά μήκος του DNA, έχοντας επιπτώσεις στη μεταγραφή των γονιδίων που ρυθμίζουν. Τα DNA microarray όπως έχει ήδη αναφερθεί είναι μια τεχνική για να μετρηθεί η έκφραση χιλιάδων mRNA ταυτόχρονα. Τέτοια πειράματα συλλέγουν τεράστια ποσά δεδομένων, τα οποία απεικονίζουν σαφώς πολλές πτυχές των βιολογικών διαδικασιών. Μια σημαντική πρόκληση είναι να αναπτυχθούν οι στατιστικές και υπολογιστικές μεθοδολογίες για την ανάλυση τέτοιου συνόλου δεδομένων και να συμπεράνουν βιολογικές αλληλεπιδράσεις από αυτά.

Τα περισσότερα από τα εργαλεία ανάλυσης που χρησιμοποιούνται αυτήν την περίοδο είναι βασισμένα μόνο σε αλγορίθμους ομαδοποίησης. Αυτοί οι αλγόριθμοι προσπαθούν να εντοπίσουν ομάδες γονιδίων που έχουν παρόμοια πρότυπα έκφρασης σε ένα σύνολο πειραμάτων. Μια τέτοια ανάλυση έχει αποδειχθεί χρήσιμη στην ανακάλυψη των γονιδίων που είναι ομο-ρυθμισμένα. Ένας πιο φιλόδοξος στόχος είναι ο συμπερασμός της δομής της μεταγραφικής διαδικασίας.

Επιπλέον, τα δεδομένα έκφρασης του mRNA δίνουν μόνο μια μερική εικόνα που δεν απεικονίζει τα βασικά γεγονότα όπως τη μετάφραση και τη πρωτεϊνική ενεργοποίηση. Τέλος, η ποσότητα δειγμάτων, ακόμη και στα μεγαλύτερα πειράματα, δεν παρέχει αρκετές πληροφορίες για να κατασκευαστεί ένα πλήρες πρότυπο με σημαντική στατιστική σημασία. Τα Μπεϋζιανά δίκτυα είναι μια προσέγγιση για την ανάλυση των προτύπων έκφρασης των γονιδίων, η οποία αποκαλύπτει τις ιδιότητες της μεταγραφικής διαδικασίας με την εξέταση των στατιστικών ιδιοτήτων της

εξάρτησης και της υπό όρους ανεξαρτησίας στα δεδομένα. Αυτά τα δίκτυα αναπαριστούν τη δομή εξάρτησης μεταξύ των αλληλεπιδρώντων ποσοτήτων (επίπεδα έκφρασης διαφορετικών γονιδίων). Τα Μπεϋζιανά δίκτυα είναι μαθηματικά μοντέλα και μπορούν εκτός των άλλων να χρησιμοποιηθούν για να συμπεράνουν αιτιότητα.

7.9.1 Αναπαριστώντας κατανομές με Μπεϋζιανά δίκτυα

Έστω ένα πεπερασμένο σύνολο τυχαίων μεταβλητών $X=\{X_1,...,X_n\}$ όπου κάθε μεταβλητή X_i μπορεί να πάρει τιμές x_i. Τα σύνολα μεταβλητών απεικονίζονται με έντονα κεφαλαία γράμματα **Z Y, X,** .Επίσης γράφουμε $I(X;Y \mid Z)$ για να δηλώσουμε ότι το Χ είναι εξαρτώμενο του Υ δοθέντος του Z

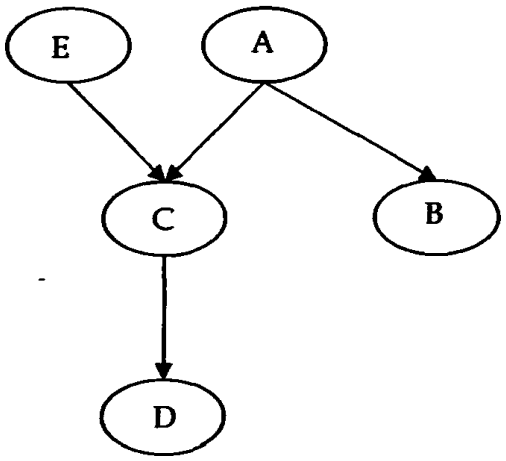
Ένα Μπεϋζιανό δίκτυο είναι μια αναπαράσταση μιας από κοινού κατανομής πιθανοτήτων. Αυτή η αναπαράσταση αποτελείται από δύο συστατικά: ένα κατευθυνόμενο άκυκλο γράφημα (DAG) G , του οποίου οι κόμβοι αντιστοιχούν στις τυχαίες μεταβλητές $X_1,...,X_n$, και το θ περιγράφει μια υπό συνθήκη κατανομή για κάθε μεταβλητή δοθέντων των γονιών του στο G., αυτά τα δύο συστατικά μαζί ορίζουν μια μοναδική κατανομή στα $X_1,...,X_n$.

Ο γράφος G αναπαριστά υποθέσεις για υπό συνθήκη ανεξαρτησίες οι οποίες επιτρέπουν στις από κοινού κατανομές να οριστούν κάνοντας χρήση μόνο των απαραίτητων παραμέτρων. Ο γράφος G εμπεριέχει την Μαρκοβιανή θεώρηση: **"Κάθε μεταβλητή X_i είναι ανεξάρτητη από τους απογόνους δοθέντων των γονέων του με βάση το *G*."**

Εφαρμόζοντας τον κανόνα της αλυσίδας των πιθανοτήτων και των ιδιοτήτων των υπό συνθήκη ανεξαρτησιών, οποιαδήποτε από κοινού κατανομή ικανοποιεί τον παραπάνω κανόνα μπορεί να αποσυντεθεί στη μορφή γινομένου

$$P(X_1,...,X_n)=\prod_{i=1}^{n} P\left(X_i \mid Pa^G(X_i)\right) \qquad (7.51)$$

Όπου *Pa* είναι το σύνολο των γονέων του *X* δεδομένου του G. Στην Εικόνα 7-24 παρουσιάζεται ένα παράδειγμα ενός γράφου G.



Εικόνα 7-24 Ένα παράδειγμα μίας απλής Μπεϋζιανής δομής δικτύων. Σε αυτή τη δομή φαίνονται διάφορες υπό όρους ανεξαρτησίες. Η δομή επίσης υποδεικνύει ότι η από κοινού κατανομή έχει τη μορφή του γινομένου $P(A,B,C,D,E)=P(A)P(E)P(B|A,E)P(C|B)P(D|A)$

Πρέπει επίσης να ορίσουμε κάθε μιας από τις υπό όρους πιθανότητες στη μορφή γινομένων. Το δεύτερο στοιχείο ενός Μπευζιανού δικτύου περιγράφει αυτές τις υπό συνθήκη κατανομές από τις παραμέτρους θ. Αν υποθέσουμε ότι οι γονείς μιας μεταβλητής X είναι $U_1,...,U_n$. Τότε

- για διακριτές τιμές, δηλαδή εάν κάθε X και $U_1,...,U_n$ παίρνουν διακριτές τιμές από ένα πεπερασμένο σύνολο τότε μπορούμε να αναπαραστήσουμε την πιθανότητα $P(X|U_1,...,U_k)$ ως ένα πίνακα ο οποίος ορίζει την πιθανότητα των τιμών του X για κάθε από συνδυασμό των $U_1,...,U_k$

- Για συνεχείς μεταβλητές, μια κατανομή που μπορεί να χρησιμοποιηθεί είναι η Gaussian κατανομή, όπου:

$$P(X|u_1,...,u_k)\sim N\left(a_0+\sum_i a_i u_i,\sigma^2\right) \qquad (7.52)$$

Δηλαδή το Χ κατανέμεται κανονικά γύρω από έναν μέσο που εξαρτάται γραμμικά από τις τιμές των γονέων του.

7.9.2 Εκπαίδευση Μπεϋζιανών δικτύων

Το πρόβλημα της εκπαίδευσης Μπεϋζιανών δικτύων μπορεί να οριστεί ως εξής. Δοθέντος ενός συνόλου εκπαίδευσης $D=\{x^1,...,x^n\}$ ανεξάρτητων στιγμιότυπων του X, βρες ένα δίκτυο $B=\langle G,\Theta\rangle$ το οποίο ταιριάζει καλύτερα στο D. Η θεωρία της εκπαίδευσης δικτύων από δεδομένα έχει ήδη εξεταστεί εκτενώς τις τελευταίες δεκαετίες. Η κοινή προσέγγιση είναι να εισαχθεί μια στατιστική συνάρτηση score η οποία θα αξιολογεί κάθε δίκτυο με βάση τα δεδομένα, και θα αναζητά το βέλτιστο δίκτυο σύμφωνα με αυτό το αποτέλεσμα. Μια συνάρτηση κόστους μπορεί να είναι:

$$P(D|G)=\int P(D|G,\Theta)P(\Theta|G)d\Theta \tag{7.53}$$

$$\begin{aligned} S(G:D) &= \log P(G|D) \\ &= \log P(D|G)+\log P(G)+C \end{aligned} \tag{7.54}$$

όπου το C είναι σταθερά ανεξάρτητη του G και

Η επιλογή των priors για κάθε G ορίζουν το ακριβές Μπεϋζιανό score. Αν αφήσουμε την ακριβή περιγραφή των priors, το μόνο που μας ενδιαφέρει είναι να αναλύσουμε περαιτέρω τη συνάρτηση score, το οποίο μπορεί να γραφεί ως εξής:

$$S(G:D)=\sum_i ScoreContribution\left(X_i,Pa^G(X_i):D\right) \tag{7.55}$$

όπου η συμβολή κάθε μεταβλητής X_i στο συνολικό δίκτυο εξαρτάται μόνο από την δική της τιμή και τις τιμές των γονέων του στο G. Αυτές οι τοπικές συνεισφορές για κάθε μεταβλητή μπορούν να υπολογιστούν χρησιμοποιώντας μια κλειστού τύπου εξίσωση. Εφόσον οι priors έχουν οριστεί και τα δεδομένα δίνονται, η εκπαίδευση σκοπεύει στην εύρεση της δομής του G που μεγιστοποιεί το score. Το πρόβλημα αυτό είναι γνωστό ότι είναι NP-hard, κατά συνέπεια προσφεύγουμε σε ευρετική αναζήτηση. Έχει υπολογιστεί ότι ο αριθμός των DAG για ένα γράφο n κόμβων δίνεται από την σχέση

$$f(1)=1$$

$$f(n)=\sum_{i=1}^{n}(-1)^{i+1}\frac{n!}{(n-1)!i!}2^{i(n-i)}f(n-i)$$

δηλαδή για n=10 υπάρχουν $f(10)$~4.2*10^{18} διαφορετικά άκυκλα γραφήματα.

Η ανάλυση του score είναι σημαντική για αυτήν την βελτιστοποίηση. Μια τοπική διαδικασία αναζήτησης που αλλάζει μια ακμή σε κάθε βήμα μπορεί να αξιολογεί αποτελεσματικά το κέρδος που υπάρχει. Ένα παράδειγμα μιας τέτοιας διαδικασίας είναι ο greedy αλγόριθμος hill-climbing που σε κάθε βήμα εκτελεί μια τοπική αλλαγή που οδηγεί σε μέγιστο κέρδος, και εκτελείται έως ότου φθάνει σε ένα τοπικό μέγιστο. Ένα Μπεϋζιανό δίκτυο είναι ένα μοντέλο εξαρτήσεων μεταξύ πολλών μεταβλητών. Στα Μπεϋζιανά δίκτυα, οι γονείς μιας μεταβλητής ερμηνεύονται ως οι άμεσες αιτίες του. Ένα αιτιώδες δίκτυο μοντελοποιεί όχι μόνο τη κατανομή των παρατηρήσεων, αλλά και τα αποτελέσματα των αλληλεπιδράσεων. Εάν το X προκαλεί το Y, τότε αλλάζοντας τις τιμές του X προκαλεί αλλαγές στις τιμές του Y. Απ' την άλλη, εάν το Y είναι αιτία του X, επηρεάζοντας το X δεν έχει επιπτώσεις στο Y. Στο πεδίο της βιολογίας υποθέτουμε ότι το X είναι ένας παράγοντας μεταγραφής του Y. Εάν μεταβάλλουμε την ισορροπία του γονιδίου X θα έχουμε επιπτώσεις στην έκφραση του γονιδίου Y. Αλλά μεταβάλλοντας την ισορροπία του γονιδίου Y δεν θα έχει καμία επίδραση στην έκφραση του γονιδίου X. Υπάρχει επίσης μια αντίστοιχη μαρκοβιανή υπόθεση στο αιτιώδες δίκτυο: δοθέντων των τιμών των άμεσων αιτιών κάποιας μεταβλητής, είναι ανεξάρτητες από τις προηγούμενες αιτίες του.

7.9.3 Αναλύοντας δεδομένα έκφρασης

Σε αυτή τη φάση σκοπός μας είναι να περιγράψουμε την κατάσταση του συστήματος (ενός κυττάρου ή ένας οργανισμού και το περιβάλλον του) χρησιμοποιώντας τις τυχαίες μεταβλητές. Αυτές οι τυχαίες μεταβλητές δείχνουν το επίπεδο έκφρασης των γονιδίων. Επιπλέον, περιλαμβάνουμε και ορισμένες τυχαίες μεταβλητές που αναφέρονται σε άλλες ιδιότητες που έχουν επιπτώσεις στο σύστημα, όπως πειραματικοί όροι, χρονικοί δείκτες, background μεταβλητές, και εξωκυτταρικές συνθήκες. Προσπαθούμε έτσι να χτίσουμε ένα μοντέλο που είναι η από κοινού κατανομή ενός συνόλου τυχαίων μεταβλητών. Εάν έχουμε ένα τέτοιο μοντέλο, θα μπορούμε να απαντήσουμε σε ένα ευρύ φάσμα ερωτήσεων για το σύστημα. Οι

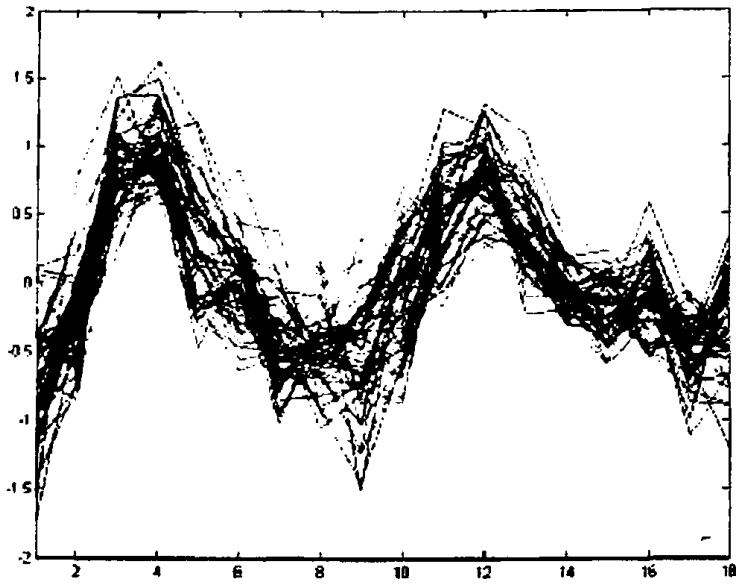
περισσότερες από τις δυσκολίες στην εκπαίδευση από δεδομένα έκφρασης οφείλεται στο γεγονός ότι τα δεδομένα έκφρασης περιλαμβάνουν επίπεδα μεταγραφής χιλιάδων γονιδίων ενώ σε πολλά άλλα προβλήματα το σύνολο δεδομένων περιλαμβάνει μερικές δεκάδες δειγμάτων. Αυτό προσθέτει προβλήματα στην υπολογιστική πολυπλοκότητα και τη στατιστική σημασία των προκυπτόντων δικτύων. Απ' την άλλη, ένα θετικό στοιχείο είναι ότι, τα γενετικά ρυθμιστικά δίκτυα είναι αραιά, δηλ., λαμβάνοντας υπόψη ένα γονίδιο, συνήθως λιγότερα από δώδεκα γονίδια έχουν επιπτώσεις άμεσα στη μεταγραφή του.

Η προσέγγιση που ακολουθήσαμε για την εκπαίδευση του δικτύου είναι βασισμένη στην μέθοδος μέγιστης πιθανοφάνειας ML, μπορεί να μην είναι βέλτιστη εξαιτίας της φύσης του προβλήματος (ύπαρξη κρυμμένων μεταβλητών και πολλές μεταβλητές), αλλά μπορεί να εκτιμήσει μια καλή λύση για το πρόβλημα αυτό σε πεπερασμένο χρόνο.

7.9.4 Εφαρμογή του αλγορίθμου εκπαίδευσης στα δεδομένα έκφρασης γονιδίων

Σε αυτή τη παράγραφο θα περιγραφεί αναλυτικά ο αλγόριθμος εκπαίδευσης Bayesian δικτύων έχοντας πλήρες σύνολο δεδομένων. Ο αλγόριθμος που εφαρμόσαμε για την εκπαίδευση του δικτύου είναι ένας απλός αλγόριθμος εύρεσης βέλτιστης δομής ο οποίος δουλεύει ως εξής. Ο γράφος αρχικοποιείται με τον περιορισμό: ο κάθε κόμβος να έχει γονείς max_fun_in γονείς. Στην πορεία δημιουργούνται όλα τα γραφήματα τα οποία προκύπτουν από το αρχικό γράφημα με εφαρμογή ενός από τους τελεστές {add, del, reverse} μια φορά. Για κάθε ένα γράφημα υπολογίζεται το BIC score, η δομή με το μεγαλύτερο score χρησιμοποιείται ως η αρχική δομή για την επόμενη επανάληψη του αλγορίθμου.

Για το πρόβλημα το οποίο εξετάζουμε το σύνολο δεδομένων είναι η 11η ομάδα που έχει παραχθεί από την εφαρμογή Gaussian Mixture Models, και αυτό διότι ήταν η ομάδα με την καλύτερη ακρίβεια.



Εικόνα 7-25 Τα δεδομένα που της 11ης ομάδας που χρησιμοποιήθηκαν για την εκπαίδευση του Μπεϋζιανού μοντέλου

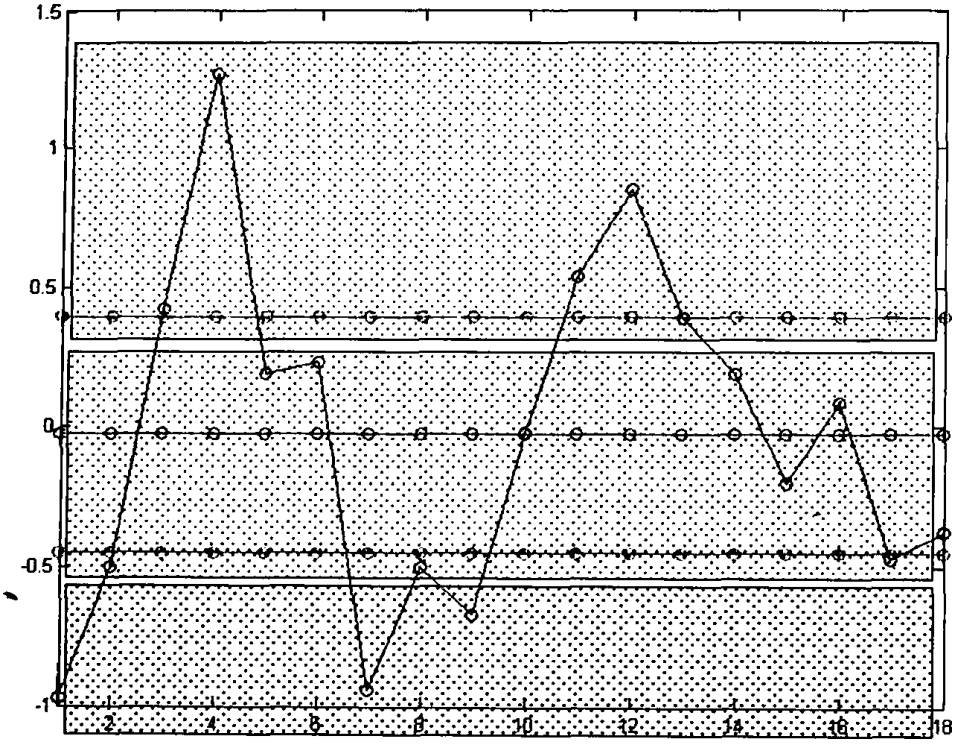
Επίσης έχουμε θεωρήσει ότι η κάθε μεταβλητή (κάθε γονίδιο) μπορεί να πάρει τρεις τιμές, άρα είναι μια διακριτή μεταβλητή που παίρνει τιμές από το σύνολο {1,2,3} τα οποία αναπαριστούν τις καταστάσεις {«υπό εκφράζεται», «δεν εκφράζεται», «υπέρ εκφράζεται»} αντίστοιχα. Για να πάρουμε τις τιμές αυτές εφαρμόσαμε τον παρακάτω τύπο στα δεδομένα έκφρασης:

$$x_{ij} = \begin{cases} 1, \text{εαν} & \min(x_i) \le x_i^j < \bar{x}_i - \sigma(x_i)/3 \\ 2, \text{εαν} & \bar{x}_i - \sigma(x_i)/2 \le x_i^j \le \bar{x}_i + \sigma(x_i)/3, \\ 3, \text{εαν} & \bar{x}_i + \sigma(x_i)/3 < x_i^j \le \max(x_i) \end{cases}$$

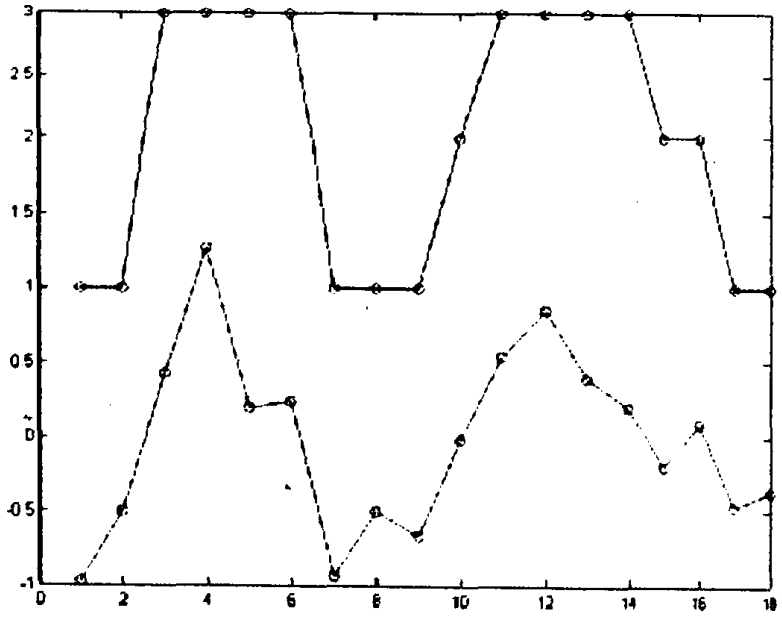
οπου (7.56)

$$\sigma = \left(1/(n-1) \sum_{j=1}^{n} \left(x_i^j - \bar{x}_i \right)^2 \right)^{1/2}$$

Όπου η κατάσταση του γονιδίου σε μια χρονική στιγμή λέμε ότι χαρακτηρίζεται ως «υπό εκφράζεται» αν η τιμή στην συγκεκριμένη χρονική στιγμή είναι μικρότερη από το μέσο όρο μείον το 1/3 της τυπικής απόκλισης.



Εικόνα 7-26 διακριτοποίηση της κατάστασης ενός γονιδίου, με κόκκινο φαίνεται η περιοχή όπου υπέρ εκφράζεται, με πράσινο η περιοχή που δεν εκφράζεται και με μπλε η περιοχή που υπό εκφράζεται.

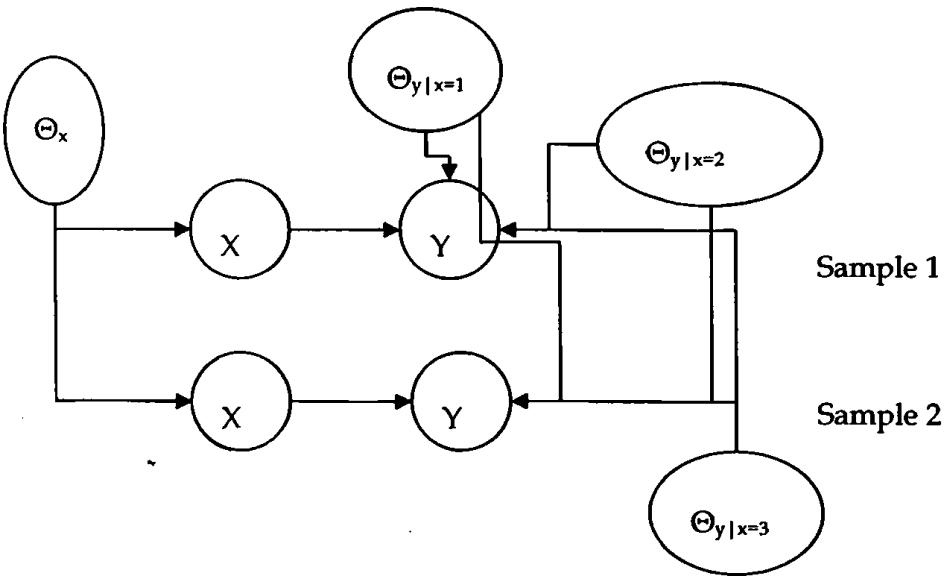


Εικόνα 7-27 Με την κόκκινη γραμμή φαίνονται οι τιμές μέτρησης της έκφρασης του γονιδίου YCL060C και με μπλε κουκίδες παρουσιάζονται οι τιμές που θα χρησιμοποιηθούν μετά την διακριτοποίηση, όπως αυτή έγινε εφαρμόζοντας τον αλγόριθμο που παρουσιάστηκε πιο πάνω.

Επίσης στο μοντέλο το οποίο προτείνουμε έχουμε εισάγει 3 μεταβλητές οι οποίες ποσοτικοποιούν τρεις καταστάσεις που επηρεάζουν το σύστημά, οι μεταβλητές αυτές εκφράζουν επιδράσεις από, θερμοκρασία, εσωκυτταρικοί παράγοντες, εξωκυτταρικοί παράγοντες και οι τιμές που παίρνουν είναι επίσης διακριτές από το σύνολο {1,2,3} το οποίο μπορεί να μεταφραστεί ως {«καταστέλλουν», «δεν επηρεάζουν», «επηρεάζουν»} το σύστημα. Τέλος η εκπαίδευση εφαρμόζεται στα πλήρη δεδομένα αφού δηλαδή έχει εφαρμοστεί ο αλγόριθμος KNN impute ο οποίος συμπληρώνει τις τιμές που λείπουν.

7.9.4.1 Εκπαιδεύοντας παραμέτρους έχοντας πλήρες σύνολο δεδομένων

Στο σχήμα έχουμε ένα απλό μοντέλο με δύο μεταβλητές, και οι δύο μεταβλητές μπορούν να πάρουν 3 τιμές από το σύνολο {1,2,3}. Στο μοντέλο αυτό θα βασιστούμε για να ορίσουμε τον τρόπο με τον οποίο εκπαιδεύονται οι παράμετροι καθώς και οι πιθανότητες των μεταβλητών του μοντέλου.



Στην εργασία που ακολουθεί θεωρούμε ότι το Bayesian Network ορίζεται από multinomial κατανομές. Αυτό σημαίνει ότι κάθε τυχαία μεταβλητή στο μοντέλο μας είναι διακριτή και οι τοπικές κατανομές είναι πολυωνυμικές , (το σύνολο των

καταστάσεων της κάθε μεταβλητής είναι 3 από το σύνολο {1,2,3}). Κατά συνέπεια αν pa_i είναι το σύνολο μεταβλητών που είναι γονείς του κόμβου x_i τότε η τοπική κατανομή δίνεται από την σχέση

$$p\left(x_i^k \mid pa_i^j, \theta_i, S\right) = \theta_{ijk} > 0 \tag{7.57}$$

$$\text{όπου } pa_i^1, \ldots, pa_i^{q_i} \quad , \left(q_i = \prod_{X_i \in Pa_i} r_i \right) \text{ και } r_i = 3 \tag{7.58}$$

$$\text{επίσης } \theta_{ij} = \left(\theta_{ij2}, \theta_{ij3}\right) \text{ και } \theta_{ij1} = 1 - \sum_{k=2}^{3} \theta_{ijk} \tag{7.59}$$

Στην πιο απλή περίπτωση θεωρούμε ότι έχουμε πλήρες σύνολο δεδομένων το οποίο σημαίνει ότι δεν λείπουν τιμές από το σύνολο δεδομένων και κατά συνέπεια δεν υπάρχουν κρυφές μεταβλητές στο μοντέλο μας. Αυτή η υπόθεση επίσης, μας επιτρέπει να θεωρήσουμε ότι τα διανύσματα των παραμέτρων θ_{ij} είναι ανεξάρτητα μεταξύ τους.

Δηλαδή:

$$p(\theta_s \mid S) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} p\left(\theta_{ij} \mid S\right) \tag{7.60}$$

όπου n είναι ο αριθμός των μεταβλητών

και q_i ο αριθμός των γονέων του x_i

θ_i είναι το διάνυσμα των παραμέτρων για την κατανομή $p(x_i \mid pa_i, \theta_i, S)$ και θ_S είναι το διάνυσμα των παραμέτρων $(\theta_1, \ldots, \theta_n)$

Αν θεωρήσουμε ότι το $p(\theta_{ij})$ ακολουθεί κατανομή $Dir(\theta_{ij} \mid a_{ij1}, a_{ij2}, a_{ij3})$ τότε και η posterior ακολουθεί επίσης κατανομή *Dirichlet* η οποία είναι: $Dir(\theta_{ij} \mid a_{ij1} + N_{ij1}, a_{ij2} + N_{ij2}, a_{ij3} + N_{ij2})$, όπου N_{ij1} είναι ο αριθμός των περιπτώσεων στο D όπου $X_i = x_i^k$ και $Pa_i = pa_i^j$, τα x_i^k, pa_i^j παίρνουν τιμές από το σύνολο {1,2,3}

Ενώ η κατανομή *Dirichlet* στην γενική της μορφή δίνεται από την σχέση $p(\theta)=\frac{\Gamma(\alpha)}{\prod_{\kappa=1}^{r}\Gamma(\alpha_\kappa)}\prod_{\kappa=1}^{r}\theta_\kappa^{\alpha_\kappa-1}, \quad \alpha=\sum_{\kappa=1}^{r}\alpha_\kappa$, η συνάρτηση Γ ικανοποιεί τις συνθήκες

$\Gamma(x+1)=x\Gamma(x)$ και $\Gamma(1)=1$ ή αλλιώς $\Gamma(x)=\int_0^\infty t^{x-1}e^{-t}dt$

Ένας εκτιμητής της θ είναι η Maximum Likelihood (ML) εκτίμηση η οποία επιλέγει τις τιμές του θ που μεγιστοποιούν την πιθανοφάνεια $p(D|\theta)$ και δίνεται από την σχέση $\theta_{ijk}=\frac{N_{ijk}}{\sum_{k=1}^{3}N_{ij}}$, όπου γίνεται μηδέν αν το N_{ijk} είναι μηδέν, (κάτι που εξαρτάται από το σύνολο δεδομένων). Επειδή, για τον λόγο αυτό, υπάρχει αβεβαιότητα για το θ μια άλλη εκτίμηση που χρησιμοποιούμε είναι η αναμενόμενη τιμή του θ σε σχέση με την posterior η οποία δίνεται από τον τύπο $E_{p(\theta|D)}(\theta)=\int\theta p(\theta|D)d\theta$ από όπου παίρνουμε τελικά ότι

$$\theta_{ijk}=\frac{a_{ijk}+N_{ijk}}{\sum_{k=1}^{3}a_{ijk}+N_{ijk}} \tag{7.61}$$

Τα a_{ijk} είναι οι παράμετροι της Dirichlet κατανομής, τα οποία επιλέγονται συνήθως $a_{ij1}=a_{ij2}=a_{ij3}>>1$,

7.9.4.2 Εκπαιδεύοντας παραμέτρους έχοντας μη πλήρες σύνολο δεδομένων

Στη συνέχεια θα δούμε πως μπορούμε να ενημερώσουμε τις παραμέτρους ενός μοντέλου στην περίπτωση που το σύνολο δεδομένων είναι μη πλήρες δηλ. σε μερικά στιγμιότυπα του συνόλου δεδομένων δεν υπάρχουν τιμές. Στην περίπτωση αυτή το σύνολο των μεταβλητών μας αποτελείται και από κρυφές μεταβλητές των οποίων οι καταστάσεις είναι ίδιες με αυτές των γνωστών μεταβλητών, δηλ. ακολουθούν πολυωνυμική κατανομή 3 καταστάσεων.

Θεωρούμε ότι έχουμε το σύνολο των παρατηρούμενων μεταβλητών $Y \subset X$ και $Z \subset X$ το σύνολο των μη πλήρως παρατηρούμενων μεταβλητών. Η posterior κατανομή για τα θ_{ij} δοθέντος του y και της δομής S δίνεται από τους τύπους:

$$\begin{aligned} p(\theta_{ij} \mid y,S) &= \sum_z p(z \mid y,S)\, p(\theta_{ij} \mid y,z,S) \\ &= \left(1 - p\left(pa_i^j \mid y,S\right)\right)\left\{p(\theta_{ij} \mid S)\right\} + \sum_{k=1}^{3} p\left(x_i^k, pa_i^j \mid y,S\right)\left\{p\left(\theta_{ij} \mid x_i^k, pa_i^j, S\right)\right\} \end{aligned} \tag{7.62}$$

Ένας τρόπος για να εκτιμήσουμε το θ είναι χρησιμοποιώντας την μέγιστη πιθανοφάνεια, δηλαδή θέλουμε το θ που μεγιστοποιεί την πιθανοφάνεια $p(D \mid \theta_s, S)$

ή $$\theta_s = \arg\max_{\theta_s} \left\{p(D \mid \theta_s, S)\right\}, \tag{7.63}$$

όπου θ το διάνυσμα των παραμέτρων του μοντέλου. Ένας τρόπος για να υπολογιστεί η ML ή η MAP είναι με χρήση του αλγορίθμου ΕΜ.

Τα βήματα που ακολουθούμε για να υπολογιστεί η ML ή η MAP, είναι τα εξής:

1. αποδίδουμε τυχαία τιμές στο διάνυσμα θ
2. υπολογίζουμε την αναμενόμενη στατιστική για ένα πλήρες σύνολο δεδομένων, όπου η αναμενόμενη τιμή προκύπτει με βάση το την από κοινού κατανομή του X δοθέντος του γνωστού διανύσματος θ και των γνωστών δεδομένων D. Για την περίπτωσή μας υπολογίζουμε το

$$E_{p(x \mid D,\theta,S)}\left(N_{ijk}\right) = \sum_{l=1}^{N} p\left(x_i^k, pa_i^j \mid y_l, \theta, S\right) \tag{7.64}$$

όπου το y_l είναι η πιθανή μη πλήρες l περίπτωση του D. Ενώ τα X_i και όλες οι μεταβλητές στο Pa_i είναι γνωστές από το σύνολο δεδομένων στην περίπτωση x_l.

3. Στην περίπτωση που υπολογίζουμε την ML παίρνουμε τις τιμές του διανύσματος θ_s μεγιστοποιώντας την πιθανοφάνεια $p(D_c \mid \theta_s, S)$ όπου για την περίπτωσή μας δίνεται από την σχέση:

$$\theta_{ijk} = \frac{E_{p(x \mid D,\theta,S)}\left(N_{ijk}\right)}{\sum_{k=1}^{3} E_{p(x \mid D,\theta,S)}\left(N_{ijk}\right)} \tag{7.65}$$

ενώ αν χρησιμοποιήσουμε την *maximum a posterior* MAP υπολογίζουμε τις τιμές του διανύσματος θ μεγιστοποιώντας την πιθανότητα $p(\theta \mid D_c, S)$ η δοθέντος ότι η prior ακολουθεί την κατανομή Dirichlet η νέες τιμές του θ_{ijk} δίνονται από τον τύπο

$$\theta_{ijk} = \frac{a_{ijk} + E_{p(x|D,\theta,S)}(N_{ijk})}{\sum_{k=1}^{3}\left(a_{ijk} + E_{p(x|D,\theta,S)}(N_{ijk})\right)} \tag{7.66}$$

Το βήμα αυτό αποτελεί το *maximization* βήμα του EM αλγορίθμου, το οποίο αποδεικνύεται ότι συγκλίνει σε τοπικό μέγιστο. Στην περίπτωσή μας τα a_{ijk} έχουν επιλεγεί ίσα με τιμή μεγαλύτερη της μονάδας ($a_{ijk} = 2$)

Στην περίπτωση που έχουμε πλήρες σύνολο δεδομένων από τον κανόνα του Bayes το MAP μοντέλο είναι εκείνο που μεγιστοποιεί την ποσότητα:

$$P(G \mid D) = \frac{P(D \mid G)P(G)}{P(D)} \tag{7.67}$$

ή

$$\log P(G \mid D) = \log P(D \mid G) + \log P(G) + c \tag{7.68}$$

όπου $G = \langle S, \theta \rangle$, $c = -\log P(D)$ το οποίο είναι σταθερά ανεξάρτητο από το G. Το $\log P(G)$ χρησιμοποιείται σαν ποινή για τα σύνθετα μοντέλα.

Στην περίπτωση που έχουμε κρυφές μεταβλητές και μη πλήρες σύνολο δεδομένων τότε η πιθανοφάνεια προσεγγιστικά δίνεται από την παρακάτω σχέση

$$\log P(D \mid G) \approx \log P(D \mid S, \theta_{MAP}) - \frac{d}{2}\log N \tag{7.69}$$

όπου N είναι ο αριθμός των δειγμάτων, d είναι ο αριθμός των παραμέτρων και θ_{MAP} η MAP εκτίμηση των παραμέτρων. Το score αυτό χρησιμοποιήθηκε για τον υπολογισμό του $Score(S : S_k)$ για κάθε ένα από τα μοντέλα που ανήκουν στο σύνολο S_k όπως αυτά έχουν προκύψει εφαρμόζοντας τους τελεστές *add, del, reverse.*

Η ποσότητα $\log P(D \mid S, \theta_{MAP})$ υπολογίζεται ως εξής:

$$\begin{aligned} & \log P(D \mid S, \theta_{MAP}) \\ &= \log P(x_1, \ldots, x_n \mid S, \theta_{MAP}) \\ &= \log \prod_{i=1}^{n} P(x_i \mid Pa_i, S, \theta_{MAP}) \\ &= \sum_{i=1}^{n} \log P(x_i \mid Pa_i, S, \theta_{MAP}) \end{aligned} \tag{7.70}$$

q_i ο αριθμός των γονέων του κόμβου x_i

<u>Αλγόριθμος εύρεση βέλτιστης δομής</u>

είσοδος: Πλήρες σύνολο δεδομένων [nxm], n: μεταβλητές, m περιπτώσεις

1. Αρχικοποίηση της δομής (S_0)
2. Επανάληψη για i = 1,...max_loop
   a. Υπολογισμός των παραμέτρων του Bayesian δικτύου
   b. Δημιουργία όλων των δομών που διαφοροποιούνται με την εφαρμογή ενός από τους τελεστές {*add, del, rev*} μια φορά
   c. Υπολογισμός του Bayesian score για όλα τα γραφήματα B^h
   d. Εάν $\max(B^h) < 0{,}99 * BIC(S_{old})$
      i. Αρχικοποίηση της δομής (S_i) (συνέχεια επανάληψης)
   e. Αλλιώς $S_{best} = S_{new}$ - συνέχεια επανάληψης
3. Επιστροφή S_{best}

Για την υλοποίηση του αλγορίθμου χρησιμοποιήθηκε το BN Toolbox και πιο συγκεκριμένα οι συναρτήσεις:

Συνάρτηση	Περιγραφή

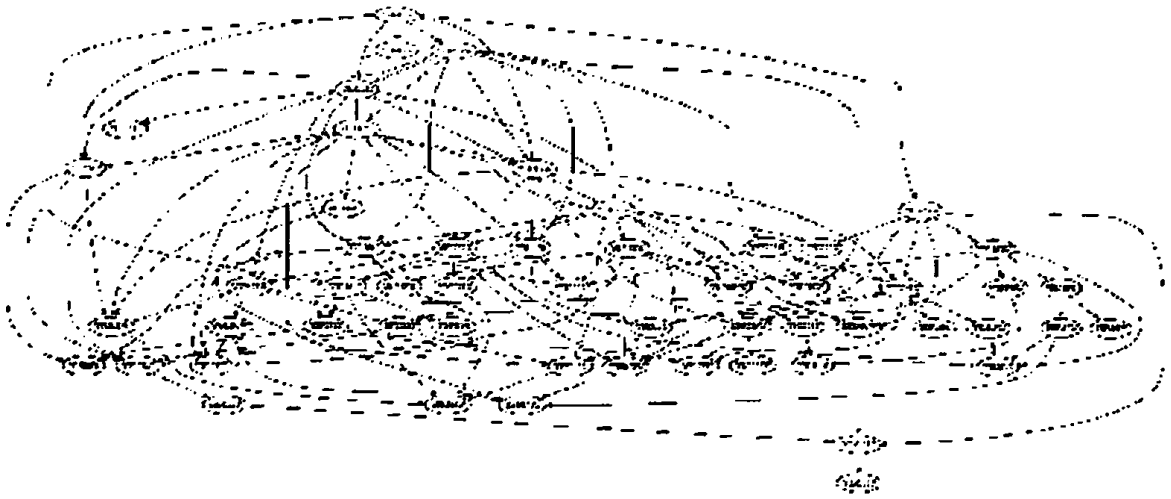
Score_dags	Υπολογίζει το BIC score για ένα η περισσότερα μοντέλα
learn_params_em	Υπολογίζει τις παραμέτρους ενός μοντελου χρησιμοποιώντας την MAP και τον αλγόριθμο ΕΜ. Συγκεκριμένα για το μοντέλο μας εάν δεν υπάρχουν κρυφές μεταβλητές υπολογίζει τις παραμέτρους από τον τύπο $\theta_{ijk} = \frac{a_{ijk} + N_{ijk}}{\sum_{k=1}^{3} a_{ijk} + N_{ijk}}$ Εάν υπάρχουν κρυφές μεταβλητές υπολογίζει τις παραμέτρους από τον τύπο $\theta_{ijk} = \frac{a_{ijk} + E_{p(x\|D,\theta,S)}(N_{ijk})}{\sum_{k=1}^{3}\left(a_{ijk} + E_{p(x\|D,\theta,S)}(N_{ijk})\right)}$
mk_nbrs_of_dag	Υπολογίζει όλες τις "γειτονικές" στο αρχικό μοντέλο δομές

Το μεγαλύτερο μέρος του χρόνου που καταναλώνεται κατά τη διάρκεια της εκτέλεσης της διαδικασίας ξοδεύεται στους υπολογισμούς των αναμενόμενων στατιστικών. Αυτή είναι η σημαντική διαφορά από τον παραμετρικό ΕΜ. Στον παραμετρικό ΕΜ, ξέρουμε εκ των προτέρων ποίες αναμενόμενες στατιστικές απαιτούνται. Δηλαδή, αυτές οι στατιστικές των γεγονότων εκφράζονται από τα X_i, pa_i όπου pa_i είναι οι "γονείς" του κόμβου x_i

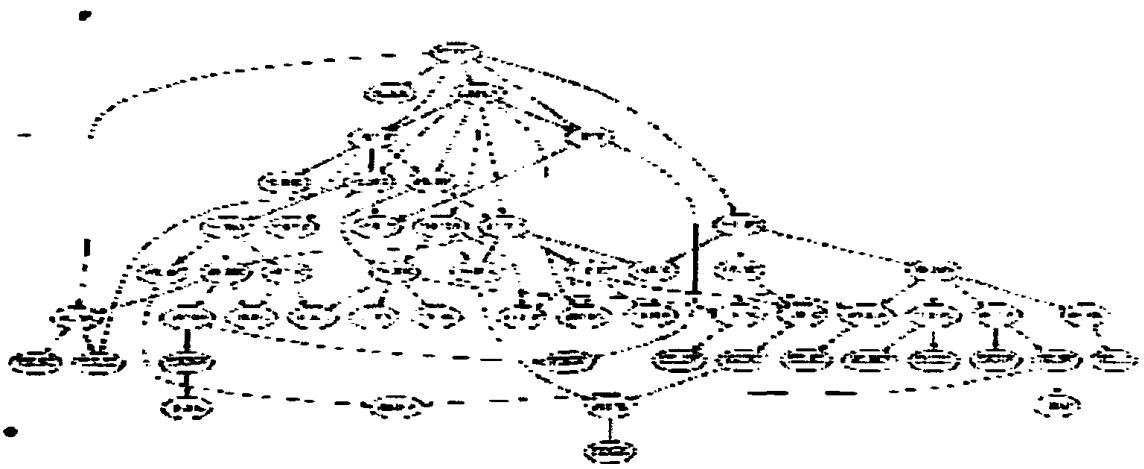
Ένα άλλο κρίσιμο σημείο είναι η επιλογή του αρχικού μοντέλου για τον αλγόριθμο. Αυτή η επιλογή καθορίζει το σημείο σύγκλισης του αλγορίθμου. Γενικά, δεν θέλουμε να επιλέξουμε μια πάρα πολύ απλή αρχική δομή, δεδομένου ότι μια τέτοια δομή ενσωματώνει πολλές ανεξαρτησίες, και προκαταλαμβάνει έτσι τις αναμενόμενες

στατιστικές για να αποφανθεί τελικά ότι οι μεταβλητές είναι ανεξάρτητες η μια από την άλλη. Αφ' ετέρου, δεν θέλουμε να επιλέξουμε μια πάρα πολύ σύνθετη αρχική δομή, δεδομένου ότι με μια τέτοια δομή θα είναι πολύ δύσκολο να εκτελεσθεί αλγόριθμοι συμπερασματολογίας. Όπως και στον παραμετρικό ΕΜ, επιλέγουμε τις αρχικές παραμέτρους για την επιλεγμένη δομή τυχαία. Επιπλέον, προκειμένου να αποφευχθεί η ανεπιτυχής πρόωρη σύγκλιση, τρέχουμε τη διαδικασία αρκετές φορές, αρχίζοντας από διαφορετικά αρχικά μοντέλα. Έπειτα επιλέγουμε το δίκτυο με το υψηλότερο score από εκείνα που βρέθηκαν στα τρεξίματα.

Ο αλγόριθμος εφαρμόστηκε σε μοντέλα τα οποία δεν ενσωματώνουν εκ των προτέρων γνώση και αποτελούνται από ένα δίκτυο αλυσίδα και σε γράφους που περιλαμβάνουν εκ των προτέρων γνώση δηλαδή γνωστές αλληλεπιδράσεις γονιδίων προερχόμενες από γνωστές βάσεις δεδομένων όπως η PIN (Protein Interaction Network). Στις παρακάτω εικόνες βλέπουμε το αρχικό μοντέλο και αυτό που προέκυψε μετά από 300 επαναλήψεις του αλγορίθμου.



Εικόνα 7-28 Ο αρχικός γράφος



Εικόνα 7-29 Το μοντέλο που προέκυψε μετά από 50 επαναλήψεις του αλγορίθμου

Τα αποτελέσματα του αλγορίθμου παρουσιάζονται λεπτομερώς στο επόμενο κεφάλαιο.

Ο κώδικας του αλγορίθμου:

```
function [BestBNet, order, BIC_score] = learn_struct_simple(data, ns, bnet, samplesM, max_loop)

tiny = exp(-700);

N = length(bnet.dag);

ncases = size(samplesM, 2);

ns = bnet.node_sizes;

prev_score = 1000;

loop = 0;

evidence = cell(1,N);

count_init_bnets = 0;

while loop<max_loop

  loop = loop + 1

  engine = jtree_inf_engine(bnet);

  bnet = learn_params_em(engine, samplesM, 30);

  if loop == max_loop-1

    figure
```

```
%plot(LOGLIKE, 'x-')

end

[nbrs, ops, nodes, orders] = mk_nbrs_of_dag_topo(bnet.dag);

nGs = length(nbrs);

cache = score_init_cache(N,1000);

[score, cache] = score_dags(data, ns, nbrs, 'scoring_fn', 'bic', 'params', [], 'discrete', [1:N], 'cache',cache);

[new_score, bestc]=max(score);

if (abs(abs(prev_score)-abs(new_score))/100 > 0.0001)

   dag1 = nbrs{bestc};

   bnet = mk_bnet(dag1, ns);        % use the best dag now to produce a new bnet, with altered nodes labels

   for j=1:N                        % randomly set the CPTs values of each CPDs

      bnet.CPD{j} = tabular_CPD(bnet, j, 'prior_type', 'dirichlet', 'dirichlet_weight', 2);

   end

   BestBNet = bnet;

else

   count_init_bnets = count_init_bnets + 1;

end

if (count_init_bnets == 5)

   dag1 = mk_rnd_dag(N, 3)

   bnet = mk_bnet(dag1, ns);        % use the best dag now to produce a new bnet, with altered nodes labels

   for j=1:N                        % randomly set the CPTs values of each CPDs

      bnet.CPD{j} = tabular_CPD(bnet, j, 'prior_type', 'dirichlet', 'dirichlet_weight', 2);

   end

   count_init_bnets = 0;

   prev_score = 1000;

end
```

```
  prev_score = new_score;

end

[BestBNet, LOGLIKE] = learn_params_em(engine, samplesM, 30);      % default set the parameter EM runs 10
iterations
```

Κεφάλαιο 8: Αποτελέσματα

Τα αποτελέσματα της μεθοδολογίας που αναπτύχθηκε μπορούν να αποτιμηθούν σταδιακά μετά από κάθε ένα από τα βασικά βήματα αλλά και συνολικά. Τα βασικά βήματα της μεθοδολογίας όπως αυτά περιγράφηκαν στο προηγούμενο κεφάλαιο είναι τα εξής:

- προεπεξεργασία
- ομαδοποίηση ομο-εκφρασμένων γονιδίων
  - ο KMeans αλγόριθμος
  - ο Fuzzy KMeans αλγόριθμος
  - ο Clustering με χρήση GMM
- ανακατασκευή γενετικού δικτύου

Στο βήμα της προεπεξεργασίας πραγματοποιείται απομάκρυνση των γονιδίων που προσθέτουν θόρυβο στο σύνολο δεδομένων αλλά και η απομάκρυνση εκείνων των γονιδίων που δεν συμμετέχουν ενεργά στις διαδικασίες του κύκλου του κυττάρου. Στα στάδια της προεπεξεργασίας αφαιρέθηκαν 1789 γονίδια από τα 6179 που περιείχε το αρχικό σύνολο δεδομένων. Τα 4390 χρησιμοποιήθηκαν για μια πρώτη αποτίμηση της διαδικασίας, σύμφωνα με αυτή συγκρίναμε τα γονίδια που γνωρίζουμε από πειραματικές μετρήσεις [Spellman et al. 1998] ότι εκφράζονται κατά την διάρκεια του κύκλου του κυττάρου καθώς και με αυτά που σύμφωνα με τον [Spellman et al. 1998] ρυθμίζονται κατά την ίδια διαδικασία. Τα αποτελέσματα παρουσιάζονται στους δύο παρακάτω πίνακες, σύμφωνα με τη πρώτη σύγκριση από τα 104 γονίδια 5 δεν έχουν συμπεριληφθεί μετά την προεπεξεργασία δηλαδή ποσοστό επιτυχίας 95%. Για το δεύτερο σύνολο το οποίο έχει προκύψει από υπολογιστικές μεθόδους και όχι πειραματικές το αντίστοιχο ποσοστό επιτυχίας είναι 99%.

Γνωστά γονίδια που ρυθμίζονται κατά την διάρκεια του κύκλου του κυττάρου [Spellman et al. 1998]	**104**
Αριθμός γονιδίων που συμπεριλαμβάνονται στα clusters	99

Ποσοστό	**95,19%**

Πίνακας 8.1 Αριθμός γονιδίων που επιλέχθηκαν μετά τα στάδια της προεπεξεργασίας και περιλαμβάνονται στο σύνολο των 104 (πειραματικά) γνωστών γονιδίων

Γονίδια που ρυθμίζονται κατά την διάρκεια του κύκλου του κυττάρου κατά [Spellman et al. 1998]	**799**
Αριθμός γονιδίων που συμπεριλαμβάνονται στα clusters	**792**
Ποσοστό	**99,12%**

Πίνακας 8.2 Αριθμός γονιδίων που επιλέχθηκαν μετά τα στάδια της προεπεξεργασίας και περιλαμβάνονται στο σύνολο των 799 (υπολογιστικά) γνωστών γονιδίων

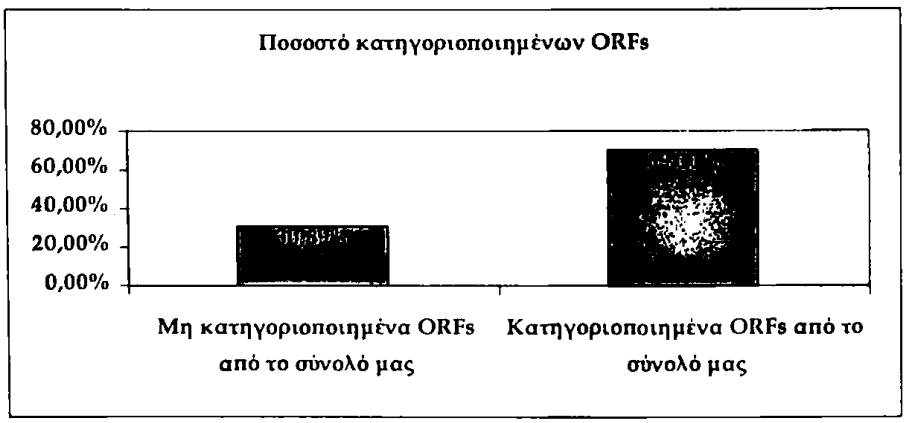
8.1 Αποτελέσματα ομαδοποίησης

Κατά την διάρκεια του δεύτερου βήματος πραγματοποιείτε ομαδοποίηση των ORFs σε ομάδες που έχουν την ίδια συμπεριφορά κατά την διάρκεια των μετρήσεων. Η ιδιότητα που θέλουμε να εκμεταλλευτούμε στο στάδιο αυτό είναι εκφράζεταιαπό την πρόταση "τα γονίδια που εκφράζονται με όμοιο τρόπο σε μια βιολογική διαδικασία έχουν και παρόμοιες ιδιότητες-λειτουργίες μέσα στο κύτταρο." Τα αποτελέσματα της ομαδοποίησης θα αποτιμηθούν κάνοντας χρήση του συνόλου των 104 γονιδίων που είναι γνωστό πειραματικά ότι εκφράζονται κατά την διάρκεια των πειραμάτων [Spellman et al. 1998], ενώ θα χρησιμοποιηθεί και η βάση δεδομένων CYGD για το yeast που παρέχεται από το MIPS (Munich Information Center for Protein Sequences) [Mewes et al. 1997], [MIPS]. Η MIPS είναι μια βάση δεδομένων στην οποία έχουν καταγραφεί τα ORFs του οργανισμού S. cerevisiae και οι λειτουργίες στις οποίες έχει βρεθεί πειραματικά ότι συμμετέχουν κατά την διάρκεια διαφόρων πειραμάτων. Οι ομάδες παρουσιάζονται στον πίνακα 2.3 με τον αριθμό των ORFs που έχει βρεθεί ότι συμμετέχουν σε αυτές.

Βιολογικές διαδικασίες	Καταγραφές
Metabolism	984
Energy	260
Cell Cycle And Dna Processing	688
Transcription	837
Protein Synthesis	380
Protein Fate (Folding, Modification, Destination)	630
Protein With Binding Function Or Cofactor Requirement (Structural Or Catalytic)	39
Protein Activity Regulation	27
Cellular Transport, Transport Facilitation And Transport Routes	718
Cellular Communication/Signal Transduction Mechanism	93
Cell Rescue, Defense And Virulence	294
Interaction With The Cellular Environment	331
Interaction With The Environment (Systemic)	2
Transposable Elements, Viral And Plasmid Proteins	118
Development (Systemic)	1
Biogenesis Of Cellular Components	447
Cell Type Differentiation	339
Subcellular Localization	11
Classification Not Yet Clear-Cut	114
Unclassified Proteins	2490

Πίνακας 8.3 Ομάδες καταγεγραμμένες από τον MIPS, στην πρώτη στήλη υπάρχουν οι "ομάδες" (βιολογικές διεργάσίες) και στην δεύτερη στήλη υπάρχει ο αριθμός των ORFs που έχουν βρεθεί ότι συμμετέχουν σε αυτές τις βιολογικές διαδικασίες

Το σύνολο δεδομένων που χρησιμοποιήσαμε στην διαδικασία της ομαδοποίησης αποτελείτε από 4390 ORFs, από αυτά το ~31% ανήκει στα μη κατηγοριοποιημένα ORFs σύμφωνα με την MIPS, ενώ τα υπόλοιπα (~69%) έχει βρεθεί ότι συμμετέχουν σε μία από τις 19 κατηγορίες όπως φαίνονται και στον πίνακα 2.3.

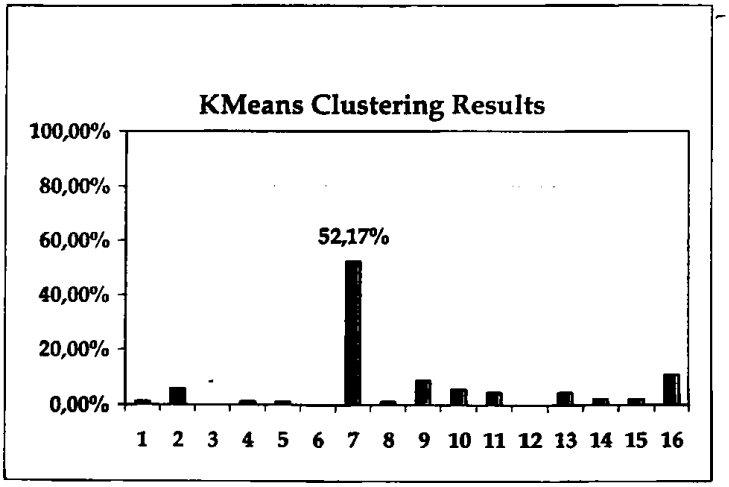


Εικόνα 8-1 Το σύνολο δεδομένων μας αποτελείται από 4390 ORFs, από αυτά με βάση τον MIPS το 30,89% δεν είναι κατηγοριοποιημένο ενώ το υπόλοιπο ~70% του

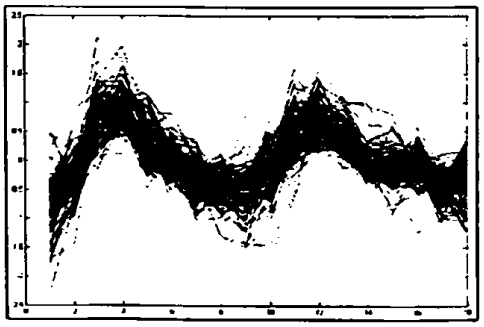
συνόλου δεδομένων μας είναι κατηγοριοποιημένο σε μία από τις βιολογικές διαδικασίες οι οποίες παρουσιάζονται στον πίνακα 2.3

Σε πρώτη φάση συγκρίναμε τα αποτελέσματα της ομαδοποίησης με το γνωστό σύνολο δεδομένων των 104 ORFs που γνωρίζουμε ότι συμμετέχουν στην διαδικασία του κύκλου του κυττάρου (βλ. Παράρτημα 4.1). Η αποτίμηση έγινε συγκρίνοντας κάθε ομάδα με το γνωστό σύνολο των 104 ORFs για κάθε αλγόριθμο ομαδοποίησης που εφαρμόσαμε, τα αποτελέσματα παρουσιάζονται γραφικά στις Εικόνες 2.2 - 2.7.

Αποτελέσματα ομαδοποίησης με χρήση KMeans

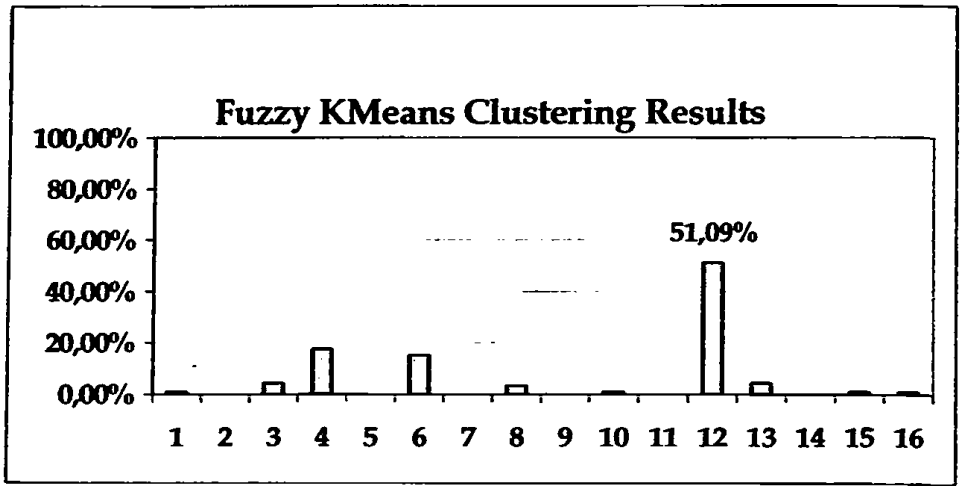


Εικόνα 8-2 Αποτελέσματα KMean αλγορίθμου, η 7η ομάδα περιέχει το 52,17% των γονιδίων που ρυθμίζονται κατά την διάρκεια του κύκλου του κυττάρου

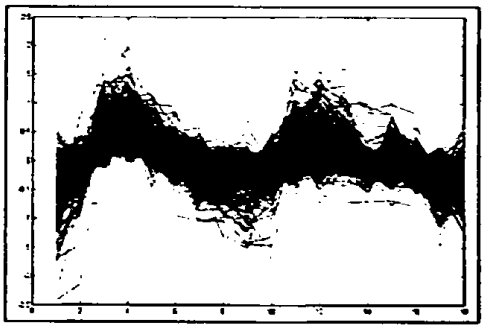


Εικόνα 8-3 Οι γραφική παράσταση των δεδομένων έκφρασης για την 7η ομάδα

Αποτελέσματα ομαδοποίησης με χρήση Fuzzy KMeans

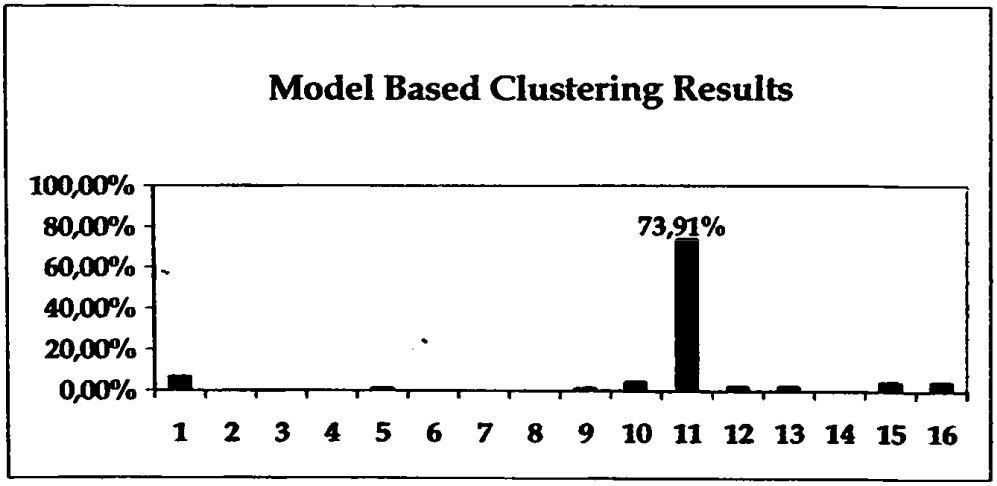


Εικόνα 8-4 Αποτελέσματα ομαδοποίησης με χρήση Fuzzy KMeans, η 12η ομάδα περιέχει το 51,09% των γονιδίων που ρυθμίζονται κατά την διάρκεια του κύκλου του κυττάρου

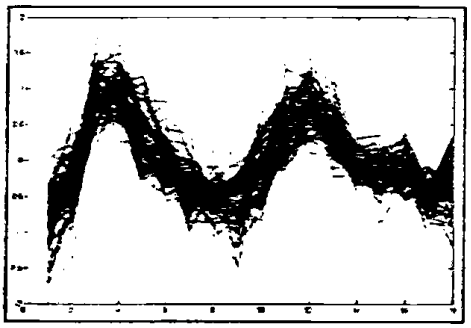


Εικόνα 8-5 Οι γραφική παράσταση των δεδομένων έκφρασης για την 12η ομάδα

Αποτελέσματα ομαδοποίησης με χρήση Gaussian Mixture Models



Εικόνα 8-6 Αποτελέσματα ομαδοποίησης με χρήση Gaussian Mixture Models, η 11η ομάδα περιέχει το 73,91% των γονιδίων που ρυθμίζονται κατά την διάρκεια του κύκλου του κυττάρου

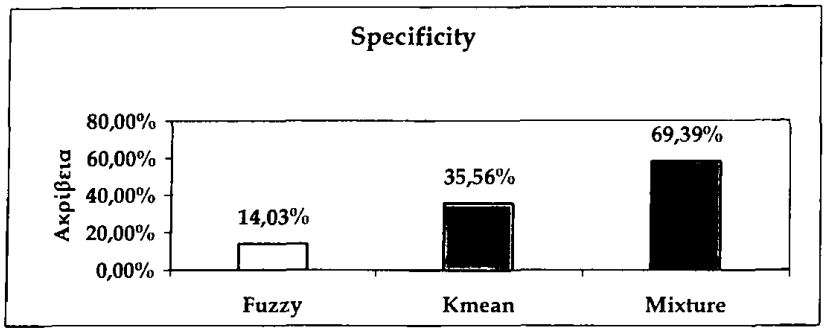


Εικόνα 8-7 Οι γραφική παράσταση των δεδομένων έκφρασης για την 11η ομάδα

Από τα παραπάνω αποτελέσματα παρατηρούμε ότι σε όλους τους αλγορίθμους ομαδοποίησης που εφαρμόσαμε υπάρχει μία ομάδα που περιέχει τα ORFs τα οποία συμμετέχουν στις βιολογικές διαδικασίες του κύκλου του κυττάρου, τα ORFs εκφράζονται με ένα συγκεκριμένο τρόπο όπως μπορούμε να δούμε και στις Εικόνες 2.3, 2.5, 2.7. Τα καλύτερα αποτελέσματα παρουσιάζονται στον αλγόριθμο GMM όπου εκτός από το ποσοστό της 11ης 74%, αρκετά σημαντικό χαρακτηριστικό είναι η ακρίβεια του αποτελέσματος, η οποία παρουσιάζεται στον παρακάτω πίνακα. Σύμφωνα με τον πίνακα η ομάδα στην οποία έχουν παρατηρηθεί τα περισσότερα "σωστά" ORFs από τον αλγόριθμο GMM παρουσιάζει ακρίβεια ~60%.

	Fuzzy KMeans	KMeans	GMM
Best Cluster	12	7	11
#ORFs in best cluster	335	135	98
Correct ORFs	47	48	68
Ακρίβεια	14,03%	35,56%	69,39%

Πίνακας 8.4 Στον πίνακα φαίνεται η ακρίβεια των αποτελεσμάτων για τη "καλύτερη" ομάδα κάθε αλγορίθμου



Εικόνα 8-8 Στην γραφική παράσταση φαίνονται τα δεδομένα του πίνακα 2.4

Παρατήρηση: Από τα πειραματικά δεδομένα γνωρίζουμε την ύπαρξη 104 ORFs που συμμετέχουν στην διαδικασία των πειραμάτων που εξετάζουμε, ωστόσο για τα υπόλοιπα (6180 - 104) ORFs που έχουν παρθεί μετρήσεις δεν υπάρχει κάποια γνώση για την συμπεριφορά τους στον κύκλο του κυττάρου. Κατά συνέπεια δεν μπορούμε να υπολογίσουμε τα μέτρα Sensitivity - Specificity για τους αλγορίθμους, αλλά ούτε και να αποτιμήσουμε τους αλγορίθμους με ακρίβεια.

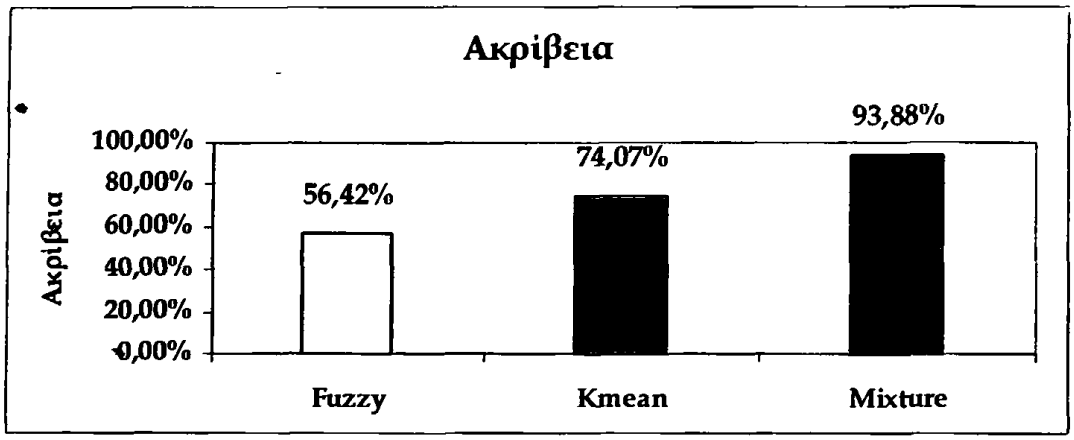
Τα ORFs που παρήχθησαν στην ομάδα 11 του GMM αλγορίθμου είναι τα εξής:

ORFs της ομάδας 11 από GMM

YAR007C	YIL066C	YEL075C	YMR179W
YBL035C	YIL140W	YEL076C	YMR199W
YBL111C	YIL141W	YEL077C	YNL082W
YBL112C	YIL177C	YER070W	YNL233W
YBL113C	YJL073W	YER095W	YNL262W
YBR070C	YJL074C	YER111C	YNL312W
YBR071W	YJL115W	YER189W	YNL339C
YBR088C	YJL225C	YER190W	YOL007C
YBR089W	YJR030C	YFL064C	YOL017W
YBR098W	YKL045W	YFL065C	YOL090W
YCL022C	YKL108W	YFL066C	YOR074C
YCL024W	YKL113C	YFL067W	YPL014W
YCL060C	YKR077W	YFL068W	YPL057C
YCL061C	YLL066C	YGL038C	YPL153C
YCR065W	YLL067C	YGR151C	YPL221W
YDL003W	YLR049C	YGR152C	YPL267W
YDL018C	YLR103C	YGR189C	YPL283C
YDL101C	YLR183C	YGR221C	YPR120C
YDL163W	YLR462W	YGR296W	YPR135W
YDL164C	YLR463C	YHL049C	YPR174C
YDR097C	YLR464W	YHL050C	YPR175W
YDR507C	YLR465C	YHR149C	YPR202W
YDR528W	YLR466W	YHR218W	YPR203W
YDR545W	YLR467W	YHR219W	YPR204W
YEL040W	YML027W		

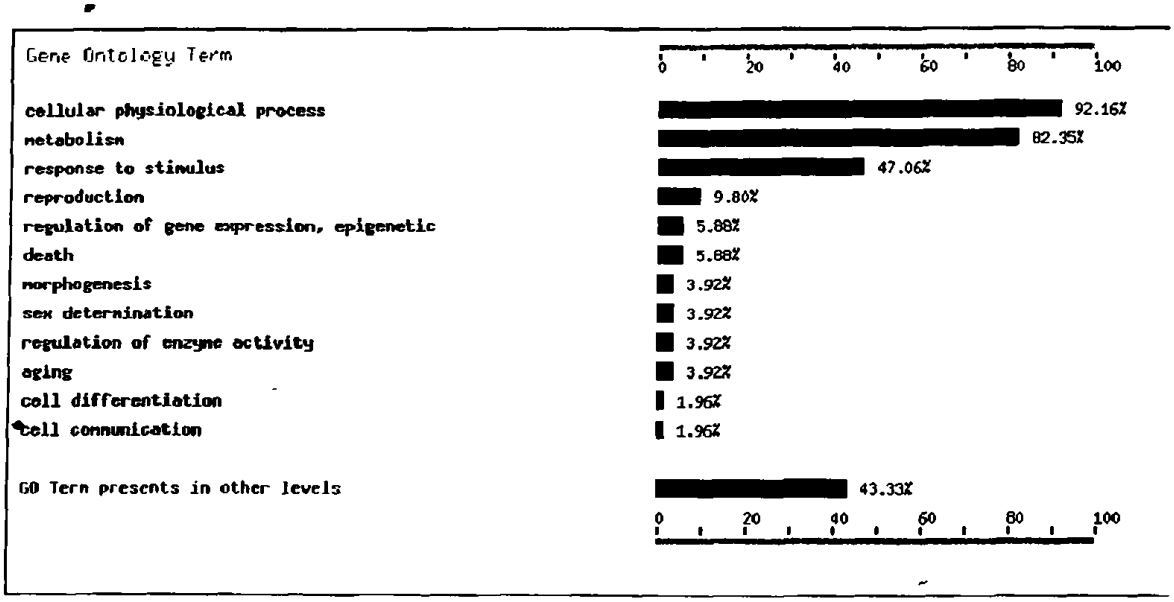
Πίνακας 8.5 **ORFs της ομάδας 11 από GMM**

Τα αντίστοιχα αποτελέσματα χρησιμοποιώντας την βάση MIPS και πιο συγκεκριμένα την κατηγορία Cell Cycle and DNA Processing παρουσιάζονται παρακάτω.

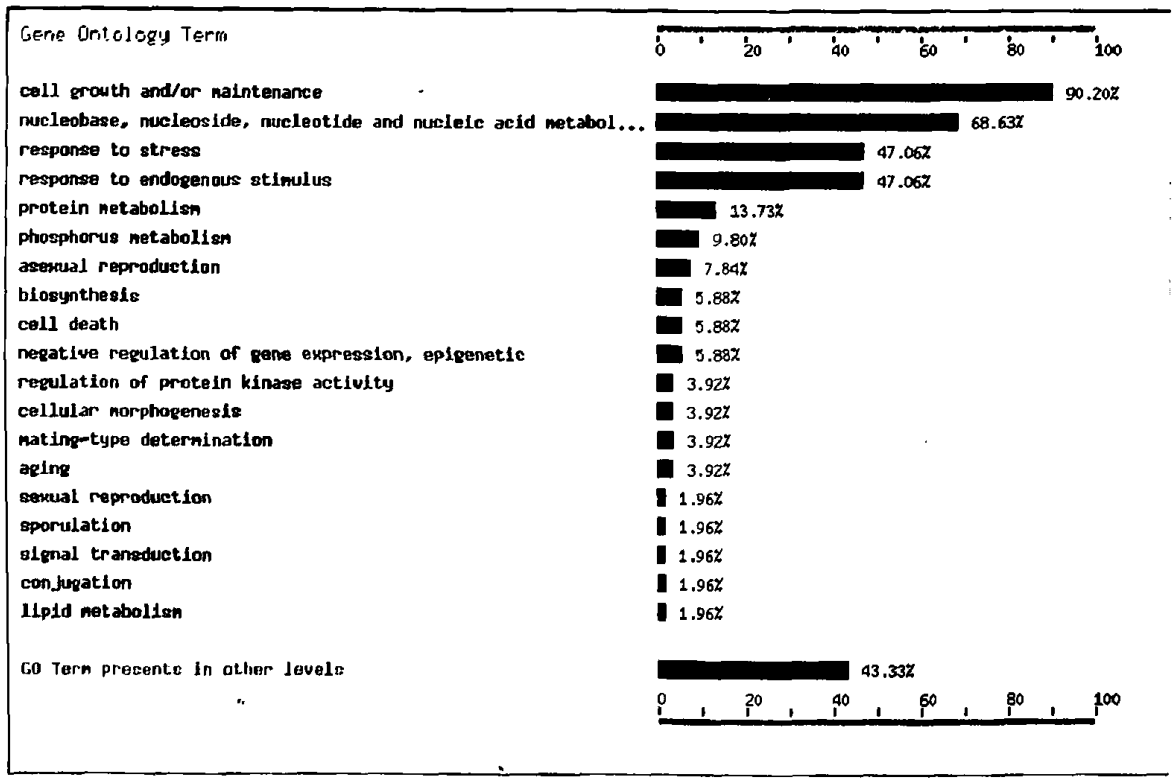


Εικόνα 8-9 Στο γράφημα φαίνεται η ακρίβεια των αποτελεσμάτων για τη "καλύτερη" ομάδα κάθε αλγορίθμου (με βάση τη MIPS)

Χρησιμοποιώντας την οντολογία Gene Ontology και το εργαλείο FatiGO (http://fatigo.bioinfo.cnio.es) μπορούμε να κάνουμε μια δεύτερη αποτίμηση των δεδομένων. Σύμφωνα με τις εικόνες (Εικόνα 2.10 εφόσον περιέχει περισσότερο εξειδικευμένες ομάδες) μπορούμε να παρατηρήσουμε ότι το 90% των ORFs της ομάδας 11 του GMM συμμετέχουν στην διαδικασία cell growth and/or maintenance (GO:0008151), αντίστοιχα το 69% στην nucleobase, nucleoside, nucleotide and nucleic acid metabolism (GO:0006139), 47% στην response to stress (GO:0006950) και 47% στην response to endogenous stimulus (GO:0009719). Στην εικόνα 2.9 παρουσιάζονται οι υπερ-ομάδες (επίπεδο 3) από όπου επίσης παρατηρούμε ότι τα ORFs της ομάδας 11 του GMM αποτελούν το 92% των ORFs που έχουν εντοπιστεί ότι συμμετέχουν στην διαδικασία cellular physiological process.

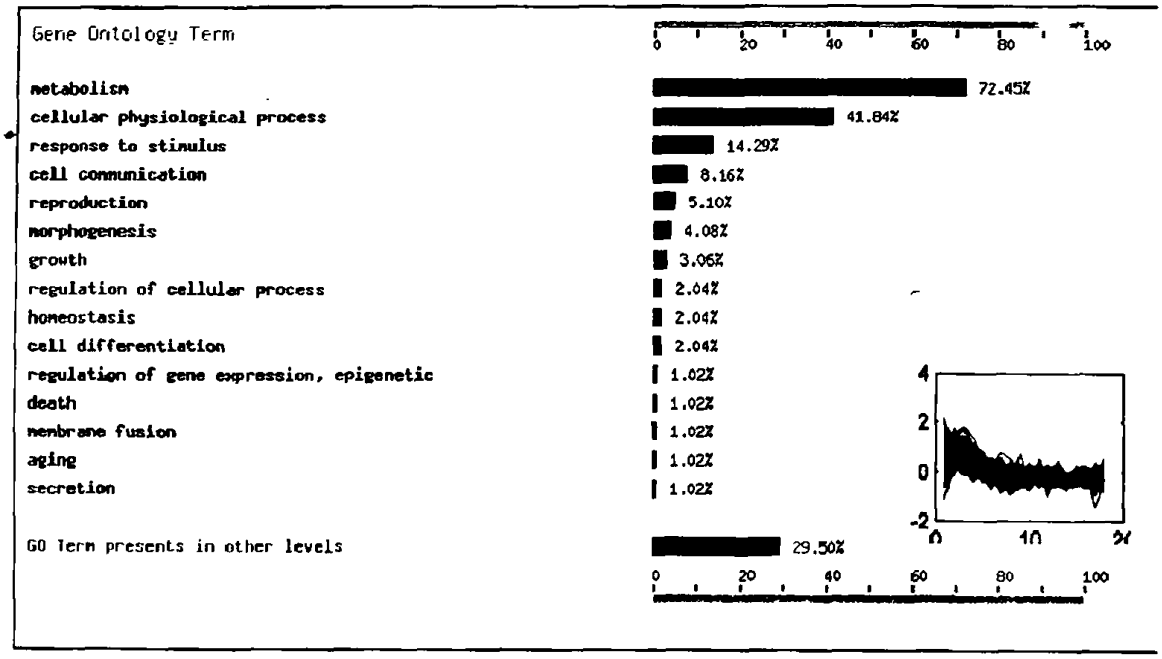


Εικόνα 8-10 Αποτελέσματα με χρήση της GO του επιπέδου 3, το 92% των ORFs της 11ης ομάδας συμμετέχουν στη διαδικασία cellular physiological process (GO:0050875)



Εικόνα 8-11 Αποτελέσματα με χρήση της GO του επιπέδου 4 (περισσότερο εξειδικευμένες ομάδες από το επίπεδο 3), το 90% των ORFs της 11ης ομάδας συμμετέχουν στη διαδικασία cell growth and/or maintenance (GO:0008151)

Ενδιαφέρον επίσης παρουσιάζουν και οι υπόλοιπες ομάδες όπως για παράδειγμα η ομάδα 2 του ίδιου αλγορίθμου όπου τα ORFs αυτής αποτελούν το ~73% των ORFs που έχει βρεθεί ότι συμμετέχουν στην διαδικασία του μεταβολισμού, όπως φαίνεται στην εικόνα



Εικόνα 8-12 Αποτελέσματα με χρήση της GO του επιπέδου 3, το 73% τνη ORFs της 2ης ομάδας συμμετέχουν στη διαδικασία metabolism

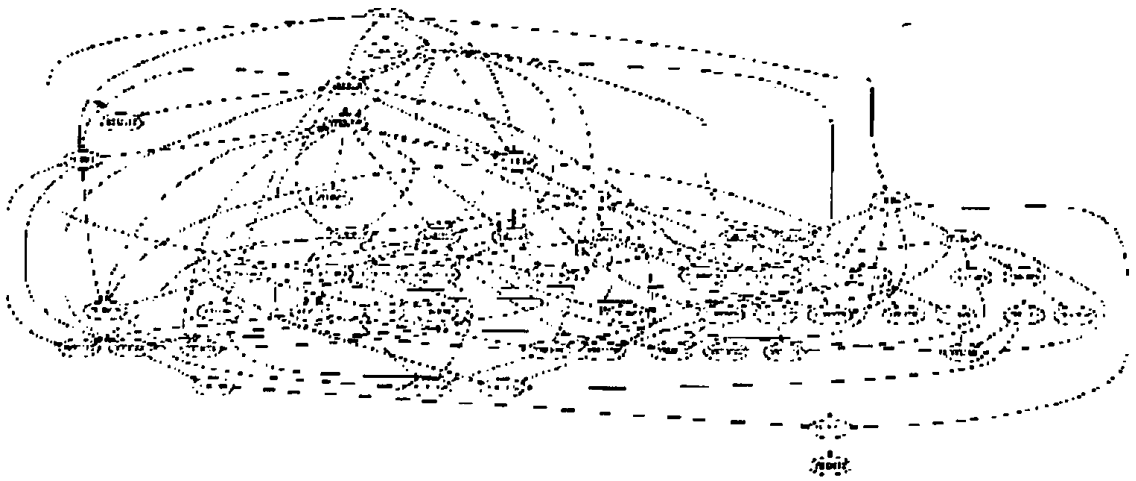
8.2 Αποτελέσματα ανακατασκευής γενετικών δικτύων

Σε αυτή τη παράγραφο αναφέρουμε τα τελικά αποτελέσματα που προκύπτουν μετά την ολοκλήρωση όλων των βημάτων της μεθόδου. Τα αποτελέσματα χωρίζονται σε δύο βασικές κατηγορίες, σε αυτά που έχουν προκύψει χωρίς εφαρμογή εκ των προτέρων γνώση σχετικά με τις αλληλεπιδράσεις των πρωτεϊνών και άλλων μορίων μέσα στο κύτταρο και στα αποτελέσματα στα οποία χρησιμοποιήθηκε κάποια γνώση στην αρχικοποίηση του μοντέλου. Η γνώση που προστέθηκε στα τελευταία πειράματα έχει προέλθει από βάσεις δεδομένων που περιέχουν πληροφορίες σχετικά με αλληλεπιδράσεις μορίων μέσα οε ένα κύτταρο. Οι βάσεις δεδομένων που χρησιμοποιήθηκαν είναι οι BIND - Biomolecular Interaction Network Database, DIP - Database of Interacting Proteins, PathCalling Yeast Interaction Database και MINT - a Molecular Interactions Database. Παρακάτω περιγράφονται τα αποτελέσματα καθώς και μια αποτίμηση αυτών δεδομένου ότι οι βάσεις που προαναφέρθηκαν δεν είναι πλήρεις, δηλαδή δεν περιέχουν όλες τις αντιδράσεις που πραγματοποιούνται

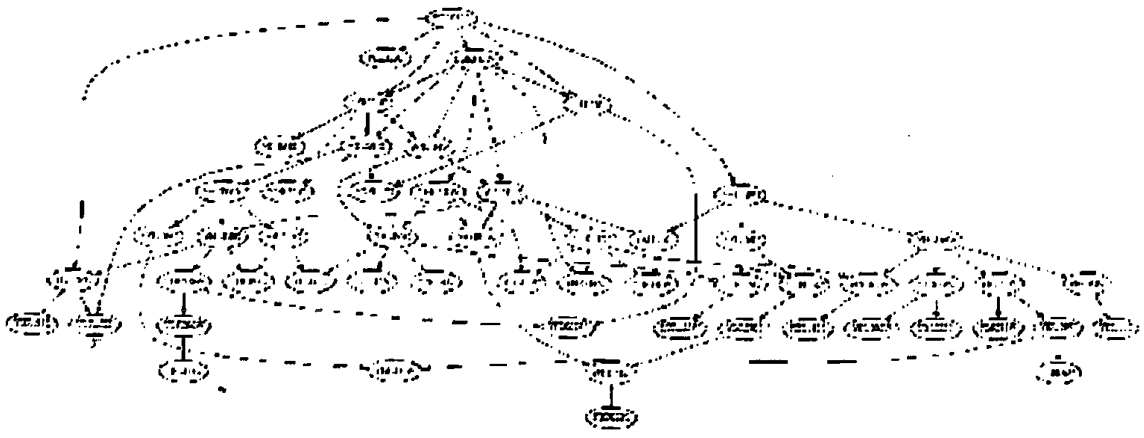
κατά την διάρκεια του κύκλου του κυττάρου του σακχαρομύκητα (yeast). Η είσοδος στον αλγόριθμο για την εκπαίδευση του Bayesian μοντέλου ήταν η 11η ομάδα από τον αλγόριθμο των μεικτών κατανομών το οποίο είχε 95 κόμβους. Για ένα μοντέλο 95 κόμβων υπάρχουν περίπου 2^{95} διαφορετικές δομές οι οποίες πρέπει να εξεταστούν αν θέλουμε να βρούμε την βέλτιστη λύση. Στις εκτελέσεις του αλγορίθμου SEM ωστόσο ξεκινώντας από ένα αρχικό μοντέλο με τοπικές αλλαγές καταλήγουμε σε μια βέλτιστη λύση για την η οποία εξαρτάται πάρα πολύ από την αρχική δομή. Για τον λόγο αυτό εκτελούμε τον αλγόριθμο πολλές φορές με αρχική δομή σε κάθε επανάληψη το μοντέλο που προέκυψε από την επόμενη. Για να συγκρίνουμε τα αποτελέσματα έχουμε εφαρμόσει τον αλγόριθμο για 15, 30, 50, 100, 150 και 300 επαναλήψεις, των οποίων τα αποτελέσματα παρουσιάζονται στα σχήματα παρακάτω.

8.2.1 Χωρίς εκ των προτέρων γνώση

Στα σχήματα της παραγράφου που ακολουθεί θα δούμε αποτελέσματα γενετικών δικτύων των οποίων το αρχικό μοντέλο ήταν ο απλούστερος γράφος δηλαδή μια αλυσίδα. Η δομή αυτή επιλέχθηκε διότι δεν αποδίδει στο μοντέλο καμία αρχική πόλωση σε κάποια μερικός γνωστή δομή. Τα αποτελέσματα παρουσιάζονται παρακάτω για 50 επαναλήψεις του αλγορίθμου.



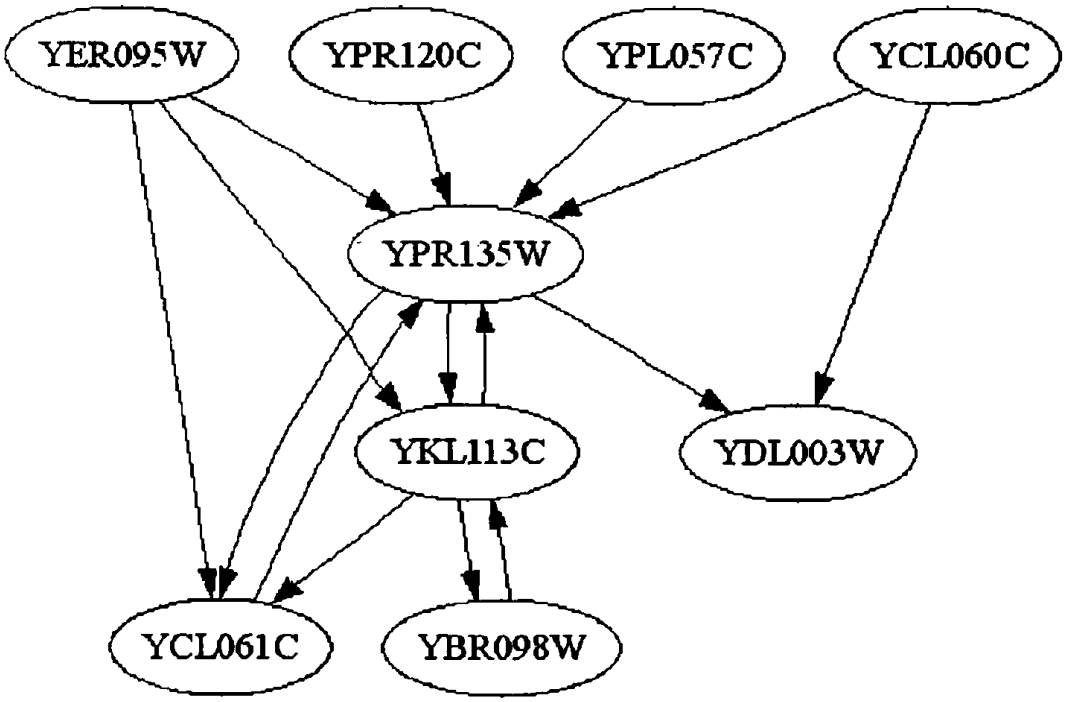
Εικόνα 8-13 Το αρχικό μοντέλο



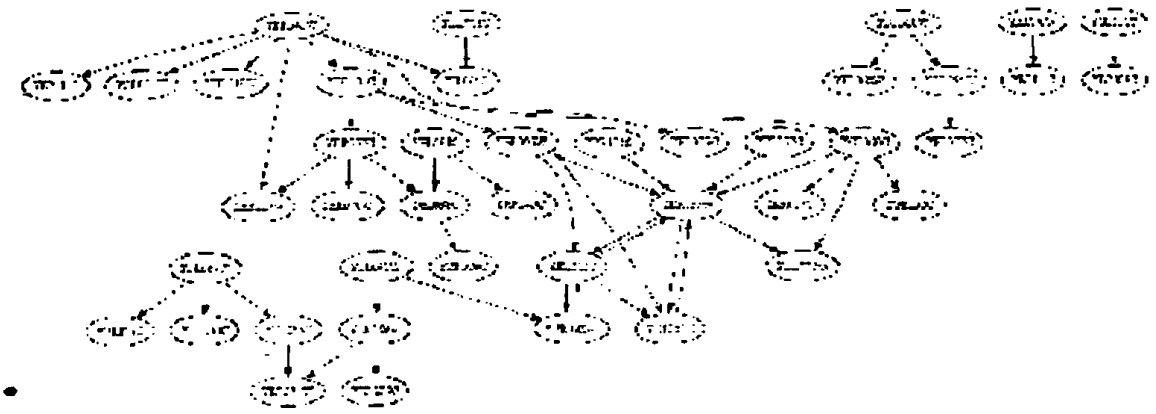
Εικόνα 8-14 Ο αλγόριθμος εκτελέστηκε για 50 φορές, με αρχικό μοντέλο εκείνο της εικόνας 8.13

8.2.2 Με εφαρμογή εκ των προτέρων γνώσης

Όπως έχουμε ήδη αναφέρει η αφετηρία μας σε αυτή τη μεθοδολογία είναι ένας πίνακας δεδομένων γονιδίων έκφρασης, όπου κάθε γραμμή αντιπροσωπεύει ένα γονίδιο και κάθε στήλη αντιπροσωπεύει ένα πείραμα. Κάθε στοιχείο, x_{ij}, του X δείχνει το επίπεδο έκφρασης ενός γονιδίου i σε μια χρονική στιγμή του πειράματος. Ο πίνακας αυτός δεν περιέχει κάποιου είδους εκ των προτέρων πληροφορία η οποία μπορεί να χρησιμοποιηθεί απευθείας στο μοντέλο. Για να μεγαλώσουμε την βεβαιότητα των τελικών μοντέλων χρησιμοποιήσαμε εκ των προτέρων γνώση που εκφράζεται σε γνωστές αλληλεπιδράσεις μεταξύ των ORFs, οι οποίες παρουσιάζονται στο παρακάτω σχήμα.



Εικόνα 8-15 Ο γράφος αναπαριστά το αρχικό μοντέλο, το οποίο αναπαριστά όλες τις γνωστές αλληλεπιδράσεις μεταξύ των ORF στο yeast. Το πεδίο των αλληλεπιδράσεων για το yeast δεν πλήρως γνωστό. Κατά συνέπεια υπάρχουν πολλές αλληλεπιδράσεις που δεν είναι γνωστές. Οι αλληλεπιδράσεις που φαίνονται στο παραπάνω γράφο αναφέρονται στα ORFs τα οποία αποτελούν το σύνολο δεδομένων μας για το βήμα την ανακατασκευής του γενετικού δικτύου. Οι κόμβοι του δικτύου αποτελούν τα γονίδια ενώ οι ακμές τις υπό συνθήκη εξαρτήσεις



Εικόνα 8-16 Ο αλγόριθμος εκτελέστηκε για 50 φορές, με αρχικό μοντέλο εκείνο της εικόνας 8.20 που απεικονίζει τις γνωστές αλληλεπιδράσεις για το σύνολο δεδομένων που έχουμε

Χωρίς αρχική γνώση στο μοντέλο	
Επαναλήψεις Αλγορίθμου	**50**
Σύνολο αλληλεπιδράσεων που βρέθηκαν	76
Σωστές αλληλεπιδράσεις	7
Γνωστές αλληλεπιδράσεις του συνόλου δεδομένων: 12	
Ευαισθησία	**58%**
Ακρίβεια	**10%**

Πίνακας 8.6 Αποτελέσματα ανακατασκευής χωρίς εφαρμογή γνώσης στην αρχικοποίηση του μοντέλου, το αρχικό μοντέλο είναι μια αλυσίδα

Με αρχική γνώση στο μοντέλο	
Επαναλήψεις Αλγορίθμου	**50**
Σύνολο αλληλεπιδράσεων που βρέθηκαν	47

Σωστές αλληλεπιδράσεις	9
Γνωστές αλληλεπιδράσεις του συνόλου δεδομένων: 12	
Ευαισθησία	**75%**
Ακρίβεια	**20%**

Πίνακας 8.7 Αποτελέσματα ανακατασκευής με εφαρμογή γνώσης στην αρχικοποίηση του μοντέλου

Κεφάλαιο 9: Συμπεράσματα

Στις προηγούμενες παραγράφους αναφερθήκαμε στην δημιουργία μοντέλων και την εκπαίδευσή τους με σκοπό την ανακατασκευή των γενετικών δικτύων από δεδομένα έκφρασης γονιδίων. Τα αποτελέσματα τα οποία παρουσιάστηκαν είναι αφενός ικανοποιητικά δεδομένου ότι τείνουν στην παρουσίαση ενός μοντέλου όμοιο με εκείνου που γνωρίζουμε για τα ρυθμιστικά δίκτυα, ωστόσο υπάρχουν ορισμένες σημαντικές παραδοχές που κάναμε πριν φτάσουμε σε αυτά. Αρχικά οι αλγόριθμοι εκπαίδευσης των Bayesian δικτύων εφαρμόστηκαν σε ένα πολύ μικρό υποσύνολο σε σχέση με το αρχικό σύνολο δεδομένων, η πολυπλοκότητα των αλγορίθμων εκπαίδευσης και το πλήθος των μεταβλητών εμποδίζει την εφαρμογή των ίδιων αλγορίθμων στο αρχικό σύνολο δεδομένων. Επίσης ένας σημαντικός περιορισμό προέρχεται από την γνώση η οποία υπάρχει για αυτά τα συστήματα η οποία είναι αρκετά γενική σε σχέση με το στόχο μας, ο οποίος είναι η ακριβής μοντελοποίηση ενός είδους γενετικού δικτύου.

Ένα άλλο επίσης σημαντικό θέμα σε αυτές τις προσεγγίσεις είναι η εξάρτηση μεταξύ των δύο στόχων όσον αφορά την μοντελοποίηση με γραφικά μοντέλα. Από την μια, στοχεύουμε στην ανακατασκευή λεπτομερών μοντέλων ενώ παράλληλα θέλουμε να είμαστε σε θέση να εκπαιδεύσουμε αυτά τα μοντέλα από τα διαθέσιμα δεδομένα. Αποτέλεσμα αυτού του tradeoff είναι η εξειδίκευση των μεθοδολογιών που προτείνονται με αποτέλεσμα να συλλαμβάνεται ένα μόνο μέρος του προβλήματος, ίσως το κομμάτι που θεωρείται πιο σημαντικό για τη βιολογία.

Η εφαρμογή των τεχνικών τεχνητής νοημοσύνης στη βιοπληροφορική είναι ακόμα σε ένα αρχικό στάδιο. Όπως έχει όμως αποδειχθεί οι τεχνικές αυτές μπορεί να έχουν μεγάλη συνεισφορά σε αυτήν την περιοχή. Η βιοπληροφορική παρέχει επίσης στον τομέα της τεχνητής νοημοσύνης μια ευκαιρία να εξεταστούν οι τεχνικές σε μια νέα περιοχή και να τροποποιηθούν στα πλαίσια των συγκεκριμένων προβλημάτων (όπως συνέβη με τα HMMs, όπου η αρχική αρχιτεκτονική των HMM για την επεξεργασία του φυσικής γλώσσας τροποποιήθηκε ώστε να καταστήσει πιο ικανή στην ανάλυση ακολουθιών DNA). Η διαφορά μεταξύ των προηγούμενων προβλημάτων όπου οι τεχνικές αυτές έχουν εφαρμοστεί, με τη βιοπληροφορική είναι ότι τα δεδομένα στην

τελευταία αυξάνονται με ένα απίστευτα γρήγορο ρυθμό, με τη δυνατότητά να δίνεται η ευκαιρία ανάλυσης και να μοντελοποίησης αυτών των δεδομένων.

Παρόλα αυτά οι υπάρχουσες προτάσεις για ανακατασκευή γενετικά ρυθμιστικών δικτύων από microarray δεδομένα έχουν αγνοήσει πολλές από τις δυσκολίες που υπάρχουν για την αξιόπιστη ανάλυση δεδομένων. Η αύξηση του πλήθους των δεδομένων και η βελτίωση της ποιότητας των δεδομένων θα βελτίωνε πολύ τις προοπτικές για επιτυχή εφαρμογή των τρεχόντων τεχνικών. Από έρευνες που έχουν γίνει δεν φαίνεται να υπάρχουν μεγάλα εμπόδια για την συλλογή των δεδομένων που θα βελτιώσουν την απόδοση των Μπεϋζιανών αλγορίθμων για την ανακατασκευή των γενετικών δικτύων

Κάποια μελλοντική εργασία πάνω στο θέμα της ανακατασκευής των ρυθμιστικών δικτύων από δεδομένα έκφρασης περιλαμβάνει την χρήση ενοποιημένων μοντέλων που συνδυάζουν δεδομένα από διαφορετικά επίπεδα του κυτταρικού μηχανισμού. Μια βασική δυναμική των γραφικών μοντέλων είναι η δυνατότητα να οριστούν τα πρότυπα που θα διαχειριστούν τέτοιες ετερογενείς παρατηρήσεις. Στην εργασία αυτή μιλήσαμε για ένα μοντέλο που βρίσκει τις περιοχές αλληλεπιδράσεων και χαρακτηρίζει ταυτόχρονα τη συμπεριφορά των γονιδίων. Μια άλλη κατεύθυνση είναι να ενσωματωθούν τα δεδομένα από πειράματα της κατηγορίας των proteomics. Για παράδειγμα, πρόσφατες εργασίες συνδυάζουν τους χάρτες αλληλεπίδρασης πρωτεΐνης-DNA και πρωτεΐνης - πρωτεΐνης για να ανακατασκευάσουν τα ρυθμιστικά δίκτυα που εξηγούν τη knockout έκφραση των γονιδίων στα πειράματα.

Ένα άλλο βασικό συστατικό για τη βελτίωση των μοντέλων είναι η κατανόησή των βιολογικών ρυθμιστικών μηχανισμών. Όταν πιθανόν, ενσωματωθούν βιολογικές αρχές στην σχεδίαση μοντέλων (π.χ., εισάγοντας περιορισμούς στην δομή του δικτύου) μπορεί να περιορίσει τους βαθμούς ελευθερίας κατά τη διάρκεια της εκπαίδευσης και να οδηγήσει σε καλύτερα αποτελέσματα. Η πρόκληση για το μέλλον στον τομέα αυτό είναι η κατανόηση των επιλογών μοντελοποίησης για τα διαφορετικά κυτταρικά συστατικά και της καταλληλότητάς τους για τους διαφορετικούς τύπους πειραμάτων, ώστε να επεκταθούν οι μέθοδοι στη συμπερασματολογία και την εκπαίδευση τέτοιων προτύπων.

Κεφάλαιο 10: Βιβλιογραφία

Η υλοποίηση των αλγορίθμων ομαδοποίησης έγιναν στην Matlab με χρήση του Clustering Toolbox, David Peter Alfred Corney το οποίο συνοδεύεται από GNU Library General Public License:

http://www.cs.ucl.ac.uk/staff/D.Corney/ClusteringMatlab.html

Στην υλοποίηση για αναπαράσταση και εκπαίδευση των Bayesian Networks, χρησιμοποιήθηκε το BNT (Bayes Net Toolbox for Matlab), Kevin Murphy το οποίο συνοδεύεται από GNU Library General Public License

http://www.ai.mit.edu/~murphyk

[Cooper et al. 1992] G.F. Cooper, and E., Herskovitz, 'A Bayesian method for the induction of probabilistic netwroks from data', Machine Learning, 9, 309-347, (1992).

[Friedman et al. 1987] Friedman Nir (1987), Learning belief networks in the presence of missing values and hidden variables. In Fourteenth Inter. Conf. on Machine Learning (ICML).

[Friedman et al. 1998] Friedman Nir (1998), The Bayesian structural EM algorithm, In Fourteenth Conf. on Uncertainty in Artificial Intelligence.

[Spellman et al. 1998] Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B: Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell 1998, 9:3273-3297

[D'haeseleer et al. 2000] D'haeseleer P, Liang S, Somogyi R: "Genetic network inference: from co-expression clustering to reverse engineering". Bioinformatics 2000, 16:707-726.

[Pe'er et al. 2001] Pe'er D, Regev A, Elidan G, Friedman N. "Inferring subnetworks from perturbed expression profiles. Bioinformatics" 2001, 17(Suppl 1):S215-S224.

[Akutsu et al. 2000] Akutsu T, Miyano S, Kuhara S: "Algorithms for inferring qualitative models of biological networks" Pac Symp Biocomput 2000, 293-304.

[Friedman et al. 2000] Friedman N, Linial M, Nachman I, Pe'er D: "Using Bayesian networks to analyze expression data" J Comput Biol 2000, 7:601-620.

[Kauffman et al. 1969] Kauffman, SA. "Metabolic Stability and Epigenesis in Randomly Connected Nets", Journal of Theoretical Biology 1969, 22:437

[Cho et al. 1998] Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L, Wolfsberg TG, Gabrielian AE, Landsman D, Lockhart DJ, Davis RW: A genome-wide transcriptional analysis of the mitotic cell cycle. Mol Cell 1998, 2:65-73.

[PubMed] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi

[YPD] http://www.proteome.com/databases/YPD/YPDsearch-quick.html

[Troyanskaya et al. 2001] Troyanskaya O. Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein R Altman R. Missing value estimation methods for DNA micorarrays. Bioinformatics 2001 17:520-525.

[Bezdek et al. 1981] Bezdek J.C. Pattern recognition with fuzzy objective function algorithms. New York 1981

[Fukayama et al. 1989] Fukayama Y., Sugeno M. "A new method of choosing the number of clusters for the fuzzy c-means method." Proc. 5th Fuzzy Syst. Symp. Pp247-250, 1989

[Xie et al. 1991] Xie N. L. Beni G. A. "A validity measure for fuzzy clustering", IEEE Trans. PAMI, vol.13 no.8, pp841-847, 1991

[Kim et al. 2001], Kim D.J., Y.W. Park and D.J. Park, "A Novel Validity Index for Determination of the Optimal Number of Clutters", IEICE Transactions on Information and Systems, Vol. E84-D, No. 2, P. 281 - 285, 2001, 5

[Heckerman et al. 1994] Heckerman D., D. Geiger, D. Chickering. Learning Bayesian networks: The Combination of Knowledge and Statistical Data. Technical Report MSR-TR-94-09, Microsoft Research, March, 1994 (revised December, 1994).

[MIPS] http://mips.gsf.de/

[Mewes et al. 1997] Mewes HW, Albermann K, Bähr M, Frishman D, Gleissner A, Hani J, Heumann K, Kleine K, Maierl A, Oliver SG, Pfeiffer F, Zollner A Overview of the yeast genome. Nature 387(6632 Suppl):7-65 (1997)

[Akutsu et al. 1999] Akutsu, T., Miyano, S., Kuhara, S.: "Identification of genetic networks from a small number of gene expression patterns under theboolean network model", Pacific Symp. Biocomp. 99 4: 17-28, 1999

[Arlinghaus et al. 1997] Arlinghaus, H.F., Kwoka, M.N., Jacobson, K.B.: Analysis of biosensor chips for identification of nucleic acids", Anal. Chem. 69: 3747-3753, 1997.

[Arnone et al. 1997] Arnone, M.I., Davidson, E.H.: "The hardwiring of development: organization and function of genomic regulatory systems", Development 124: 1851-1864, 1997.

[Brazma et al. 1998] Brazma, A., Jonassen, I., Vilo, J., Ukkonen, E.: "Predicting gene regulatory elements in silico on a genomic scale", Genome Res. 8: 1202-1215, 1998.

[Wessels et al. 2001] Wessels, L.F., van Someren, E.P., Reinders, M.J.: "A comparison of genetic network models", Pac Symp Biocomput 2001 6: 508-519, 2001.

[Yuh et al. 2001] Yuh C-H, Bolouri H, Davidson EH Cis-regulatory logic in the endo16 gene: switching from a specification to a differntiation mode of control. Development 128:617-629, 2001

[Yuh et al. 1998] Yuh C-H, Bolouri H, Davidson EH Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene. Science 279:1896-1902, 1998

[Friedman et al. 2000] Friedman N, Linial M, Nachman I, Pe'er D: "Using Bayesian networks to analyze expression data" J Comput Biol 2000, 7:601-620.

[Meyers & Friedland 1984] Meyers S and Friedland P "Knowledge-based simulation of genetic regulation in bacteriophage lambda" Nucleic Acids Res. 1984 12: 1-9.

[Poolman et al. 2001] Poolman, M. G., Ölçer, H., Lloyd, J. C., Raines, C. A., and Fell, D. A. Computer modelling and experimental evidence for two steady states in the photosynthetic Calvin cycle. Eur. J. Biochem. 268, 2810-2816, 2001

[Hynne et al. 2001] Hynne, F., Danø, S., and Sørensen, P. G. Full-scale model of glycolysis in Saccharomyces cerevisiae. Biophysical Chemistry 94, 121-163, 2001

[Kierzek et al. 2002] Kierzek, A. M. STOCKS: STOChastic Kinetic Simulations of biochemical systems with gillespie algorithm. Bioinformatics 18, 470-481, 2002

[Bray et al. 2001], Bray, D., Firth, C., Le Novere, N., and Shimizu, T. StochSim, 2001

[Meinhardt et al. 2001], Meinhardt, H., and de Boer, P. A. J. Pattern formation in Escherichia coli: A model for the pole-to-pole oscillations of Min proteins and the localization of the division site. PNAS 98, 14202-14207, 2001

[Kauffman et al. 1993], Kauffman, S. A. "The Origins of Orders: Self-organization and Selection in Evolution". New York: Oxford University Press, 1993.

[GO] http://www.geneontology.org

[Acid et al. 1996], Acid, S. and Campos, L.M., "BENEDICT: An algorithm for learning probabilistic belief networks," Proceedings of the sixth International Conference IPMU'96, 1996.

[Cheng et al. 1997], J., Bell, D.A., and Liu, W., "Learning belief networks from data: An information theory based approach," Proceeding of the sixth ACM International Conference on Information and Knowledge Management, 1997.

[Chow et al. 1968], Chow, C.K., Liu, C.N., "Approximating discrete probabilities distributions with dependence trees," IEEE Transactions on Information Theory, 1968; 14(3):462-467.

[Cooper et al. 1992], Cooper, G.F., Herskovits, E. "A Bayesian method for the induction of probabilistic networks from data." Machine Learning, 1992; 9:309-347.

[Cowell et al. 1999] Cowell, R.G., Dawid, A. P., Lauritzen, S.L., Spiegelhalter, D.J., "Probabilistic Networks and Expert Systems," Springer-Verilag, New York, 1999.

[Friedman et al. 1998], Friedman, N. "The Bayesian structural EM algorithm," in G.F. Cooper and S. Moral (Eds.), Proceedings of Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI '98), San Francisco, CA: Morgan Kaufmann, 1998.

[Heckerman et al. 1995], Heckerman, D., "A tutorial on learning Bayesian networks," Technical Report MSR-TR-95-06, Microsoft Corporation, 1995.

[Heckerman et al. 1994], Heckerman, D, Geiger, D., Chickering, D.M. "Learning Bayesian Networks: the combination of knowledge and statistical data," 1994 pages: 293-301.

[Krause et al. 1996], Krause, P., "Learning probabilistic networks," Technical Report, Philips research laboratories, UK, 1996.

[Lam et al. 1994], Lam, W., Bacchus, F. "Learning Bayesian belief networks: An approach based on the MDL principle," Computational Intelligence, 1994; 10(3):269-293.

[Lauritzen et al. 1995], Lauritzen, S.L., "The EM algorithm for graphical association models with missing data," Computational Statistics and Data Analysis, vol. 19, pp. 191-201, 1995.

[Neapolitan et al. 1990], Neapolitan, R. E. "Probabilistic Reasoning in Expert Systems, Theory and Algorithms," John Wiley and Sons, Inc., 1990.

[Oliver et al. 1990], Oliver, R.M., Smith, J.Q. "Influence Diagrams, Belief Nets and Decision Analysis," New York: Wiley, 1990.

[Onisko et al. 1997], Onisko, A. and Druzdzel, M.J. , "Application of Bayesian Belief Networks to Diagnosis of Liver Disorders," Proceedings of Third Conference on Neural Networks and their Applications, Poland, 1997:730-736

[Pearl et al. 1998], Pearl, J. "Probabilistic Reasoning in Intelligent Systems," Morgan Kaufmann, San Mateo, California, 1988.

[Ramoni et al. 1997], Ramoni, M. and Sebastiani, P., "Parameter estimation in Bayesian networks from incomplete databases," Technical Report KMi-TR-57, Knowledge Median Institute, The Open University, November 1997.

[Ramoni et al. 1997], Ramoni, M. and Sebastiani, P., "Learning Bayesian networks from incomplete data," Technical Report KMi-TR-43, Knowledge Median Institute, The Open University, February 1997.

[Ravi et al. 1985], Ravi, R. "A necessary and sufficient condition for acyclic graphs," Proceedings of the IEEE, 1985; 73(2):369-370.

[Russel et al. 1995], Russel ,S.J., Binder, J., Koller, D. and Kanazawa, K. "Local learning in probabilistic networks with hidden variables," Proceedings of the 14th International Joint Conference on Artificial Intelligence (ed. C.S. Mellish), Morgan Kaufmann, San Mateo, California.1995;1146-1152.

[Sebastiani et al. 1997], Sebastiani, P. and Ramoni, M., "Bayesian inference with missing data using bound and collapse," Technical Report KMi-TR-58, Knowledge Median Institute, The Open University, November 1997.

[Spiegelhalter et al. 1993], Spiegelhalter, D.J., et al. "Bayesian analysis in expert systems," Statistical Science, 1993; 8(3): 219-283.

[Suzuki et al. 1996], Suzuki, J., "Learning Bayesian belief networks based on the MDL principle: An efficient algorithm using the branch and bound technique," Proceedings of the international conference on machine learning, Bari, Italy, 1996

[Wagner et al. 2001] Wagner, A.. "How to Reconstruct a Genetic Network from n single-gene perturbations in fewer than n2 easy steps", Bioinformatics 17: 1183-1197, 2001.

[Golub et al. 1999] Golub T.R., Slonim D. K., Tamayo P., Huard C., Gaasenbeek M., Mesirov J.P., Coller H., Loh M. L. Downing J. R., Caligiuri M. A., Bloomfield C. D., Lander E. S.. (1999) "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring" Science Vol 286 pp 531-536

[Su et al. 2002] Su, T., Basu, M., Toure, A., "Multi-Domain Gating Network for Classification of Cancer Cells using Gene Expression Data" Proceedings of the International Joint Conference on Neural Networks (IJCNN'02), Honolulu, Hawaii, pp 286-289, ISBN 0-7803-7279-4, 2002

[D'Haeseleer et al. 1999] D'Haeseleer P., Liang S., Somogyi R., (1999) "Gene Expression Analysis and Modelling", Proceedings of Pacific Symposium on Biocomputing, Hawaii, (PSB99)

[Kauffman et al. 1996] Kauffman, S. (1996) At Home in the Universe: The Search for Laws of Self-Organization and Complexity. Penguin Books

[Wuensche et al. 1998] Wuensche,A., "Genomic Regulation Modeled as a Network with Basins of Attraction", in "Pacific Symposium on Biocomputing '98" eds. R.B.Altman, A.K.Dunker, L.Hunter, T.E.Klien. World Scientific, Singapore, 1998

[Ando et al. 2001a] Ando, S., Iba H.,. (2001a) "Inference of Gene Regulatory Model by Genetic Algorithms", Proceedings of Conference on Evolutionary Computation 2001 pp712-719

[Ando et al. 2001b] Ando, S., Iba H., (2001b) "The Matrix Modeling of Gene Regulatory Networks -Reverse Engineering by Genetic Algorithms-", Proceedings of Atlantic Symposium on Computational Biology, and Genome Information Systems & Technology 2001.

[Ando et al. 2000] Ando, S., Iba H., (2000) "Inference of Gene Regulatory Model by Genetic Algorithms", Proceeding of International Symposium on Adaptive Systems

[Toh et al. 2000] Toh, H., Horimoto, K., (2000) Inference of Genetic Networks from Expression Profile by Graphical Gaussian Modeling Genome Informatics 11: 242-244 (2000)

[Spirtes et al. 2000] Spirtes P., Glymour C., Scheines R., Kauffmann S., Aimale V., Wimberly F. (2000) "Constructing Bayesian Network Models of Gene Expression Networks from Microarray Data" In Proceedings of the Atlantic Symposium on Computational Biology, Genome Information Systems & Technology, 2000.

Κεφάλαιο 11: Παράρτημα

11.1 Γνωστά ρυθμιστικά γονίδια

Υπάρχουν 104 γονίδια που έχει αποδειχθεί με πειραματικές μεθόδους ότι εκφράζονται κατά την διάρκεια του κύκλου του κυττάρου.

ORFs γνωστών ρυθμιζόμενων γονιδίων κατά την διάρκεια του κύκλου του κυττάρου

YAL040C	YGR108W	YDR033W	YLR342W	YBR202W	YKL096W
YAR007C	YGR109C	YDR077W	YLR452C	YCL027W	YKL096W-A
YAR018C	YHR152W	YDR097C	YML021C	YCL055W	YKL101W
YBL002W	YIL106W	YDR146C	YMR001C	YDL003W	YKL113C
YBL003C	YJL092W	YDR150W	YMR198W	YDL055C	YKL185W
YBL035C	YJL115W	YDR224C	YMR199W	YDL102W	YLR079W
YBR009C	YJL157C	YDR225W	YMR307W	YDL127W	YLR103C
YBR010W	YJL173C	YDR309C	YNL030W	YDL164C	YLR131C
YBR054W	YJL187C	YDR356W	YNL031C	YDL179W	YLR210W
YBR067C	YJL194W	YER001W	YNL082W	YDL197C	YLR274W
YBR083W	YKL042W	YER070W	YNL102W	YDL227C	YLR286C
YBR088C	YKL045W	YER095W	YNL145W	YPR119W	YPR159W
YER111C	YNL192W	YGL163C	YNL312W	YGR092W	YOL090W
YFL026W	YNL262W	YGL225W	YNL327W	YPR120C	YOR058C
YGL116W	YNL289W	YGR044C	YNR044W	YPR141C	YOR074C
YPR175W	YPL153C	YPL256C			

M/G1 Boundary (SWI5 or ECB (MCM1) or STE12/MCM1 dependent):

AGA1[1], ASH1[2], CDC46[3], CDC47[3], CDC6[3, 4], CHS1[5], CLN3[3], CTS1[6], EGT2[7], FUS1[8, 9], MFA2[1], PCL2[10], PCL9[10], RME1[11], SIC1[12, 13], SST2[1], STE2[8], SWI4[3, 14], TEC1[15].

Late G1, SCB regulated:

CLN1[16], CLN2[16], CSD2/CHS3[17], FKS1/CWH53[17, 18], GAS1[17, 18], HO[19], KAR4[20], KRE6[17], MNN1[17], PCL1[21], PSA1[22], SWE1[23], TIP1[24], VAN2/GOG5[17].

Late G1, MCB regulated:

ASF1[25], ASF2[25], CDC21[26, 27], CDC45[28], CDC8[29], CDC9[30], CLB5[31], CLB6[32], DBF4[33], DPB2[34], DPB3[35], GIC2[36], MCD1[37], MSH2[38, 39], MSH6[39], NIK1/HSL1[40], PDS1[41], PMS1[39, 42], POL1[43], POL12[44], POL2[45], POL3/CDC2[46], POL30[46], PRI1[47], PRI2[48], RAD17[49], RAD27[50], RAD51[51], RAD54[52], RFA1[53], RFA2[53], RFA3[53], RNR1[54], RNR3[55], SPC110/NUF1[56], SPC42[57], SPK1[58], SRS2/HPR5[59], UNG1[52].

S-phase:

Histones: HHT1, HHT2, HHF1, HHF2[60], HTA1, HTA2, HTB1, HTB2[61, 62].

S/G2-phase:

CDC14[63], CIK1[20], CLB3[64], CLB4[64], CWP1[24], CWP2[24], KAR3[20], NUM1[65], TIR1[24].

G2/M-phase:

ACE2[6], ASE1[66], CDC20[67], CDC5[68], CLB1[69, 70], CLB2[69, 70], DBF2[71], FAR1[72], KIN3[73], MOB1[74], YRO2(MST1)[75], YDR033w(MST2)[73], SED1[24], SPO12[76], SWI5[77].

1. Oehlen, L.J., McKinney, J.D., and Cross, F.R. (1996). Ste12 and Mcm1 regulate cell cycle-dependent transcription of FAR1. *Mol Cell Biol* 16, 2830-7.

2. **Bobola, N., Jansen, R.P., Shin, T.H., and Nasmyth, K. (1996).** Asymmetric accumulation of Ash1p in postanaphase nuclei depends on a myosin and restricts yeast mating-type switching to mother cells. *Cell* **84**, 699-709.

3. **McInerny, C.J., Partridge, J.F., Mikesell, G.E., Creemer, D.P., and Breeden, L.L. (1997).** A novel Mcm1-dependent element in the SWI4, CLN3, CDC6, and CDC47 promoters activates M/G1-specific transcription. *Genes Dev* **11**, 1277-88.

4. **Zwerschke, W., Rottjakob, H.W., and Kuntzel, H. (1994).** The Saccharomyces cerevisiae CDC6 gene is transcribed at late mitosis and encodes a ATP/GTPase controlling S phase initiation. *J Biol Chem* **269**, 23351-6.

5. **Pammer, M., Briza, P., Ellinger, A., Schuster, T., Stucka, R., Feldmann, H., and Breitenbach, M. (1992).** DIT101 (CSD2, CAL1), a cell cycle-regulated yeast gene required for synthesis of chitin in cell walls and chitosan in spore walls. *Yeast* **8**, 1089-99.

6. **Dohrmann, P.R., Butler, G., Tamai, K., Dorland, S., Greene, J.R., Thiele, D.J., and Stillman, D.J. (1992).** Parallel pathways of gene regulation: homologous regulators SWI5 and ACE2 differentially control transcription of HO and chitinase. *Genes Dev* **6**, 93-104.

7. **Kovacech, B., Nasmyth, K., and Schuster, T. (1996).** EGT2 gene transcription is induced predominantly by Swi5 in early G1. *Mol Cell Biol* **16**, 3264-74.

8. **Zanolari, B. and Riezman, H. (1991).** Quantitation of alpha-factor internalization and response during the Saccharomyces cerevisiae cell cycle. *Mol Cell Biol* **11**, 5251-8.

9. **Oehlen, L.J. and Cross, F.R. (1994).** G1 cyclins CLN1 and CLN2 repress the mating factor response pathway at Start in the yeast cell cycle. *Genes Dev* **8**, 1058-70.

10. **Aerne, B.L., Johnson, A.L., Toyn, J.H., and Johnston, L.H. (1998).** Swi5 controls a novel wave of cyclin synthesis in late mitosis. *Mol Biol Cell* **9**, 945-56.

11. **Toone, W.M., Johnson, A.L., Banks, G.R., Toyn, J.H., Stuart, D., Wittenberg, C., and Johnston, L.H. (1995).** Rme1, a negative regulator of meiosis, is also a positive activator of G1 cyclin gene expression. *EMBO J* **14**, 5824-32.

12. **Knapp, D., Bhoite, L., Stillman, D.J., and Nasmyth, K. (1996).** The transcription factor Swi5 regulates expression of the cyclin kinase inhibitor p40SIC1. *Mol Cell Biol* **16**, 5701-7.

13. **Donovan, J.D., Toyn, J.H., Johnson, A.L., and Johnston, L.H. (1994).** P40SDB25, a putative CDK inhibitor, has a role in the M/G1 transition in Saccharomyces cerevisiae. *Genes Dev* **8**, 1640-53.

14. **Breeden, L. and Mikesell, G.E. (1991).** Cell cycle-specific expression of the SWI4 transcription factor is required for the cell cycle regulation of HO transcription. *Genes Dev* **5**, 1183-90.

15. **Oehlen, L. and Cross, F.R. (1998).** The mating factor response pathway regulates transcription of TEC1, a gene involved in pseudohyphal differentiation of Saccharomyces cerevisiae . *FEBS Lett* **429**, 83-8.

16. **Wittenberg, C., Sugimoto, K., and Reed, S.I. (1990).** G1-specific cyclins of S. cerevisiae: cell cycle periodicity, regulation by mating pheromone, and association with the p34CDC28 protein kinase. *Cell* **62**, 225-37.

17. **Igual, J.C., Johnson, A.L., and Johnston, L.H. (1996).** Coordinated regulation of gene expression by the cell cycle transcription factor Swi4 and the protein kinase C MAP kinase pathway for yeast cell integrity. *EMBO J* **15**, 5001-13.

18. **Ram, A.F., Brekelmans, S.S., Oehlen, L.J., and Klis, F.M. (1995).** Identification of two cell cycle regulated genes affecting the beta 1,3- glucan content of cell walls in Saccharomyces cerevisiae. *FEBS Lett* **358**, 165-70.

19. **Nasmyth, K. (1983).** Molecular analysis of a cell lineage. *Nature* **302**, 670-6.

20. **Kurihara, L.J., Stewart, B.G., Gammie, A.E., and Rose, M.D. (1996).** Kar4p, a karyogamy-specific component of the yeast pheromone response pathway. *Mol Cell Biol* **16**, 3990-4002.

21. **Ogas, J., Andrews, B.J., and Herskowitz, I. (1991).** Transcriptional activation of CLN1, CLN2, and a putative new G1 cyclin (HCS26) by SWI4, a positive regulator of G1-specific transcription. *Cell* **66**, 1015-26.

22. **Benton, B.K., Plump, S.D., Roos, J., Lennarz, W.J., and Cross, F.R. (1996).** Over-expression of S. cerevisiae G1 cyclins restores the viability of alg1 N-glycosylation mutants. *Curr Genet* **29**, 106-13.

23. **Ma, X.J., Lu, Q., and Grunstein, M. (1996).** A search for proteins that interact genetically with histone H3 and H4 amino termini uncovers novel regulators of the Swe1 kinase in Saccharomyces cerevisiae. *Genes Dev* **10**, 1327-40.

24. **Caro, L.H., Smits, G.J., van Egmond, P., Chapman, J.W., and Klis, F.M. (1998).** Transcription of multiple cell wall protein-encoding genes in Saccharomyces cerevisiae is differentially regulated during the cell cycle . *FEMS Microbiol Lett* **161**, 345-9.

25. **Le, S., Davis, C., Konopka, J.B., and Sternglanz, R. (1997).** Two new S-phase-specific genes from Saccharomyces cerevisiae. *Yeast* **13**, 1029-42.

26. **McIntosh, E.M., Gadsden, M.H., and Haynes, R.H. (1986).** Transcription of genes encoding enzymes involved in DNA synthesis during the cell cycle of Saccharomyces cerevisiae. *Mol Gen Genet* **204**, 363-6.

27. **McIntosh, E.M., Ord, R.W., and Storms, R.K. (1988).** Transcriptional regulation of the cell cycle-dependent thymidylate synthase gene of Saccharomyces cerevisiae. *Mol Cell Biol* **8**, 4616-24.

28. **Hardy, C.F. (1997).** Identification of Cdc45p, an essential factor required for DNA replication. *Gene* **187**, 239-46.

29. **White, J.H., Green, S.R., Barker, D.G., Dumas, L.B., and Johnston, L.H. (1987).** The CDC8 transcript is cell cycle regulated in yeast and is expressed coordinately with CDC9 and CDC21 at a point preceding histone transcription. *Exp Cell Res* **171**, 223-31.

30. **White, J.H., Barker, D.G., Nurse, P., and Johnston, L.H. (1986).** Periodic transcription as a means of regulating gene expression during the cell cycle: contrasting modes of expression of DNA ligase genes in budding and fission yeast. *EMBO J* **5**, 1705-9.

31. **Epstein, C.B. and Cross, F.R. (1992).** CLB5: a novel B cyclin from budding yeast with a role in S phase. *Genes Dev* **6**, 1695-706.

32. **Schwob, E. and Nasmyth, K. (1993).** CLB5 and CLB6, a new pair of B cyclins involved in DNA replication in Saccharomyces cerevisiae. *Genes Dev* 7, 1160-75.

33. **Chapman, J.W. and Johnston, L.H. (1989).** The yeast gene, DBF4, essential for entry into S phase is cell cycle regulated. *Exp Cell Res* **180**, 419-28.

34. **Araki, H., Hamatake, R.K., Johnston, L.H., and Sugino, A. (1991).** DPB2, the gene encoding DNA polymerase II subunit B, is required for chromosome replication in Saccharomyces cerevisiae. *Proc Natl Acad Sci U S A* **88**, 4601-5.

35. **Araki, H., Hamatake, R.K., Morrison, A., Johnson, A.L., Johnston, L.H., and Sugino, A. (1991).** Cloning DPB3, the gene encoding the third subunit of DNA polymerase II of Saccharomyces cerevisiae. *Nucleic Acids Res* **19**, 4867-72.

36. **Jaquenoud, M., Gulli, M.-P., Peter, K., and Peter, M. (1998).** The Cdc42 effector Gic2 is targeted for ubiquitin-dependent degradation by the SCF^{GRR1} complex. *EMBO J* 17, 5360-5373.

37. **Guacci, V., Koshland, D., and Strunnikov, A. (1997).** A direct link between sister chromatid cohesion and chromosome condensation revealed through the analysis of MCD1 in S. cerevisiae . *Cell* **91**, 47-57.

38. **Yang, Y., Johnson, A.L., Johnston, L.H., Siede, W., Friedberg, E.C., Ramachandran, K., and Kunz, B.A. (1996).** A mutation in a Saccharomyces cerevisiae gene (RAD3) required for nucleotide excision repair and transcription increases the efficiency of mismatch correction. *Genetics* **144**, 459-66.

39. **Kramer, W., Fartmann, B., and Ringbeck, E.C. (1996).** Transcription of mutS and mutL-homologous genes in Saccharomyces cerevisiae during the cell cycle. *Mol Gen Genet* **252**, 275-83.

40. **Tanaka, S. and Nojima, H. (1996).** Nik1: a Nim1-like protein kinase of S. cerevisiae interacts with the Cdc28 complex and regulates cell cycle progression. *Genes Cells* **1**, 905-21.

41. **Yamamoto, A., Guacci, V., and Koshland, D. (1996).** Pds1p is required for faithful execution of anaphase in the yeast, Saccharomyces cerevisiae. *J Cell Biol* **133**, 85-97.

42. **Morrison, A., Johnson, A.L., Johnston, L.H., and Sugino, A. (1993).** Pathway correcting DNA replication errors in Saccharomyces cerevisiae. *EMBO J* **12**, 1467-73.

43. **Johnston, L.H., White, J.H., Johnson, A.L., Lucchini, G., and Plevani, P. (1987).** The yeast DNA polymerase I transcript is regulated in both the mitotic cell cycle and in meiosis and is also induced after DNA damage. *Nucleic Acids Res* **15**, 5017-30.

44. **Toyn, J.H., Toone, W.M., Morgan, B.A., and Johnston, L.H. (1995).** The activation of DNA replication in yeast. *Trends Biochem Sci* **20**, 70-3.

45. **Araki, H., Ropp, P.A., Johnson, A.L., Johnston, L.H., Morrison, A., and Sugino, A. (1992).** DNA polymerase II, the probable homolog of mammalian DNA polymerase epsilon, replicates chromosomal DNA in the yeast Saccharomyces cerevisiae. *EMBO J* **11**, 733-40.

46. **Bauer, G.A. and Burgers, P.M. (1990).** Molecular cloning, structure and expression of the yeast proliferating cell nuclear antigen gene. *Nucleic Acids Res* **18**, 261-5.

47. **Johnston, L.H., White, J.H., Johnson, A.L., Lucchini, G., and Plevani, P. (1990).** Expression of the yeast DNA primase gene, PRI1, is regulated within the mitotic cell cycle and in meiosis. *Mol Gen Genet* **221**, 44-8.

48. **Foiani, M., Santocanale, C., Plevani, P., and Lucchini, G. (1989).** A single essential gene, PRI2, encodes the large subunit of DNA primase in Saccharomyces cerevisiae. *Mol Cell Biol* 9, 3081-7.

49. **Siede, W., Nusspaumer, G., Portillo, V., Rodriguez, R., and Friedberg, E.C. (1996).** Cloning and characterization of RAD17, a gene controlling cell cycle responses to DNA damage in Saccharomyces cerevisiae. *Nucleic Acids Res* **24**, 1669-75.

50. **Reagan, M.S., Pittenger, C., Siede, W., and Friedberg, E.C. (1995).** Characterization of a mutant strain of Saccharomyces cerevisiae with a deletion of the *RAD27 gene*, a structural homolog of the RAD2 nucleotide excision repair gene. *J Bacteriol* **177**, 364-71.

51. **Basile, G., Aker, M., and Mortimer, R.K. (1992).** Nucleotide sequence and transcriptional regulation of the yeast recombinational repair gene RAD51. *Mol Cell Biol* **12**, 3235-46.

52. **Johnston, L.H. and Johnson, A.L. (1995).** The DNA repair genes RAD54 and UNG1 are cell cycle regulated in budding yeast but MCB promoter elements have no essential role in the DNA damage response. *Nucleic Acids Res* **23**, 2147-52.

53. **Brill, S.J. and Stillman, B. (1991).** Replication factor-A from Saccharomyces cerevisiae is encoded by three essential genes coordinately expressed at S phase. *Genes Dev* **5**, 1589-600.

54. **Elledge, S.J. and Davis, R.W. (1990).** Two genes differentially regulated in the cell cycle and by DNA- damaging agents encode alternative regulatory subunits of ribonucleotide reductase. *Genes Dev* **4**, 740-51.

55. **Huang, M. and Elledge, S.J. (1997).** Identification of RNR4, encoding a second essential small subunit of ribonucleotide reductase in Saccharomyces cerevisiae. *Mol Cell Biol* **17**, 6105-13.

56. **Kilmartin, J.V., Dyos, S.L., Kershaw, D., and Finch, J.T. (1993).** A spacer protein in the Saccharomyces cerevisiae spindle poly body whose transcript is cell cycle-regulated. *J Cell Biol* **123**, 1175-84.

57. **Donaldson, A.D. and Kilmartin, J.V. (1996).** Spc42p: a phosphorylated component of the S. cerevisiae spindle pole body (SPD) with an essential function during SPB duplication. *J Cell Biol* **132**, 887-901.

58. **Zheng, P., Fay, D.S., Burton, J., Xiao, H., Pinkham, J.L., and Stern, D.F. (1993).** SPK1 is an essential S-phase-specific gene of Saccharomyces cerevisiae that encodes a nuclear serine/threonine/tyrosine kinase. *Mol Cell Biol* **13**, 5829-42.

59. **Heude, M., Chanet, R., and Fabre, F. (1995).** Regulation of the Saccharomyces cerevisiae Srs2 helicase during the mitotic cell cycle, meiosis and after irradiation. *Mol Gen Genet* **248**, 59-68.

60. **Cross, S.L. and Smith, M.M. (1988).** Comparison of the structure and cell cycle expression of mRNAs encoded by two histone H3-H4 loci in Saccharomyces cerevisiae. *Mol Cell Biol* **8**, 945-54.

61. **Hereford, L.M., Osley, M.A., Ludwig, T.R.d., and McLaughlin, C.S. (1981).** Cell-cycle regulation of yeast histone mRNA. *Cell* **24**, 367-75.

62. **Hereford, L., Bromley, S., and Osley, M.A. (1982).** Periodic transcription of yeast histone genes. *Cell* **30**, 305-10.

63. **Wan, J., Xu, H., and Grunstein, M. (1992).** CDC14 of Saccharomyces cerevisiae. Cloning, sequence analysis, and transcription during the cell cycle. *J Biol Chem* **267**, 11274-80.

64. **Fitch, I., Dahmann, C., Surana, U., Amon, A., Nasmyth, K., Goetsch, L., Byers, B., and Futcher, B. (1992).** Characterization of four B-type cyclin genes of the budding yeast Saccharomyces cerevisiae. *Mol Biol Cell* **3**, 805-18.

65. **Farkasovsky, M. and Kuntzel, H. (1995).** Yeast Num1p associates with the mother cell cortex during S/G2 phase and affects microtubular functions. *J Cell Biol* **131**, 1003-14.

66. **Pellman, D., Bagget, M., Tu, Y.H., and Fink, G.R. (1995).** Two microtubule-associated proteins required for anaphase spindle movement in Saccharomyces cerevisiae [published erratum appears in J Cell Biol 1995 Oct;131(2):561]. *J Cell Biol* **130**, 1373-85.

67. **Shirayama, M., Zachariae, W., Ciosk, R., and Nasmyth, K. (1998).** The Polo-like kinase Cdc5p and the WD-repeat protein Cdc20p/fizzy are regulators and substrates of the anaphase promoting complex in Saccharomyces cerevisiae. *EMBO J* **17**, 1336-49.

68. **Kitada, K., Johnson, A.L., Johnston, L.H., and Sugino, A. (1993).** A multicopy suppressor gene of the Saccharomyces cerevisiae G1 cell cycle mutant

gene dbf4 encodes a protein kinase and is identified as CDC5. *Mol Cell Biol* **13**, 4445-57.

69. **Ghiara, J.B., Richardson, H.E., Sugimoto, K., Henze, M., Lew, D.J., Wittenberg, C., and Reed, S.I. (1991).** A cyclin B homolog in S. cerevisiae: chronic activation of the Cdc28 protein kinase by cyclin prevents exit from mitosis. *Cell* **65**, 163-74.

70. **Surana, U., Robitsch, H., Price, C., Schuster, T., Fitch, I., Futcher, A.B., and Nasmyth, K. (1991).** The role of CDC28 and cyclins during mitosis in the budding yeast S. cerevisiae. *Cell* 65, 145-61.

71. **Johnston, L.H., Eberly, S.L., Chapman, J.W., Araki, H., and Sugino, A. (1990).** The product of the Saccharomyces cerevisiae cell cycle gene DBF2 has homology with protein kinases and is periodically expressed in the cell cycle. *Mol Cell Biol* **10**, 1358-66.

72. **McKinney, J.D., Chang, F., Heintz, N., and Cross, F.R. (1993).** Negative regulation of FAR1 at the Start of the yeast cell cycle. *Genes Dev* **7**, 833-43.

73. **Schuster, T., Price, C., Rossoll, W., and Kovacech, B. (1997).** New cell cycle-regulated genes in the yeast Saccharomyces cerevisiae. *Recent Results Cancer Res* **143**, 251-61.

74. **Komarnitsky, S.I., Chiang, Y.C., Luca, F.C., Chen, J., Toyn, J.H., Winey, M., Johnston, L.H., and Denis, C.L. (1998).** DBF2 protein kinase binds to and acts through the cell cycle-regulated MOB1 protein. *Mol Cell Biol* **18**, 2100-7.

75. **Price, C., Nasmyth, K., and Schuster, T. (1991).** A general approach to the isolation of cell cycle-regulated genes in the budding yeast, Saccharomyces cerevisiae. *J Mol Biol* **218**, 543-56.

76. **Parkes, V. and Johnston, L.H. (1992).** SPO12 and SIT4 suppress mutations in DBF2, which encodes a cell cycle protein kinase that is periodically expressed. *Nucleic Acids Res* **20**, 5617-23.

77. **Nasmyth, K., Seddon, A., and Ammerer, G. (1987).** Cell cycle regulation of SW15 is required for mother-cell-specific HO transcription in yeast. *Cell* **49**, 549-58.