COMPUTATIONAL METHODS FOR KNOWLEDGE DISCOVERY FROM HETEROGENEOUS DATA SOURCES: METHODOLOGY AND IMPLEMENTATION ON BIOLOGICAL AND MOLECULAR SOURCES

by

Lefteris Koumakis

A dissertation submitted in partial fulfilment of the requirements for the degree

Doctor of Philosophy

Technical University of Crete,

School of Production Engineering and Management

September 2014

Dissertation is approved

1. **Professor Vassilis S. Moustakis** ……………………………………

2. **Dr. George Potamias** ……………………………………

3. **Professor Michael Zervakis** ……………………………………

4. **Professor Nikolaos Bilalis** ……………………………………

5. **Professor Manolis Tsiknakis** ……………………………………

6. **Professor Dimitrios Fotiadis** ……………………………………

7. **Dr. Dimitris Kafetzopoulos** ……………………………………

This research has been co-financed by the European Union (European Social Fund – ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: Heracleitus II. Investing in knowledge society through the European Social Fund.

Η παρούσα έρευνα έχει συγχρηματοδοτηθεί από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο - ΕΚΤ) και από εθνικούς πόρους μέσω του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» του Εθνικού Στρατηγικού Πλαισίου Αναφοράς (ΕΣΠΑ) – Ερευνητικό Χρηματοδοτούμενο Έργο: Ηράκλειτος II . Επένδυση στην κοινωνία της γνώσης μέσω του Ευρωπαϊκού Κοινωνικού Ταμείου.

Δημοσιεύσεις

Δημοσιεύσεις και ανακοινώσεις σε συνέδρια που προέκυψαν κατά την εκπόνηση της διατριβής

1. Koumakis L., Potamias G., Tsiknakis M., Zervakis M. and Moustakis V. “Integrating Microarray Data and GRNs.” Methods in Molecular Biology (under review).
2. Koumakis L., Potamias G., Sfakianakis S., Moustakis V., Zervakis M., Graf N. and Tsiknakis M. “miRNA based pathway analysis tool in nephroblastoma as a proof of principle for other cancer domains.” Under review for the 14th IEEE International Conference on BioInformatics and BioEngineering (BIBE-2014).
3. Kalantzaki, K., Lefteris Koumakis, Ekaterini S. Bei, M. Zervakis, George Potamias, and Dimitris Kafetzopoulos. “Experimental model construction and validation of the ErbB signaling pathway.” In Bioinformatics and Bioengineering (BIBE), 2013 IEEE 13th International Conference on, pp. 1-4. IEEE, 2013.
4. Koumakis, L., Moustakis, V., Zervakis, M.E., Kafetzopoulos, D., & Potamias, G.A. “Coupling Regulatory Networks and Microarays: evealing Molecular Regulations of Breast Cancer Treatment Responses.” Artificial Intelligence: Theories and Applications. Lecture Notes in Computer Science, 7297, 239-246 (2012).
5. Koumakis, L., Potamias, G.A., Zervakis, M.E., & Moustakis, V.A. (2011). “Integrating microarray data and gene regulatory networks: Survey and critical considerations.” 10th International Workshop on Biomedical Engineering. Kos, Greece 5-7 October 2011.

Abstract

More than a decade after the completion of the Human Genome Project, advances in genome research and biotechnology have influenced drastically the concept of disease diagnosis and treatment. In this context, the improvement of high throughput technologies, such as microarrays, caused a fundamental transformation in the research of various diseases (e.g. cancer). Microarrays present a powerful tool to study the molecular basis of the genesis and progression of diseases, and has advanced life scientists' ability not only to detect but also to quantify simultaneously the expression of thousands of genes for various diseases and phenotypes.

Initial expectation was that microarrays would reveal specific gene co-expression patterns (gene signatures or, gene-biomarkers) for various phenotypes, but the utility of gene-expression profiles seems to be bounded by a number of limitations, mainly related to: (a) the variation and heterogeneity of the examined tissues - when comparing two different tissue samples, the potential differences in gene-expression levels is a manifestation of all the cell types present in that sample, making the induced gene-signatures amenable to the specific tissues examined; (b) the different microarray platforms utilised as well as the different experimental protocols followed are facts that make really difficult to combine gene-expression datasets form heterogeneous platforms and different studies; and (c) the great imbalance between the huge number of transcripts and genes (tens of thousands) and the relatively small number of available sample cases (hundreds). In addition, the utilization of 'knowledge-ignorant' feature-selection approaches does not guarantee the 'biological validity' of the result (selected gene-biomarkers). In other words, focusing just on highly differential genes might not be the optimal process to follow. The aforementioned observations have being reported and justified by various studies in the literature.

Currently bioinformatics community focuses on more 'knowledge-aware' and enhanced methods for selecting genes from microarray data. These methods, aim to guide the gene-selection process by taking advantage and 'amalgamating' knowledge from other established biological sources, such as molecular pathways, and especially gene regulatory networks (GRNs). In cells thousands of genes are expressed and work in concert to ensure the cell's function, fitness, and survival. The gene relationships have been mapped onto GRNs that can be interrogated to gain insight into the mechanisms of differential gene expression at a systems level. These networks can also be used to understand the flow of

information in a biological system, to identify circuits that may be used for a specific purpose, and to model changes in gene expression under different conditions. The study of the function, structure and evolution of GRNs in combination with microarray gene-expression profiles has become essential for contemporary biology research.

The most prominent research line in the respective fields, called pathway analysis, focus on the identification of the most discriminant GRNs (pathways), or parts of GRNs (sub-paths) that differentiate between specific phenotypes by integrating and coupling the underlying gene regulatory machinery of GRNs and gene-expression profiles from microarray data. The relevant approaches and methodologies increased significantly over the past years, a fact that indicates the importance of such an integration endeavour. In addition, all reported methodologies and developed tools have significantly contributed to the identification of informative associations between GRNs and target phenotypes. One critical drawback of these tools comes from the way the methodologies handle the knowledge encoded in GRNs. In most cases each GRN is represented and manipulated just as the set of the genes engaged in the network. With this approach, and following the gene enrichment analysis (GEA) algorithmic processes, one can determine which biological pathways are significantly over-represented (i.e., more than expected by chance) for a specific phenotype. So, the GEA-like methodologies, are unable to access and do not provide information for parts (i.e., sub-paths) of the pathway. Recently, some enhanced GEA-like tools, take advantage and utilize in their analysis the topology of the GRNs (based on graph-theoretic approaches and network visualization techniques) but only a limited number of the reported so-far methodologies take advantage of the signalling information present in a GRN i.e., the topology and the type of involved interactions such as the activation or inhibition relations holding between genes.

The work reported in this thesis introduces and presents a novel pathway-analysis methodology. The whole methodology is implanted in a system called MinePath (www.minepath.org), a web-based platform aiming to facilitate and ease the identification and visualization of differentially active paths or sub-paths within a GRN, using gene-expression data. The methodology takes advantage of the topology and the underlying regulatory mechanisms of GRNs, including the direction and the type of the engaged interactions (e.g. activation/expression, inhibition). Each GRN sub-path is interpreted according to Kauffman's principles and semantics: (i) the network is a directed graph with genes (inputs and outputs) being the graph nodes and the edges between them representing the causal links between them, i.e., the regulatory reactions; (ii) each node can be in one of the two states, 'ON', the gene is expressed or up-regulated (i.e., the respective substance being present) or, 'OFF', the gene is not-expressed or targeted from a specific gene; and (iii) time is viewed as proceeding in discrete steps - at each step the new state of a node is a Boolean function of the prior states of the nodes with arrows pointing towards it.

The method of MinePath unfolds into five modular steps:

I. Gene expression values are discretized into two states with values 1 and 0 for up-regulated and down-regulated genes, respectively, and the respective samples' binary gene-expression sample matrix is formed;

II. each target GRN is decomposed into its constituent sub-paths, e.g., the path A → B —| C is decomposed into three sub-paths, A → B, B —| C and A → B —| C (note that the overlapping sub-paths are also identified and formed);

III. Each sub-path is interpreted on the basis of its functional active-state, and it is represented by a binary ordered-vector with active states, e.g., sub-path A → B —| C is considered functional when A↑ and B↑ are up-regulated and C↓ is down-regulated, resulting into its active-state ordered vector <1,1,0> for the corresponding genes;

IV. The binary ordered-vector of each sub-path is aligned and matched against all (discretized) binary gene-expression sample profiles. A sub-path is considered to match a sample if and only if all the corresponding genes in the sub-path exhibit the same active-state in the sample, i.e. genes A, B are up-regulated and gene C is down-regulated, resulting into the corresponding sample ordered-vector <1,1,0>, which matches the sub-path vector. In addition, a binary sub-path expression matrix is formed with rows the sub-paths, columns the input samples, and cell-values 1, 0 for the respective sub-path being functional and active (or hold) for the corresponding sample or not. In other words, the sub-paths are taking the place of sample descriptor features and are utilized for the construction of sub-path based phenotype prediction models.

V. Finally, the differential power of each sub-path is computed and appropriate parameterized (users may adjust them to his/her exploratory needs). The highly ranked (best matching) sub-paths are kept according to user-defined thresholds. Subsequently each sub-path is characterized about its phenotype inclination; sub-paths with positive differential power values are characterized as inclined to phenotype 1, and those with negative power as phenotype 2. These sub-paths present putative evidential molecular mechanisms that govern the disease itself, its type, its state or other targeted disease phenotypes (e.g., histopathological characterization, positive or negative response to specific drug treatment). The system also identifies the sub-paths that are functional and always active in both phenotypes. The result is a binary sub-path expression matrix analogue to the gene-expression matrix where the sub-paths are taking the place of genes playing the role of sample descriptors. Then the prediction performance of the selected sub-paths is assessed and reported – the reported prediction performance follows a 10-fold cross-validation mode on machine-learning algorithms, such as C4.5 decision-tree, Naïve Bays, or support vector machines (SVMs); as all relevant sub-path expression matrices are saved and stored, the user may utilize them to build other prediction models based on his/her preferences and needs.

MinePath uses binary data structures and Boolean algebra for the calculations, a framework that makes it capable to operate in real time even on big datasets with hundreds of pathways and tens of thousands of sub-paths.

Apart from the MinePath methodology, only four other tools/methodologies take advantage of the underlying GRN gene regulation mechanisms, namely GGEA, SPIA, TEAK and PATHOME. The main difference that contrasts MinePath with these approaches resides in the handling of the gene regulatory mechanisms. To our knowledge, all aforementioned methodologies score with +1 the activations and -1 the inhibitions relations between genes, and each sub-path gets a final rank. Contrary MinePath methodology strictly checks and assess the differential power of the sub-paths that are functional and hold in one of the phenotypes (as exemplified in step IV, above).

Another limitation of the aforementioned tools is that they lack of a productive environment with efficient, interactive and user-friendly visualization operations that offers rich exploratory capabilities to the research biomedical scientists towards their quest to reveal and get insight to key phenotype regulatory mechanisms. A key innovation of MinePath, contrary to similar approaches that visualize just the state of genes in a GRN, rest in its exploratory capabilities and especially in the visualization of active gene–to–gene regulatory relations that differentiate between the target phenotypes. In addition, MinePath supports active interaction and re-adjustment of the visualized network and is equipped with special operational features enabling live interaction, immediate visualization of regulatory relations and the reduction of GRN's complexity using special topological and network-adjustment functionalities.

Furthermore, MinePath is the only tool that takes also into account and visualizes sub-paths that are fully functional and hold for both phenotypes. These sub-paths possess no differential power but they may be utilised to link the gap (functional interaction) between two sub-paths and reveal long and more complex functional routes in molecular pathways, the interpretation and validation of which is biologically more profound e.g. link the gap between extracellular gene interactions and final biological reaction such as apoptosis. This feature serves the biomedical researchers' exploratory needs to reveal and interpret the regulatory mechanisms that underlie and putatively govern the expression of target phenotypes.

MinePath methodology and the web-platform aim to effectively address all the aforementioned issues. MinePath has been thoroughly tested for its stability and the methodology was applied on gene-expression and miRNA expression data with the target of identifying mechanisms that underlie the expression of specific phenotypes (e.g. breast cancer patients according to their ER-status profiles, or Wilms' tumour prediction). The results are quite indicative and strongly supported by the relevant biomedical literature. In addition, the prediction performance of MinePath, using the selected differential sub-paths as sample descriptors, was tested and contrasted with the corresponding performance when the original gene-expression data are used – the results are quite satisfactory.

Περίληψη

Υπολογιστικές Προσεγγίσεις για την Ανακάλυψη και Παραγωγή Γνώσης από Ετερογενείς Πήγες: Μεθοδολογία και Εφαρμογή σε βάσεις Βιολογικών και Μοριακών Δεδομένων

Οι σύγχρονες κατευθύνσεις στον τομέα της υγείας και της ιατρικής θέτουν τη πρόληψη, και την εξατομικευμένη ιατρική ως κύριες προτεραιότητες. Ωστόσο αποτελεί κοινή διαπίστωση το γεγονός ότι για να κινηθούμε προς αυτή τη κατεύθυνση πρέπει να ενσωματώσουμε τη γενετική πληροφορία στη καθημερινή πρακτική των επιστημών υγείας. Καθώς εισερχόμαστε στη μεταγονιδωματική εποχή όπου η ακολουθία του ανθρώπινου γονιδιώματος έχει αποκωδικοποιηθεί εξολοκλήρου, η βιολογία διαθέτει πλέον μεθόδους όχι μόνο για την λεπτομερειακή απεικόνιση των αλληλεπιδράσεων των γονιδίων αλλά και την δυνατότητα να επεμβαίνει ώστε να μεταβάλει και να καθορίζει, σε τεχνικό επίπεδο, τη φυσιολογία του ανθρώπινου οργανισμού μέσω των κυττάρων και συνεπώς των ιστών. Για να μπορέσουμε να εκμεταλλευτούμε στο μέγιστο αυτές τις επαναστατικές τεχνολογικές εξελίξεις πρέπει πρώτα να κατανοήσουμε και να αποτυπώσουμε τους χαοτικούς δρόμους που ακολουθεί η γονιδιακή έκφραση, καθώς μια απλή γονιδιακή μετάλλαξη, ή ένας φαινομενικά ασήμαντος περιβαλλοντικός παράγοντας μπορεί να οδηγήσει σε σημαντικές παθολογικές καταστάσεις. Η ευέλικτη, λοιπόν, και αποτελεσματική διαχείριση και επεξεργασία της γονιδιωματικής πληροφορίας με σκοπό την εξατομικευμένη ιατρική είναι η νέα πρόκληση που καλούμαστε να αντιμετωπίσουμε.

Τα παραπάνω μαζί με την πρόοδο στον γενικότερο συστημικό και υπολογιστικό τρόπο που διαχειρίζονται οι ερευνητές όλα τα στοιχεία της μοριακής βιολογίας (όπως γονίδια, πρωτεΐνες, ένζυμα, μεταγραφικούς παράγοντες, μεταβολικά και κανονιστικά δίκτυα) έχουν δημιουργήσει μία νέα περιοχή έρευνας, την βιοπληροφορική. Η βιοπληροφορική είναι ο τομέας της θετικής επιστήμης ο οποίος μελετάει τη συμπεριφορά βασικών μονάδων της βιολογικής λειτουργίας μέσω υπολογιστικών μεθόδων. Σκοπός της είναι η εύρεση πρωτότυπων και η εφαρμογή ήδη υπαρχόντων αποδοτικών και ευέλικτων αλγορίθμων επεξεργασίας γενομικών δεδομένων ώστε να εξαχθεί η γνώση που 'ελλοχεύει' σε αυτά.

Η πρόοδος της βιοπληροφορικής διευρύνθηκε με την πλήρη χαρτογράφηση του ανθρώπινου γονιδιώματος και την εφεύρεση των μικροσυστοιχίων (microarrays). Οι μικροσυστοιχίες είναι συσκευές οι οποίες επιτρέπουν την ταυτόχρονη μέτρηση της έκφρασης δεκάδων χιλιάδων γονιδίων. Μέσω αυτών μπορούμε να μετρήσουμε τη ποσοτική συμμέτοχη ενός μεγάλου μέρους του γονιδιώματος

ενός οργανισμού σε κάποιο συγκεκριμένο ιστό. Ο ιστός αυτός μπορεί να είναι υγιείς, καρκινικός, υπό θεραπεία, υπό την επίδραση κάποιου φαρμάκου ή τα κύτταρά του να υποβάλλονται σε κάποια βιολογική διεργασία όπως διαίρεση ή απόπτωση. Σε πειράματα που μετέχουν διαφορετικοί τύποι ιστών μπορούμε να εντοπίσουμε και να μετρήσουμε τη διαφορική έκφραση των γονιδίων. Από την ανακάλυψη των μικροσυστοιχιών (1996) μέχρι σήμερα έχει γίνει μία τεράστια ερευνητική προσπάθεια για την βελτίωση της ακρίβειας τους, την εφαρμογή τους σε περισσότερους ιστούς κάτω από ποικίλες συνθήκες αλλά και για την ολοκλήρωση της γνώσης που παράγεται με άλλα βιολογικά ευρήματα. Αρχικά η προσδοκία ήταν ότι οι μικροσυστοιχίες θα αποκάλυπταν μοναδικά μοτίβα γονίδιων (γονιδιακές υπογραφές) για διάφορους φαινοτύπους, όμως η επαλήθευση των γονιδιακών υπογραφών είναι περιορισμένη, κυρίως λόγω της πολυπλοκότητας και των ετερογένειών που εμφανίζονται σε αυτές. Λόγω των διαφορετικών πλατφορμών που χρησιμοποιούνται στα διάφορα πειραματικά πρωτόκολλα και κυρίως σε πειράματα με μικρά μεγέθη δειγμάτων, η υψηλή διαφορική έκφραση ενός γονιδίου δεν απηχεί κατ’ ανάγκη σε μια μεγαλύτερη πιθανότητα το γονίδιο να σχετίζεται με τη νόσο και, ως εκ τούτου, εστιάζοντας μόνο στα υποψήφια γονίδια με υψηλές διαφορικές εκφράσεις μπορεί να μην είναι η βέλτιστη διαδικασία για τον διαχωρισμό ή την πρόβλεψη ετερογενών φαινοτύπων.

Στις μέρες μας η βιοπληροφορική επικεντρώνεται σε πιο ανεπτυγμένες μεθόδους για την επιλογή γονιδίων από μικροσυστοιχίες κυρίως με την προσθήκη και την επεξεργασία γνώσης από άλλες πηγές, όπως τα γονιδιακά ρυθμιστικά δίκτυα (ΓΡΔ) (Gene Regulatory Networks), τα οποία μοντελοποιούν τις αλληλεπιδράσεις των γονιδίων κατά τη διάρκεια βιολογικών διεργασιών. Στο κύτταρο εκατοντάδες ή χιλιάδες γονίδια εκφράζονται και συνεργάζονται από κοινού για να εξασφαλιστεί η λειτουργία και η επιβίωση του. Οι σχέσεις των γονιδίων έχουν χαρτογραφηθεί σε ΓΡΔ τα οποία μπορούν να προσφέρουν γνώση σχετικά με τους μηχανισμούς της γονιδιακής έκφρασης σε επίπεδο συστήματος. Αυτά τα δίκτυα μπορούν επίσης να χρησιμοποιηθούν για την κατανόηση της ροής των πληροφοριών σε ένα βιολογικό σύστημα, για τον εντοπισμό μονοπατιών που μπορούν να χρησιμοποιηθούν για συγκεκριμένο σκοπό, και να μοντελοποιήσουν αλλαγές στην έκφραση γονιδίων κάτω από διαφορετικές συνθήκες. Η μελέτη της λειτουργίας, της δομής και της εξέλιξης των ΓΡΔ σε συνδυασμό με το προφίλ γονιδιακής έκφρασης από μικροσυστοιχίες έχει γίνει απαραίτητη για τη σύγχρονη βιολογική έρευνα.

Οι περισσότερες προσπάθειες για την ολοκλήρωση της γνώσης που εμπεριέχουν οι παραπάνω πηγές (μικροσυστοιχίες και ΓΡΔ) αντιμετωπίζουν τα δίκτυα σαν μονοδιάστατες πηγές πληροφορίας όπου οι συσχετίσεις των γονιδίων, όπως αυτά μοντελοποιούνται, δεν εμπερικλείονται και συνεπώς δεν αξιοποιούνται. Πρόσφατα, όλο και περισσότερες μέθοδοι επωφελούνται από την τοπολογία των δικτύων χρησιμοποιώντας μεθόδους της θεωρίας γράφων, αλλά μόνο ένας περιορισμένος αριθμός των επί του παρόντος διαθέσιμων μεθοδολογιών, μπορεί να αξιοποιήσει τις πληροφορίες ρύθμισης εντός των ΓΡΔ όπως η αλληλεπίδραση μεταξύ γονιδίων. Η αλληλεπίδραση αυτή μπορεί να χωριστεί σε πολλές κατηγορίες, με δύο από αυτές να θεωρούνται οι πιο σημαντικές. Η πρώτη είναι η ενεργοποίηση/έκφραση (activation), όπου ένα γονίδιο ενεργοποιεί κάποιο άλλο, και η δεύτερη η αναστολή (inhibition), όπου ένα γονίδιο σταματάει την ενεργοποίηση κάποιου άλλου. Είναι χαρακτηριστικό ότι υπάρχουν γονίδια των οποίων η πρωτεΐνη που κωδικοποιούν δεν έχει κάποιο βιολογικό ρόλο πέρα από την ενεργοποίηση ή απενεργοποίηση άλλων γονιδίων. Τα γονίδια αυτά ονομάζονται μεταγραφικοί παράγοντες (transcription factors).

Η παρούσα εργασία στόχο έχει στο να συμβάλει στους σχετικά πρόσφατους τομείς της υπολογιστικής βιολογίας και της βιοπληροφορικής με την υλοποίηση μεθόδων για μοντελοποιήση της συμπεριφοράς των ΓΡΔ και εισαγωγή τρόπων εξόρυξης γνώσης από αυτά. Ο κύριος θεματικός τομέας της διατριβής είναι η υπολογιστική μοντελοποίηση των δυναμικών και συστημικών ιδιοτήτων των ΓΡΔ καθώς και η δυνατότητα εκμετάλλευσης της πληροφορίας που εμπεριέχουν σε συνδυασμό με άλλες σύγχρονες έννοιες της μοριακής βιολογίας όπως είναι η γενετική έκφραση. Ποιο συγκεκριμένα: τα μονοπάτια που εκφράζονται ή υποεκφράζονται σε έναν ιστό όπως αυτό αποτυπώνεται από πειράματα με μικροσυστοιχίες θα εντοπιστούν μέσω μεθόδων ανίχνευσης διαφορικής έκφρασης. Χρησιμοποιώντας σύγχρονες τεχνικές βελτιστοποίησης δικτύων για ανίχνευση διαφορικών μονοπατιών από ΓΡΔ αναμένουμε να απαντήσουμε σε ένα σύνολο από βιολογικά ερωτήματα όπως:

- Ποια δίκτυα ή μονοπάτια «λειτουργούν» και ποια όχι μεταξύ διαφορετικών τύπων ιστών/φαινοτύπων.
- Ποιες διαδρομές είναι αυτές που ακολουθούνται, και ποιοι παράγοντες/γονίδια ευθύνονται για διαδρομές που δεν φαίνεται να ακολουθούνται σε παθογενείς ιστούς ή ακολουθούνται με διαφορετικό τρόπο.
- Πως μπορούμε τεχνικά να επέμβουμε με σκοπό την επιτάχυνση μίας διαδρομής που παράγει κάποια επιθυμητή ένωση (π.χ. ινσουλίνης) ή την αποτροπή μίας μη επιθυμητής διαδρομής (π.χ. απόπτωση).

Η παρούσα διατριβή δημιούργησε και παρουσιάζει το MinePath (www.minepath.org), μια διαδικτυακή πλατφόρμα, που υλοποιεί μια νέα μεθοδολογία για τον προσδιορισμό και την οπτικοποίηση των διαφορικά ενεργών δικτύων ή μονοπατιών μέσα σε ένα ΓΡΔ, χρησιμοποιώντας δεδομένα γονιδιακής έκφρασης. Η πλατφόρμα εκμεταλλεύεται την τοπολογία και τους ρυθμιστικούς μηχανισμούς των ΓΡΔ, συμπεριλαμβανομένης της κατεύθυνσης και του τύπου των γονιδιακών αλληλεπιδράσεων (π.χ. ενεργοποίηση / έκφραση, αναστολή). Η μεθοδολογία εντοπίζει όλα τα λειτουργικά μονοπάτια που εμφανίζονται σε (επιλεγμένα και στοχευμένα) ΓΡΔ και εξάγει τα συμβατά με τις τιμές έκφρασης των γονιδίων των δειγμάτων που ανήκουν σε διαφορετικό κλινικό φαινότυπο (π.χ., νοσούντα εναντίον υγιούς). Η διαφορική δυναμική των επιλεγμένων μονοπατιών υπολογίζεται και η βιολογική σημασία τους αξιολογείται.

Το MinePath λειτουργεί με ΓΡΔ από τη βάση δεδομένων KEGG (Kyoto Encyclopedia of Genes and Genomes). Από την πρώτη τους εμφάνιση το 1995 τα δίκτυα της KEGG έχουν χρησιμοποιηθεί ευρέως ως βάση γνώσεων αναφοράς για την κατανόηση των βιολογικών μονοπατιών και την λειτουργία των κυτταρικών διαδικασιών. Κάθε ΓΡΔ περιγράφεται ως γράφημα, όπου οι κόμβοι αντιπροσωπεύουν γονίδια, ομάδες γονιδίων, ενώσεων ή άλλων δικτύων και οι ακμές αντιπροσωπεύουν γνωστές βιολογικές αλληλεπιδράσεις γονιδίων όπως ενεργοποίηση, αναστολή, έκφραση, φωσφορυλίωση, ένωση, διάσπαση κλπ. Η επεξεργασία των ΓΡΔ στο MinePath λαμβάνει υπόψη όλες τις πιθανές λειτουργικές αλληλεπιδράσεις του δικτύου. Διαφορετικές αλληλεπιδράσεις αντιστοιχούν σε διαφορετικά λειτουργικά μονοπάτια που μπορεί να ακολουθούνται για την ρύθμιση ενός γονιδίου.

Κάθε μονοπάτι από τα ΓΡΔ ερμηνεύεται σύμφωνα με τις αρχές και τη σημασιολογία του Kauffman όπου: (i) το δίκτυο είναι ένας κατευθυνόμενος γράφος με κόμβους (γονίδια) και οι ακμές μεταξύ αυτών εκπροσωπούν τις αλληλεπιδράσεις μεταξύ τους, δηλαδή τις ρυθμιστικές αντιδράσεις (ii) κάθε κόμβος μπορεί να αναπαρίσταται με μία από τις δύο καταστάσεις, «ON», το γονίδιο εκφράζεται (δηλαδή το γονίδιο είναι ενεργό), ή «OFF», το γονίδιο δεν εκφράζεται, ή αναστέλλεται από ένα άλλο γονίδιο και (iii) ο χρόνος θεωρείται ως διαδικασία σε διακριτά βήματα - σε κάθε βήμα η νέα κατάσταση ενός κόμβου είναι μια δυαδική λειτουργία των πρότερων καταστάσεων των γονιδίων με ακμές που δείχνουν προς την κατεύθυνση αυτή.

Η μεθοδολογία του MinePath μοντελοποιείται σε πέντε διακριτά βήματα:

I. Οι τιμές έκφρασης των γονιδίων από τις μικροσυστοιχίες διακριτοποιούνται σε τιμές 1 και 0 για τα εκφρασμένα και υπο-εκφρασμένα γονίδια αντίστοιχα, και σχηματίζεται μια δυαδική μήτρα γονιδίων και φαινοτύπων

II. Κάθε ΓΡΔ αναλύεται σε όλα τα δυνατά μονοπάτια; για παράδειγμα το μονοπάτι A → B —| C αναλύεται σε τρία μονοπάτια, τα A → B, B —| C και A → B —| C

III. Κάθε μονοπάτι χαρακτηρίζεται από την λειτουργική ενεργή κατάσταση του με τη χρήση δυαδικού διανύσματος. Για παράδειγμα το μονοπάτι A → B —| C θεωρείται ενεργό όταν A↑ και B↑ (εκφρασμένα γονίδια) και C↓ (υπο-εκφρασμένο γονίδιο), που μας δίνει το δυαδικό διάνυσμα <1,1,0> για το μονοπάτι A → B —| C

IV. Τα δυαδικά διανύσματα για όλα τα μονοπάτια αντιπαραβάλλονται με την δυαδική έκφραση των γονιδίων από τις μικροσυστοιχίες για κάθε δείγμα. Ένα μονοπάτι θεωρείται ότι είναι ενεργό σε ένα δείγμα, αν και μόνο αν όλα τα αντίστοιχα γονίδια στο μονοπάτι έχουν την ίδια ενεργό κατάσταση στο δείγμα, δηλαδή, τα γονίδια A, B είναι εκφρασμένα και το γονίδιο C υπο-εκφρασμένο, που αντιστοιχεί στο διάνυσμα <1,1,0> για τα γονίδια <A,B,C> στο δείγμα. Επιπλέον, μια δυαδική μήτρα σχηματίζεται με τις σειρές να αναπαριστούν μονοπάτια, τις στήλες τα δείγματα, και οι τιμές των κελιών να είναι δυαδικές (1, 0) όπου 1 όταν το αντίστοιχο μονοπάτι είναι ενεργό για το αντίστοιχο δείγμα ή 0 αν δεν είναι. Με άλλα λόγια, τα μονοπάτια παίρνουν τη θέση χαρακτηριστικών του δείγματος και χρησιμοποιούνται για την κατασκευή μοντέλων πρόβλεψης φαινοτύπων.

V. Στο τελικό βήμα, η διαφορική δυναμική κάθε μονοπατιού υπολογίζεται χρησιμοποιώντας ειδικά διαμορφωμένες φόρμουλες. Τα μονοπάτια με τη μέγιστη διαφορική δυναμική και πάνω από ένα όριο θεωρούνται τα μονοπάτια που μπορούν να διαχωρίσουν τους δύο φαινοτύπους. Επιπρόσθετα, τα μονοπάτια με θετική διαφορική δυναμική χαρακτηρίζουν τον ένα φαινότυπο (π.χ. ασθενής) ενώ τα μονοπάτια με αρνητική διαφορική δυναμική χαρακτηρίζουν τον δεύτερο φαινότυπο (π.χ υγιής). Το αποτέλεσμα είναι ένας πίνακας μονοπατιών με δυαδικές τιμές για κάθε δείγμα. Στη συνέχεια υπολογίζουμε την ικανότητα πρόβλεψης των επιλεγμένων μονοπατιών χρησιμοποιώντας την τεχνική αξιολόγησης 10 fold cross validation σε αλγόριθμους μηχανικής μάθησης, όπως C4.5 δέντρο αποφάσεων, naïve Bays, ή support vector machine. Το σύστημα επίσης αναγνωρίζει και εξάγει και τα μονοπάτια που είναι πάντα ενεργά (και για τους δύο φαινοτύπους) χωρίς να τα λαμβάνει υπόψιν του στα μοντέλα πρόβλεψης.

Το MinePath χρησιμοποιεί δυαδικές δομές δεδομένων και άλγεβρα Μπουλ για τους υπολογισμούς, καθιστώντας το ικανό να αναλύσει σε πραγματικό χρόνο δεδομένα από μεγάλες κλινικές δοκιμές (με μικροσυστοιχίες) σε συνδυασμό με εκατοντάδες ΓΡΔ και δεκάδες χιλιάδες μονοπάτια. Η μεθοδολογία αυτή αναδεικνύει τα ενεργά και μη ενεργά μονοπάτια σε ΓΡΔ ανα φαινότυπο. Αυτά τα μονοπάτια αναδεικνύουν μοριακούς μηχανισμούς που διέπουν την ίδια την ασθένεια, τον τύπο, την κατάσταση ή άλλους εστιασμένους φαινοτύπους όπως απόκριση ή μη σε ειδικές θεραπείες.

Εκτός από την προτεινόμενη μεθοδολογία, μόνο τέσσερα άλλα εργαλεία / μέθοδοι εκμεταλλεύονται τους μηχανισμούς γονιδιακής ρύθμισης στα ΓΡΔ, τα GGEA, SPIA, TEAK και PATHOME. Η κύρια διαφορά της προτεινόμενης μεθοδολογίας από αυτά τα τέσσερα συστήματα είναι ο χειρισμός των γονιδιακών ρυθμιστικών μηχανισμών. Όλες οι άλλες μεθοδολογίες μετράνε με +1 τις ενεργοποιήσεις και -1 τις αναστολές. Κάθε μονοπάτι παίρνει ένα τελικό αποτέλεσμα το οποίο χρησιμοποιείται επίσης ως μια φόρμουλα κατάταξης. Αντίθετα, η προσέγγιση μας ελέγχει και λαμβάνει υπόψη μόνο μονοπάτια που είναι πλήρως λειτουργικά (σύμφωνα με τις σχέσεις των γονιδίων και τις εκφράσεις τους).

Ένας άλλος βασικός περιορισμός με τη χρήση αυτών των μεθόδων είναι η έλλειψη ενός παραγωγικού περιβάλλοντος με αποτελεσματικό, δια-δραστικό και φιλικό προς το χρήστη τρόπο απεικόνισης που να προσφέρει διερευνητικές ικανότητες για την κατανόηση των ρυθμιστικών μηχανισμών των φαινοτύπων. Σε αντίθεση με παρόμοιες προσπάθειες, οι οποίες απεικονίζουν την κατάσταση των γονιδίων σε ένα ΓΡΔ, μια βασική καινοτομία της πλατφόρμας MinePath έγκειται στις δυνατότητες απεικόνισης και ειδικά, στην οπτικοποίηση των ενεργών γονιδιακών ρυθμιστικών σχέσεων που διαφοροποιούν τους υπό μελέτη φαινοτύπους. Το MinePath υποστηρίζει ενεργή αλληλεπίδραση με τα οπτικοποιημένα δίκτυα όπως η εκ νέου ρύθμιση της τοπολογίας τους και είναι εξοπλισμένο με ειδικά λειτουργικά χαρακτηριστικά που επιτρέπουν άμεση αλληλεπίδραση, άμεση απεικόνιση των ρυθμιστικών σχέσεων και τη μείωση της πολυπλοκότητας των ΓΡΔ χρησιμοποιώντας ειδικές λειτουργίες τοπολογίας.

Επιπρόσθετα, η προτεινόμενη μεθοδολογία είναι η μόνη που λαμβάνει υπόψη και οπτικοποιεί μονοπάτια πλήρως λειτουργικά και για τους δύο φαινοτύπους. Αυτά τα μονοπάτια δεν έχουν καμία διακριτική αξία αλλά μέσα σε ένα ΓΡΔ τα μονοπάτια που είναι πάντα ενεργοποιημένο μπορεί να συνδέσουν το κενό (λειτουργική αλληλεπίδραση) μεταξύ δύο μονοπατιών και να αποκαλύψουν ένα πλήρες λειτουργικό μονοπάτι που είναι βιολογικά πολύτιμο όπως για παράδειγμα η σύνδεση του χάσματος μεταξύ λειτουργικών εξω-κυτταρικών γονιδίων και ενός τελικού μηχανισμού κυτταρικής λειτουργίας (απόπτωση, νέκρωση, πολλαπλασιασμός, κτλ).

Η μεθοδολογία του MinePath και η διαδικτυακή της υλοποίηση έχει ως στόχο την αποτελεσματική αντιμετώπιση αυτών των ζητημάτων. Η μεθοδολογία εφαρμόστηκε σε μικροσυστοιχίες γονιδίων και miRNAs με στόχο την ανάδειξη πιθανών μηχανισμών που διέπουν και ρυθμίζουν την ανταπόκριση σε θεραπεία συγκεκριμένων φαινοτύπων (π.χ. ασθενείς με καρκίνο του μαστού, σύμφωνα με το προφίλ τους σε υποδοχείς οιστρογόνων, ή την πρόβλεψη της ασθένειας Wilms' tumor). Τα αποτελέσματα είναι αρκετά ενθαρρυντικά και υποστηρίζονται από τη σχετική βιοϊατρική βιβλιογραφία. Οπλισμένο με τα παραπάνω χαρακτηριστικά, το MinePath εξυπηρετεί διερευνητικές ανάγκες ερευνητών για την ανακάλυψη ρυθμιστικών μηχανισμών που αποτελούν τη βάση και ορίζουν την έκφραση συγκεκριμένων φαινοτύπων.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Key word
BRCA	Breast Cancer mutation in either of the genes BRCA1 and BRCA2
C4.5	Decision tree learning developed by Ross Quinlan
DNA	Deoxyribonucleic acid
ER	Estrogen Receptor
GEO	Gene Expression Omnibus
GRNs	Gene Regulatory Networks
GSA	Gene Set Analysis
GUI	Graphical User Interface
KEGG	Kyoto Encyclopaedia of Genes and Genomes
LOOCV	Leave One Out Cross Validation
miRNA	microRNA
MTI	miRNA–target interactions
mRNA	Messenger Ribonucleic acid
NB	Naïve Bays
NLP	Natural Language Processing
ORA	Over Representation Analysis
RNA	Ribonucleic acid
SMO	Sequential Minimal Optimization SVM
StD	Standard Deviation
SVM	Support Vector Machines

1. Introduction

More than a decade after the completion of the Human Genome Project[1], advances in genome research and biotechnology (*omics*-comprehensive analysis platforms) have influenced drastically the concept of disease diagnosis and treatment. Genome sequencing identified approximately 22.000 genes in human Deoxyribonucleic acid (DNA) and determined the sequence of the about 3,2 billion chemical base pairs that make up human DNA. To overcome complexity, scientists developed tools and techniques to map and handle the massive volumes of data.

The two of the most important and significant genomic data sources come from microarray gene-expression experiments and respective databanks and from molecular pathways and gene regulatory networks (GRNs) stored and curated in public as well as in commercial repositories. The association of these two sources aims to give new insight in disease understanding and reveal new molecular targets in the treatment of specific phenotypes.

1.1. Microarrays

A DNA microarray (also commonly known as DNA chip or biochip) is a collection of microscopic DNA spots attached to a solid surface. Scientists use DNA microarrays to measure expression levels of large numbers of genes simultaneously or to genotype multiple regions of a genome. DNA microarray is a widely used tool to analyse genome-wide messenger ribonucleic acid (mRNA) expression levels within a particular sample.

Most common type of microarrays is the two colour, which measures tens of thousands of expressions on a single chip and use two colours to differentiate [1]. Applications of microarrays include measuring gene expression in different developmental stages, identifying biomarkers for particular phenotypes or diseases and monitoring treatment response. The process of expression data analysis encompasses three major categories:

I. The first one is “class comparison”, in which expression levels from two or more different types of samples are compared in order to identify differentially expressed genes between these classes. Most such experiments are of the case/control type and try to identify the genes that contribute to a particular phenotype, for example breast cancer tissue versus normal

[1] http://www.genome.gov/10001772 (last day visited 11/08/2014)

tissue [2]. Other experiments may focus on the differences in downstream gene expression following a gene deactivation due to a mutation, or artificial gene silencing methods, in order to gain insight into the function of that particular gene [3].

II. The second one is "class discovery", which can be applied to a collection of samples that share a common phenotype. Clustering techniques such as hierarchical clustering or k-means are used to generate molecular subgroups that share common features and can be used as diagnostic classifiers [4]. A well-known example is the classification of breast cancer into distinct phenotypes [5].

III. The last category is "class prediction". Two or more predefined classes of samples are needed in order to construct the classifier using their expression profiles. Unknown samples can then be matched to one of the classes, by comparing their expression profile to the profiles of the known ones. Common class assignment techniques are nearest neighbour algorithms, support vector machines and decision trees. Such an example is the prediction of the existence of BRCA1 and BRCA2 mutations in breast cancer samples [6].

A limitation of microarrays is that most of the datasets contain noisy data or various types of systematic errors [7]. Another limitation relates to the learning deficiencies of inference algorithms where we have (i) the '*curse of dimensionality*'- the number of features characterizing these data is in the thousands or tens of thousands and (ii) the' *curse of sparse dataset*'- the number of samples is limited [8]. Nevertheless, a lot of experiments and algorithms have been published trying to identify the most promising group of genes for specific phenotypes.

1.2. Gene Regulatory Networks

System biology is an area that studies the interactions between the components of biological systems and the behaviour of the systems into specific functions. It provides a global view of the dynamic interactions in a biological system. On the molecular level the purpose of systems biology is to ascertain the interactions and dynamic behaviour of molecules within a cell. The molecular mechanisms determine how cells interact and how they develop and maintain higher levels of organization and function. Systems biology tries to formulate these mechanisms in mathematical models.

Biological pathways represent complex reactions at the molecular level in living cells. Based on the overall effect they have on the functioning of an organism, pathways may be divided into several different categories. Three main categories are:

- metabolic pathways
- gene regulatory networks/pathways
- signal transduction pathways

Current study focuses on the gene regulatory networks but can be extended to other pathway categories too.

A GRN is a collection of DNA segments in a cell that interact with each other (indirectly through their RNA and protein expression products) and with other substances in the cell, thereby governing the rates that genes in the network are transcribed into mRNA.

Typically GRNs are represented as graphs, consisting of nodes and edges. The network by itself acts as a mechanism that determines cellular behaviour where the nodes are genes and edges are functions that represent the molecular reactions between the nodes. Each gene is represented by a node in a directed graph. Each node (gene) can have two states: on or off where on corresponds to a gene been expressed and off corresponds to a gene not expressed. An edge in a pathway usually represents a relationship or some form of interaction between the nodes. The interaction could be of many types such as activation, inhibition, catalysis, binds to, co-cited. An indicative example of the “pathways in cancer” GRN from the Kyoto Encyclopaedia of Genes and Genomes (KEGG) database is shown in Figure 1. More details regarding the KEGG pathways notations can be found in the

Appendix I (KEGG pathways).

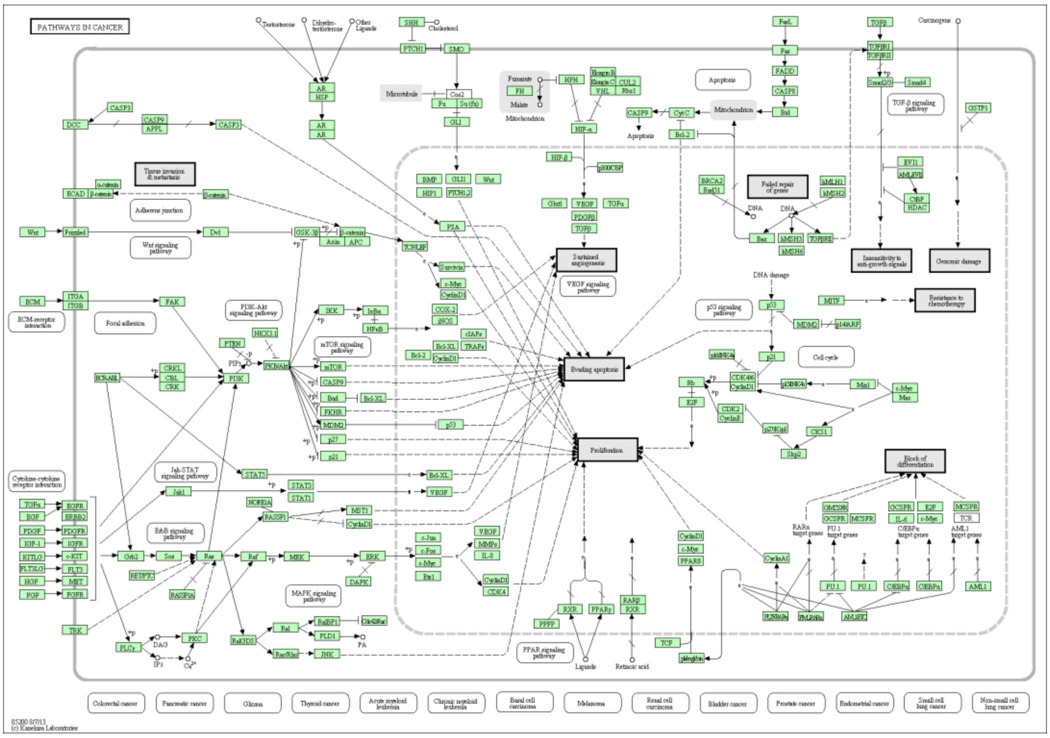


Figure 1: Pathways in Cancer GRN from KEGG

1.3. Integrating microarrays and gene regulatory networks

In recent years, high throughput data capture technologies such as microarray experiments have vastly improved life scientists' ability to detect and quantify gene, protein and metabolite expression. Furthermore, systems biology studies the behaviour of biological components such as molecules, cells, organisms or entire species. The primary aim of systems biology is to use and discover a computational model with genes, proteins and cells interacting with each other and reproducing the organism's function. GRNs are part of systems biology dealing with the modelling of genes interactions in a cell. These models have been developed to capture the GRNs in a mathematical way. Most of the gene regulatory networks are based on laboratory experimental observations, which make the generation and validation of such networks a very difficult and time-consuming task.

An important requirement for the biologists is the need to associate microarray data with gene regulatory networks diagrams to get the most biologically relevant insights from the data. Using GRNs information in microarray data analysis, scientists aim to extract more accurate and meaningful results. In a general setting, given a certain network or part of it (a sub-network), a particular gene-selection processes could focus just on the genes participating in the network, or the network participating genes could be as-signed prioritized.

On the other side, systems biology community took advantage of the human genome and the microarray technology to reconstruct and validate gene regulatory networks in an automatic way. Strong associations of genes in microarray data could be candidates for gene to gene interactions in a regulatory network.

Another area that combines GRNs and microarray data, tries to identify the most discriminant GRNs for specific phenotypes. The phenotype information is extracted from microarrays and the evaluation of the most discriminant GRNs is based on the value of each gene in the GRNs as it is expressed in microarray data. Figure 2 captures the main areas that combine microarray data and GRN knowledge, i.e., their topology and the gene to gene underlying interactions.

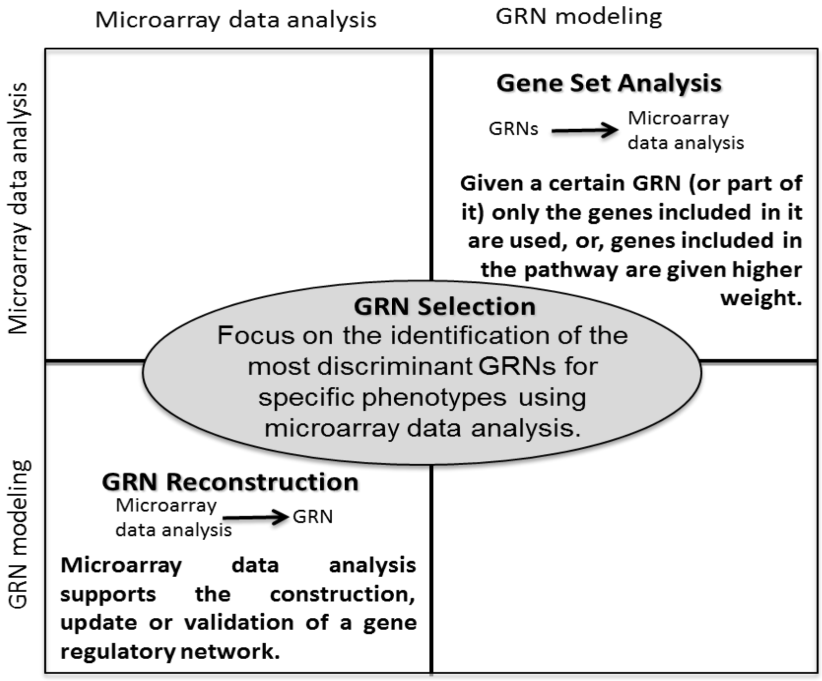
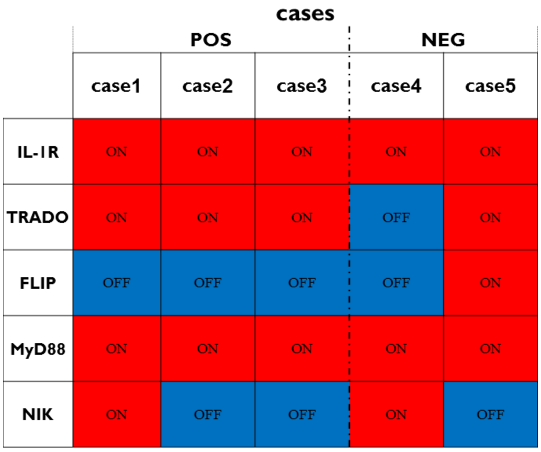

Microarray data analysis
GRN modeling
Microarray data analysis
GRN modeling
Gene Set Analysis
GRNs
Microarray data analysis
Given a certain GRN (or part of it) only the genes included in it are used, or, genes included in the pathway are given higher weight.
GRN Selection
Focus on the identification of the most discriminant GRNs for specific phenotypes using microarray data analysis.
GRN Reconstruction
Microarray data analysis
GRN
Microarray data analysis supports the construction, update or validation of a gene regulatory network.

Figure 2: Scientific areas that combine GRNs and microarrays.

1.4. Problem definition

Microarray technology has advanced life scientists' ability not only to detect but also to quantify gene-expressions for targeted phenotypes. Initial expectation was that microarrays would reveal specific gene co-expression patterns (gene signatures) for various phenotypes, but the utility of gene-expression profiles seems to be bounded to a number of limitations, mainly because of the complexity and the individual variations and heterogeneities associated with the induced gene-signatures [9], [10].

Figure 3 provides an artificial but indicative example, of the limitations in analysing solely gene-expressions data. Sample cases 1, 2 and 3 are assigned to the 'POS' class and samples case4, 5 to the 'NEG' class. At first sight we may observe that no sole gene or no group of genes can discriminate between the two classes (‘POS’ and ‘NEG‘). Inducing an un-pruned decision-tree could prove this; all the tree-branches conclude to multi-class assignments.



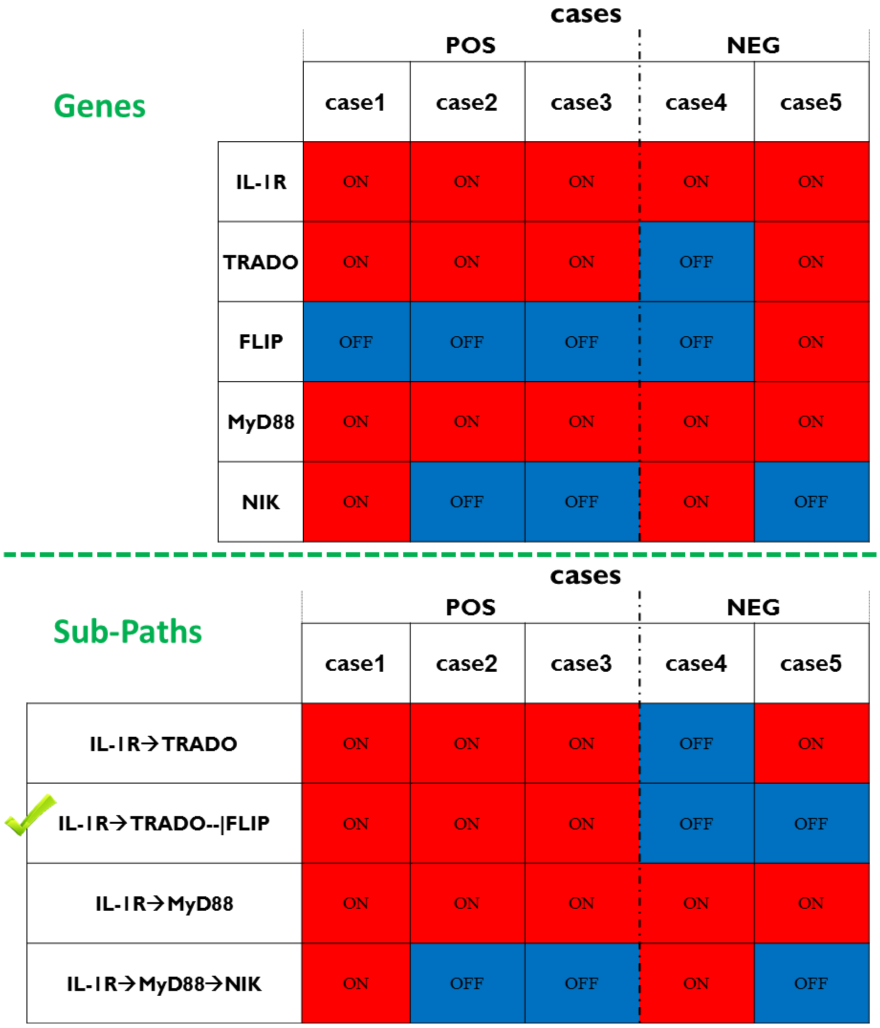
	cases				
	POS			NEG	
	case1	case2	case3	case4	case5
IL-1R	ON	ON	ON	ON	ON
TRADO	ON	ON	ON	OFF	ON
FLIP	OFF	OFF	OFF	OFF	ON
MyD88	ON	ON	ON	ON	ON
NIK	ON	OFF	OFF	ON	OFF

Figure 3: Gene expression data example. Rows represent genes, columns cases in two categories (POS and NEG), ON represents up-regulated gene for the specific case and OFF down-regulated gene.

Since the initial expectations have been limited, bioinformatics and systems biology research communities focus on more enhanced methods that utilize knowledge from known and established molecular pathways, especially in the form of gene regulatory networks and try to combine and couple such knowledge with gene-expression data.

A performance evaluation of such methods concluded that GRNs encompass additional biological features, such as the network’s topology and the underlying gene to gene interactions and may efficiently address statistical barriers in gene selection [11]. In particular, gene interaction knowledge solves the major problem of conflicting constrains when two significantly up-regulated genes increase the enrichment of the gene-set in expression data, even if the first gene inhibits the other in a GRN.

Figure 4 highlights the paradigm shift from the mining of differential genes to the mining of GRN functional sub-paths. Using the previous example we match our samples against known sub-paths of GRNs. The same gene expression example is shown in the upper part and the calculation based on sub-paths is shown on the bottom of the figure. The first sub-path (IL-1R → TRADD) satisfies cases 1,2,3,5. Second sub-path (IL-1R → TRADD **--|** FLIP) satisfies the cases case1, case2, case3 only. Third sub-path satisfies all samples and the forth sub-path does not satisfy any case. The √ symbol indicates that the second sub-path (Sub-path2) yields the maximum differential power and it contains a potential function differentiation since it is consisted only with samples that belong to the 'POS' class. In the figure, '→' represents an activation (if source gene is "ON" then the target gene is "ON" too) and '—|'an inhibition (target gene has the opposite expression of the source gene e.g. "ON"—| "OFF" or "OFF"—| "ON"). Furthermore, the regulatory fingerprint reflected by this sub-path could be considered to cause and in a way to 'govern' the specific expression status of the genes.



Genes	cases				
	POS			NEG	
	case1	case2	case3	case4	case5
IL-1R	ON	ON	ON	ON	ON
TRADO	ON	ON	ON	OFF	ON
FLIP	OFF	OFF	OFF	OFF	ON
MyD88	ON	ON	ON	ON	ON
NIK	ON	OFF	OFF	ON	OFF

Sub-Paths	cases				
	POS			NEG	
	case1	case2	case3	case4	case5
IL-1R→TRADO	ON	ON	ON	OFF	ON
✓ IL-1R→TRADO--|FLIP	ON	ON	ON	OFF	OFF
IL-1R→MyD88	ON	ON	ON	ON	ON
IL-1R→MyD88→NIK	ON	OFF	OFF	ON	OFF

Figure 4: Matching functional sub-paths and gene-expression profiles. Upper part the example from figure 3 and on the bottom the shift from genes to sub-paths and the expressions in the specific cases

Barabási et al in their review [12] stated that "*Given the functional interdependencies between the molecular components in a human cell, a disease is rarely a consequence of an abnormality in a single gene, but reflects the perturbations of the complex intracellular and intercellular network that links tissue and organ systems*". The authors concluded that there is progress towards a reliable network-based approach to disease but is currently limited by the incompleteness of the available interactome map (the whole set of molecular interactions in a particular cell) identifying also the limitations of the existing methodologies and tools to explore the role of networks in the molecular understanding of the disease.

GRNs knowledge, as it relates to specific phenotype, necessarily implies that a key molecular target should be considered within the framework of its network. A network focus enables us to more effectively infer key transcriptional changes related to the specific phenotype by examining multiple downstream (or cross-talk) effectors of the target [13]. The current gene set analysis (GSA) tools utilize mainly over-representation analysis (ORA), which reports the enrichment of functional groups (for example, gene sets) for the genes of interest. Such tools compromise the connectivity in favour of computational simplicity that is based on cellular components and not their connectivity (topology and the type of interactions) [14]. Most pathway analysis tools use the expression changes measured in high-throughput experiments only to identify pathways with unexpectedly high number of differentially expressed genes using ORA approaches or pathways whose genes are clustered in the ranked list of differentially expressed genes, but not to directly estimate the impact of such changes on specific pathways [15]. So, ORA techniques cannot distinguish cases that a subset of genes is differentially expressed just above the detection threshold from cases that the same genes are changing by many orders of magnitude.

Furthermore, probably the most important current limitation is that the knowledge embedded in GRNs concerning the genes interactions is largely unexploited. The very purpose of the pathway diagrams is to capture our current knowledge of how genes interact and regulate each other on various pathways. However, the existing analysis approaches consider only the sets of genes involved on these pathways, without taking into consideration their topology [15]. And last but not least, some genes have multiple functions and are involved in several pathways but with different roles.

1.5. MinePath approach

Our methodology, called MinePath, relies on a novel GRN processing approach that takes into account all possible functional interactions present in the network. We are inspired and guide our approach by a statement made by Geistlinger et al [16], namely: "*As the sign of gene expression changes and the direction of regulatory interactions are so far not taken into account, substantial features of the data are still ignored and the dynamics of the transcriptomic system are not realistically reflected. Activation and inhibition are essential regulatory mechanisms in the transcriptional machinery of the cell and are causes for up- and* *down-regulation of particular genes.*” In this setting, gene-expression profiles and their phenotype assignments are extracted form microarray data and all sub-paths of the GRNs are assessed and evaluated for their differential ability to discriminate between the target phenotypes, and the selection of the most informative ones.

Having in our disposal the sub-paths resulted from the functional decomposition of the GRN and the gene expression data we can precede to the identification of the sub-paths that are functionally differential. By functionally differential we define the sub-paths that are functional in one phenotype and non-functional in the other. Our purpose is to locate those paths that exhibit a high differential ability and power to discriminate between the phenotypes assigned to the sample cases of a microarray experiment.

MinePath takes advantage of interactions between genes (e.g. activation, inhibition, association etc.). A sketch outline of our approach goes as follows (as shown in Figure 5): initially we locate all functional sub-paths encoded in GRNs and we try to assess which of them are compatible with the expression status of the genes for the input samples that belong to different phenotypes (clinical/histopathological categories, diseases, prognostic states etc.); then the differential power of the selected sub-paths is computed and their biological relevance is assessed. The whole approach is applied on a set of microarray studies with the target of revealing putative regulatory mechanisms that govern the treatment responses of specific phenotypes.

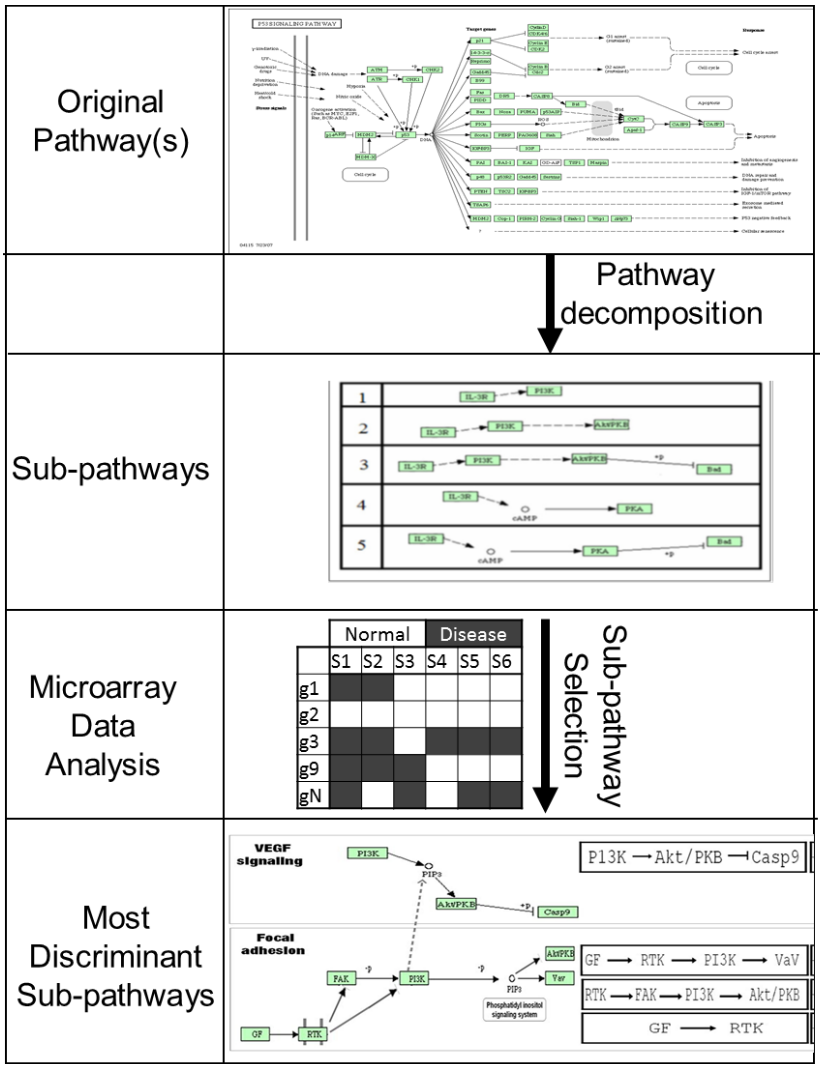


Figure 5: MinePath abstract flow of operations. From top to bottom, we start with pathways decomposition into sub-paths, we enrich with microarray expression data and we identify the most discriminant sub-paths.

In other words, the quest is for the sub-paths that exhibit high matching scores for one of the phenotypic class and low matching scores for the other. This is a paradigm shift from the mining of differential genes to the mining of GRN functional sub-paths. We applied our coupled GRN and gene-expression data analysis methodology on a set of microarray studies with the target of revealing putative regulatory mechanisms that govern the targeted phenotypes.

1.5.1. Contribution beyond the state of the art

MinePath (www.minepath.org) is a web-based platform that implements a novel methodology for the identification and visualization of differentially active paths or sub-paths within a GRN, by coupling and analysing gene-expression data in the light of the regulatory machinery reflected in the network. The platform takes advantage of the topology and the regulatory mechanisms of GRNs, including the direction and the type of the involved interactions. The methodology initially locates all functional sub-paths encoded in selected and targeted GRNs and tries to identify which of them are compatible with the expression status of genes in the given sample cases assessing at the same time the differential ability of these sub-paths to discriminate between the cases' phenotypes.

Apart from the proposed methodology, only a limited number of tools take advantage of the underlying GRN gene regulation mechanisms. The main difference of MinePath from these methodologies is the handling of the gene regulatory mechanisms. In general, all relevant existing systems and tools follows a scoring methodology in which each gene to gene network relation is scored according to its status in the gene-expression data, with activations to receive a '+1' and inhibitions a '-1' score depending if they hold in the gene-expression data. A final score is calculated and the sub-paths are accordingly ranked. In the contrary, the MinePath approach strictly checks and takes into account only sub-paths that are functional according to the gene relations and the expression values status in the given sample cases. Each sub-path is considered as functional according to its structure and type of the interactions it involves. For example, the simple activation relation A → B between two hypothetical genes A and B is considered as functional only and only if gene A is up-regulated ('ON') and gene B is down-regulated ('OFF'). More complex patterns of gene expression statuses could be formed for more complex paths, i.e., the sub-path A → B —| C is considered as functional only and only if genes A and B are in the 'ON' status and gene C in the 'OFF' status. That is, as gene B is up-regulated and the relation states that it inhibits gene C, then for the inhibition relation to holds, gene C should be 'OFF'. Then, the samples are scanned to check and count the number of samples in which the gene A is 'ON' and gene B is 'OFF'. Finally, a class-inclination formula is applied to assess if the relation, or the whole sub-path, holds mostly (even exclusively) for one class (phenotype) or the other.

MinePath uses binary data structures and Boolean algebra for the calculations, so that it is capable of operating in real time even on large datasets with hundreds of pathways and tens of thousands of sub-paths. This approach is quite innovative, and according to our knowledge no other similar system applies it. We consider functional sub-paths to present evidential molecular mechanisms that govern the phenotype itself, and in a way uncover putative regulatory fingerprints for it.

Furthermore existing differentially expressed pathway analysis systems suffer from insufficient visualization features, a fact that does not facilitate inspection of results and limit the users' exploratory potentials. Some systems utilize pathway visualization approaches to overcome this problem but since these are based on a gene-oriented approach, are unable to handle differentially expressed pathways or even differentially expressed sub-paths. Such methodologies visualize just the pathway genes using some colour scale or colour-coding schema and neglect the gene interactions. This problem is apparent even for small pathways. For example, the inhibition relation A —| B (A inhibits B; A, B represent genes) could be considered as active in two cases: when A↑ and B↓ (*up-regulation of A inhibits B and makes it down-regulated*), and when A↓ and B↑ (*down-regulation of A leaves B unaffected and/ or turns it up-regulated*). For such different cases, different colors should be assigned to the genes. The situation becomes even more complicated when one has to visualize the phenotype inclination of an interaction, e.g., an inhibition being active for one phenotype and not for another. MinePath overcomes the aforementioned problems offering an effective Web-based platform for the identification and visualization of differentially active GRN sub-paths in real time. MinePath supports live interaction, immediate visualization of regulatory relations and it is equipped with special topological and network-adjustment functionalities. To the best of our knowledge, MinePath is the only tool that visualizes differentially expressed relations instead of just differential genes.

Furthermore the MinePath methodology is the only one that takes also into account and visualizes sub-paths that are functional in both phenotypes. Even if such sub-paths possess no discriminant power their presence can link the gap (functional interaction) between two different sub-paths and reveal a complete functional route that is biologically valuable (e.g. link the gap between extracellular gene interactions and the final result of the pathway such as apoptosis). This feature serves the users' exploratory needs to reveal the regulatory mechanisms that underlie and putatively govern the expression of specific phenotypes.

More details and examples can be found in the Methodology chapter.

1.6. Dissertation structure

The dissertation is organized as follows:

- ***Chapter 2:*** A literature review of existing methods and tools supporting the integration of gene expressions and gene regulatory networks is provided. We start with gene regulatory network reconstruction methods focusing on the reconstruction using microarray experiments. Then we report methods and tools for gene expression data analysis based on gene regulatory network knowledge (Gene Set Analysis section) and we review algorithms and tools for gene regulatory networks selection using microarray data (Discriminant pathways and sub-pathways section).

- ***Chapter 3:*** A detailed description of the proposed methodology and the overview of the MinePath approach are provided. All the learning techniques and the tools utilised and appropriately customised for this work are introduced. The goal of MinePath is to identify a set of sub-paths that differentiate two experimental groups (for example, healthy vs diseased) by considering both prior knowledge about gene regulations and experimental gene expression data.
- ***Chapter 4:*** A discussion of the experiments, including testing and evaluation, is presented along with results that clearly highlight the effectiveness of the MinePath approach towards molecular mechanisms identification. The evaluation scenarios and their implementation on experimental data are described and a discussion of the results is reported.
- ***Chapter 5:*** Summarizes the conclusions of this work. Future work is discussed as well as the contributions made by this work and the scientific publications that have resulted out of it.

2. Literature

In this chapter we survey existing methods that support the different types of gene-expression and GRN integration with a focus on methodologies that aim to identify phenotype-discriminant GRNs or sub-networks. We present all the related tools and algorithms in a unified way, using standardized notations in order to reveal their technical details and to highlight their common characteristics as well as their particularities. Extensive literature search and analysis led us to the conclusion that relevant methodologies increased significantly over the past years, a fact that indicates the importance of such an integration endeavour. In addition, all reported methodologies have significantly contributed to the identification of informative associations between GRNs and target phenotypes.

Currently bioinformatics community focuses on more enhanced methods for gene selection on microarrays mainly by adding and amalgamating knowledge from other sources, such as GRNs. Integrating GRN information into the class comparison, discovery and prediction process is an important issue in bioinformatics, mainly because the provided information possesses a true biological content. By changing the focus from individual genes to a set of genes or pathways, the gene set analysis (GSA) approach enables the understanding of cellular processes as an intricate network of functionally related components. A performance evaluation of GSA methodologies [11] concluded that the inclusion of additional biological features such as topology or covariates would be more useful than simple gene selection approaches. In addition, utilizing more domain knowledge is likely to reveal more insights in the analysis.

Similarly to bioinformatics, systems biology community took advantage of the human genome and the microarray technology to reconstruct and validate gene regulatory networks in an automatic way. GRN reconstruction or reverse engineering aims toward the inference GRN models from data (in most of the cases from gene expression data). In the literature a large number of computational methods are reported with the target of inferring gene regulatory networks from expression data [17].

A special focus of the review reported here concerns a relatively new line of research in the field: the identification of the most discriminant GRNs, or parts of GRNs (i.e., sub-networks) that differentiate between specific phenotypes by coupling GRNs and microarray data. Assessment of the discriminant power of (sub)-networks is based on the identification of those genes whose expression values are consistent, i.e., could be justified, by their corresponding interaction pattern in the target GRN. Figure 6 captures and illustrates the main research areas that combine microarray data and GRN knowledge.

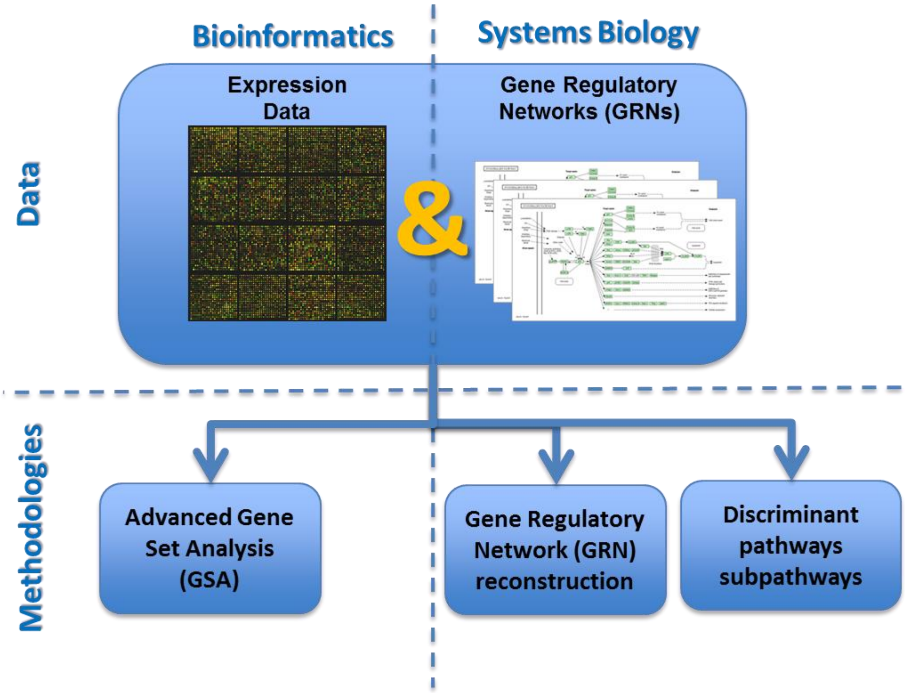

Bioinformatics
Systems Biology
Data
Expression Data
&
Gene Regulatory Networks (GRNs)
Methodologies
Advanced Gene Set Analysis (GSA)
Gene Regulatory Network (GRN) reconstruction
Discriminant pathways subpathways

Figure 6: Integration of microarray data with GRNs. Columns represent the two scientific areas, while rows map the data from these areas and the respective methodologies based on combinations of the data.

The initial search revealed that a lot of publications come from specific journals. We identified these journals and screened in depth the respective published articles. The journals that we focused are: (a) the annual Web Server issue of Nucleic Acids Research[2] and (b) the BMC Systems Biology (Software articles)[3]. We also identified that quite a few methodologies take advantage of the Cytoscape[4] platform to visualize and analyse gene regulatory networks. Thus we searched all the Cytoscape plugins in order to identify more tools/applications related to the identification and assessment of discriminant pathways.

After removing duplicates from the combined searches, the screening of the two journals and the screening of the related Cytoscape plugins, we came up with more than 100 unique citations. Most of the citations fall into the advanced Gene Set Analysis (GSA) or, into the GRN reconstruction categories (Figure 6). Out of these citations, 48 are related to GSEA with the utilization of GRN knowledge, 54 are related to GRN reconstruction using microarray data and 25 are related to discriminant pathways or, sub-pathways. Since this review focuses on methodol-

[2] http://nar.oxfordjournals.org/content/41/W1.toc#WebServices (last day visited 11/08/2014)

[3] http://www.biomedcentral.com/bmcsystbiol (last day visited 11/08/2014)

[4] http://www.cytoscape.org (last day visited 11/08/2014)

ogies that target the identification of the most phenotype-discriminant GRNs, the citations from the two first categories were rejected and our final pool of methodologies is limited to 25 citations. Here we have to note that such a distribution is expected since GSA and GRN reconstruction is the earliest research line in which coupling of gene-expression data and gene regulatory networks is utilised. As GSA and GRN-reconstruction methods are out of the scope we refer the interested reader to the related literature reviews [11], [17].

2.1. Gene regulatory networks reconstruction

Biologists use pathways to integrate results from literature, formulate hypotheses, capture empirical results, share current understanding and even run simulations. A common goal of research in the life sciences is to develop pathway models for biological processes of many different organisms.

Many studies focus on the problem of GRN reconstruction or reverse engineering of GRNs, which is how to construct, update or validate a network from other data sources.

2.1.1. Reconstruction using literature

Natural language processing (NLP) is a set of techniques that can help facilitate analysis, retrieval and integration of textual and electronic information. Recently the field of molecular biology has enjoyed an explosive development. As a result more and more publications on this field are available to the researchers. Taking advantage of the gowning size of documents related to gene interactions many researchers have propose automatic pathway identification using scientific publications.

Leroy et al [18] proposed a shallow parser, based on natural language processing, which captures the relations between noun phrases automatically from free text. The corpus of the parser consists of biomedical abstracts stored in a document warehouse. Evaluation of the parser has been done from 3 experts of the area. Park at el [19] and Daraselia et al [20] proposed two different systems to support parsing from MEDLINE. Park at el [19] extracts information about protein-to-protein interactions. The methodology of the parser is based on combinatory categorical grammar using appositions and compound nouns and anaphoric expressions. Daraselia et al [20] introduced a commercial software called MedScan, which uses natural language processing to extract interactions between proteins from related paper abstracts. The system validated using 3.5 million MEDLINE abstracts dated after 1988 and extracted 3601 interactions corresponding to 2976 distinct protein–protein interactions.

Other methodologies for gene regulatory network reconstruction have been proposed using text mining on complete text articles (publications) such as Friedman et al [21] who proposed the GENIES system. An NLP parser called MedLEE, which was applied to the domain of molecular biology for the extraction of molecular pathways from journal articles. MedLEE has been adapted to the molecular biology domain using a special molecular tag generator called term tagger. Another methodology in that direction proposed by Gaizauskas et al [22] called Protein Active Site Template Acquisition (PASTA), aims to extract information about the role of residues in protein molecules using text mining techniques.

2.1.2. Reconstruction using microarrays

The study of the function, structure and evolution of GRNs in combination with microarray gene-expression profiles and data is essential for contemporary biology research. Having in mind that differential expression analysis is a well-established strategy to screen genes or sets of genes associated with specific phenotypes, a lot of efforts focused on the reconstruction of GRNs by exploring gene-expression data have been done. Strong associations between genes found in microarray analysis can be candidates for gene interactions in a GRN.

According to microarray analysis new genes and gene associations are proposed to be added or deleted in the GRN. Figure 7 gives an example of GRN reconstruction where microarray data analysis identified a new path from RTK to P13K via PAK gene, an association between P13K and PIP3 and an activation of AkPKB from PIP3. At the top of Figure 7 we can see the original GRN and at the bottom the revised GRN according to a specific microarray dataset.

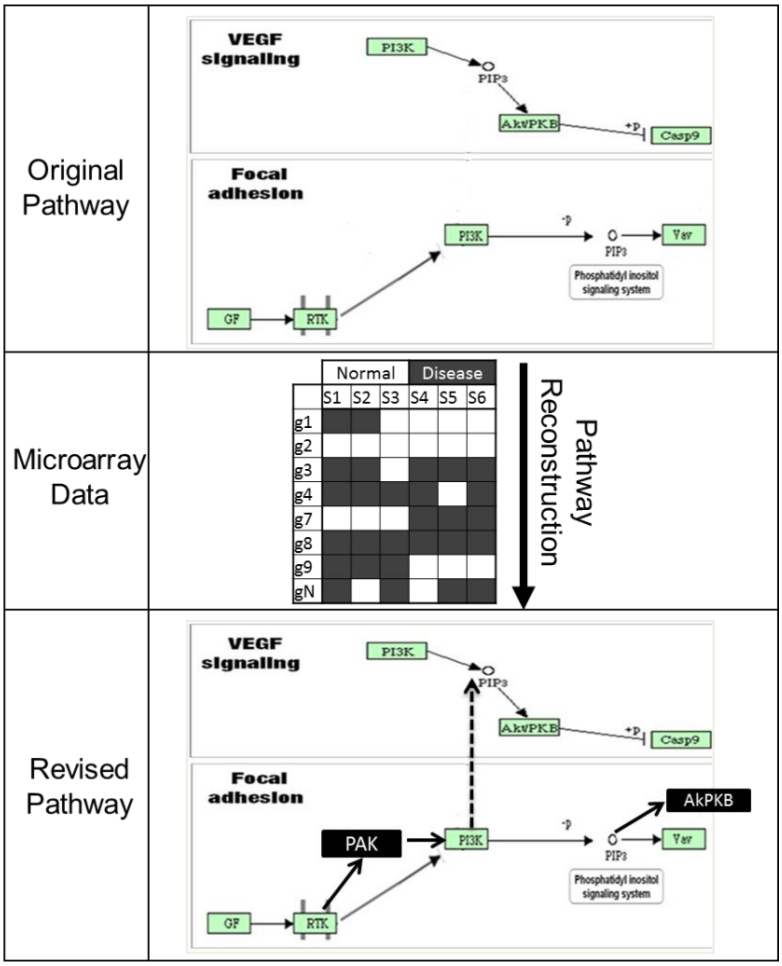


Figure 7: GRN reconstruction using microarray data. From top to bottom: Using known pathways and microarray expression data GRN reconstruction methodologies propose new gene to gene relations (e.g. PAK and AkPKB in the revised pathway).

An example of GRNs reconstruction using microarrays is RankGRN [23]. RankGRN evaluates a number of alternative hypothesises about the structure of a regulatory network against microarray data. RankGRN is a useful tool for evaluating the merits of different hypothesises on the structure of gene regulatory network using existing microarray data. It ranks the hypothetical gene network models based on their capability of explaining the microarray data.

Huang et al [24] proposed two scalable gene regulatory network learning algorithms: a modified information- theory-based Bayesian network algorithm and a modified association rule mining algorithm. Two types of evaluation were used to assess the practical value of these two techniques in helping researchers analyse large amounts of gene expression data. The simulation-based evaluation results indicated that the two techniques could infer about 20% of the relations in pre-defined network models.

Another methodology related to differentially expressed genes through microarray data and using interactome-transcriptome analysis was proposed by [25]. The paper concludes that the up-regulated genes in cancer samples tend to be "central hubs" in a network and the genes that are differentially expressed in contrast to the surrounding normal tissue, are essential for survival and proliferation.

A slightly different approach into that area proposed by Dutta et al [26] called PathNet. PathNet is a method for identifying enrichment and association between canonical pathways in the context of gene expression data. It takes into account topological information present in pathways to reveal biological information and is available as an R workspace image. PathNet utilizes the connectivity information in canonical pathway descriptions to help identify study-relevant pathways and characterize non-obvious dependencies and connections among pathways using gene expression data. It considers both the differential expression of genes and their pathway neighbours to strengthen the evidence that a pathway is implicated in the biological conditions characterizing the experiment. As an adjunct to this analysis, the system uses the connectivity of the differentially expressed genes among all pathways to score pathway contextual associations and statistically identify biological relations among pathways.

Very few methods of gene regulatory inference are considered superior, mainly because of the intrinsically noisy property of the data, '*the curse of dimensionality*' and the lack of knowledge about the 'true' underlying structure of the networks.

2.2. Gene Set Analysis

Gene set analysis (GSA), also called pathway inference, is a widely used strategy for gene expression data analysis based on pathway knowledge. GSA focuses on sets of related genes and has established major advantages over individual gene analyses, including greater robustness, sensitivity and biological relevance. GSA methods are better able to detect biologically relevant signals and give more coherent results across different studies. GSA incorporates prior knowledge of biological pathways and other experimental results in the form of gene sets.

Recently a lot of effort has been done in order to enrich the microarray analysis results with other biological data sources. One common approach is the combination of GRNs with microarray analysis for gene selection. Many methods use GRN information as groups (plain list) of associated genes in order to identify the most discriminant genes within microarray data (Figure 8 upper left part). Biological pathways are effectively reduced to sets of gene sets using a GSA approach with GRNs as a list of genes.

Although pathways maps carry important information about the structure of correlation among genes that should not be neglected, the currently available methods for gene set analysis do not fully exploit it. Recently, more and more methods take advantage of the topology of the gene regulatory network based on the graph theory and network visualization toolkits. Most of these tools take advantage of network visualization toolkits and display the discriminant genes from GSA methods on predefined gene regulatory networks (Figure 8 middle).

To our knowledge only a limited number of the published methodologies take advantage of the signalling information within the gene regulatory networks (e.g. the topology and the type of association between genes activation/inhibition) and can provide more biologically accurate interpretation of the data (Figure 8: downright part).

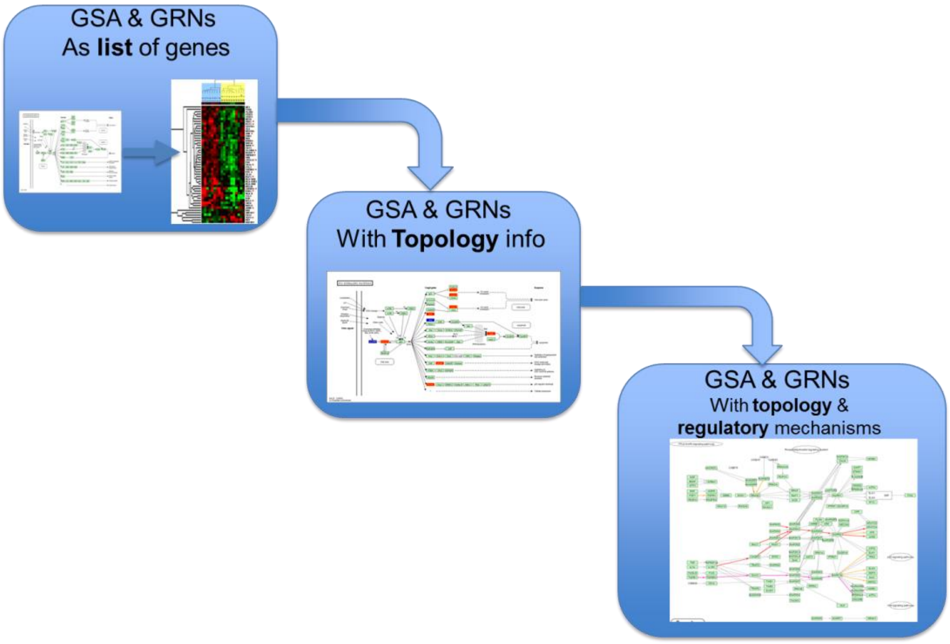
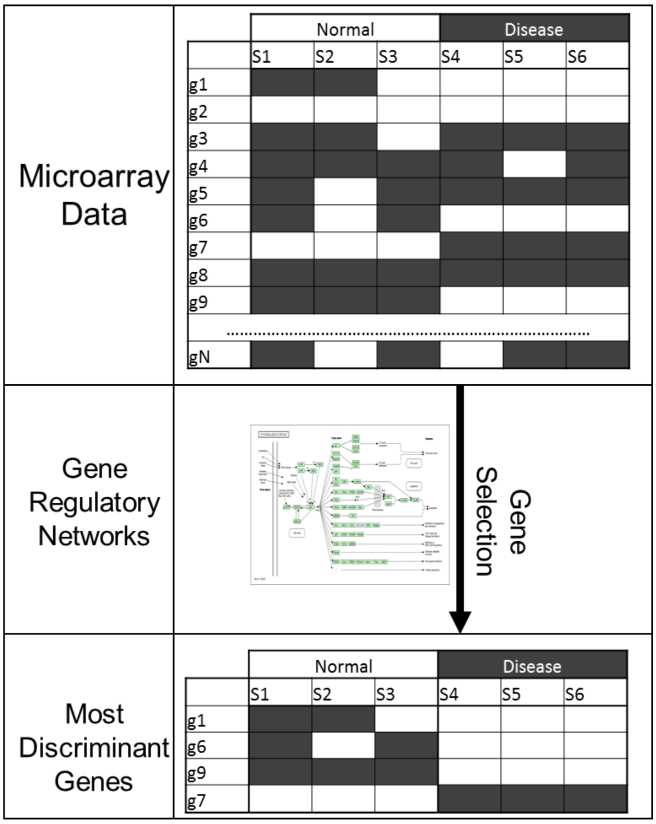


Figure 8: Evolution of Gene Set Analysis using GRNs. Initially the GRNs were treated as list of genes (left part) then the knowledge of the GRNs topology is taken into account (centre) and currently more and more methodologies take advantage of the regulatory mechanisms (right part).

The following sub-sections report methods, tools or algorithms that use microarray studies and GRN information as lists, topology or regulatory mechanisms in order to perform better accuracy at phenotype classification. A table which summarizes the gene set analysis methodologies according to main features such as input/data usage, output/purpose of use and type of application and visualization functionalities can be found in the end of the sub-section 2.2.

2.2.1. GSA & gene list from GRNs

Most of the methods proposed for gene set analysis use GRNs as group of genes to find differentially expressed group of genes on phenotype. Even though the knowledge from GRNs improves the efficiency of the selection algorithms, these methods do not take advantage of the topology of the network and the reactions/relations between genes. Gene regulatory networks are considered to be only group/list of genes (Figure 9) and such tools limit down the full list of genes in microarrays into the known list of genes from GRNs.



	Normal			Disease		
	S1	S2	S3	S4	S5	S6
g1	■	■	□	□	□	□
g2	□	□	□	□	□	□
g3	■	■	□	■	■	■
g4	■	■	■	■	□	■
g5	■	□	■	■	■	■
g6	■	□	■	□	□	□
g7	□	□	□	■	■	■
g8	■	■	■	■	■	■
g9	■	■	■	□	□	□
....						
gN	■	□	■	□	■	■

	Normal			Disease		
	S1	S2	S3	S4	S5	S6
g1	■	■	□	□	□	□
g6	■	□	■	□	□	□
g9	■	■	■	□	□	□
g7	□	□	□	■	■	■

Figure 9: Gene set analysis. From top to bottom: Having microarray expression data, we use GRNs to identify the genes that participate into known GRNs and we filter (narrow down) the microarray matrix.

Siu et al [27] proved that correlations among genes in a pathway are valuable and cannot be ignored in a gene expression analysis. The methodology is based on three statistical algorithms able to combine dependent P-Values of genes within a pathway.

Wang et al [28] proved that differential expression between two groups of samples is significantly different for genes in the pathway compared with the rest of the genes. Wang et al used linear mixed models for the analysis of microarray data at the pathway-level. The information used from the pathways is if a gene belongs to a pathway or not (basically if a pair of genes belong to the same pathway without taking into account the reactions of the pathway).

One common approach to combine microarray data with pathways is to incorporate known pathway information to reduce the dimensionality of gene interactions, such as [29] or [30]. Braun et al [29] proposed method identifies pairs of gene-pathway that are considered to be highly discriminant on microarray datasets. This method defines the expression of a known pathway via a summary value based on principal component analysis and uses KEGG pathways and handles the pathways as group of genes. Tai et al [30] proposed several versions of a modified linear discriminant analysis, group regularized discriminant analysis that aims to take advantage of existing gene functional groups. The algorithms make the assumption that the genes within the same pathway are correlated to each other. Methods were tested with simulated and real data and perform well compared to other known linear discriminant analysis algorithms for microarray analysis.

Sfakianakis et al [31] proposed a model for integration of gene annotations and pathways in order to guide the cluster analysis of gene expression data. The model gets information from the Gene Ontology (GO) and KEGG. The methodology takes advantage of the knowledge of pathways and creates a covariance matrix according to their existence or absence in pathways. Then an Expectation-Maximization algorithm is used for the identification of maximum likelihood solutions hidden variables in the model.

Another model that compares microarray experiments at the pathway level have been proposed by Beltrame et al [32] where the authors use pathways as a list of genes and computes the probability of a set of pathways to be related to some clinical/biological outcome. The proposed methodology for pathway signatures is based on the Eu.Gene application.

2.2.2. GSA with topology information from GRNs

The web based KEGG tool, Colour and mapper[5] is the simplest form of topology information on GRNs. The user can set colour to any gene within the gene regulatory network (Figure 10). Graph colour coding is a well-known approach that's used to simplify larger problems. The topology of a gene regulatory network is essential since the value of specific genes (drug targets) mainly because these genes can easily be used / manipulated using existing or new drugs. For example, a deviation from normal regulatory network topology may reveal the mechanism

[5] http://www.genome.jp/kegg/tool/map_pathway2.html (last day visited 11/08/2014)

of pathogenesis [33] and the genes that undergo the most network topological changes may serve as biomarkers or drug targets.

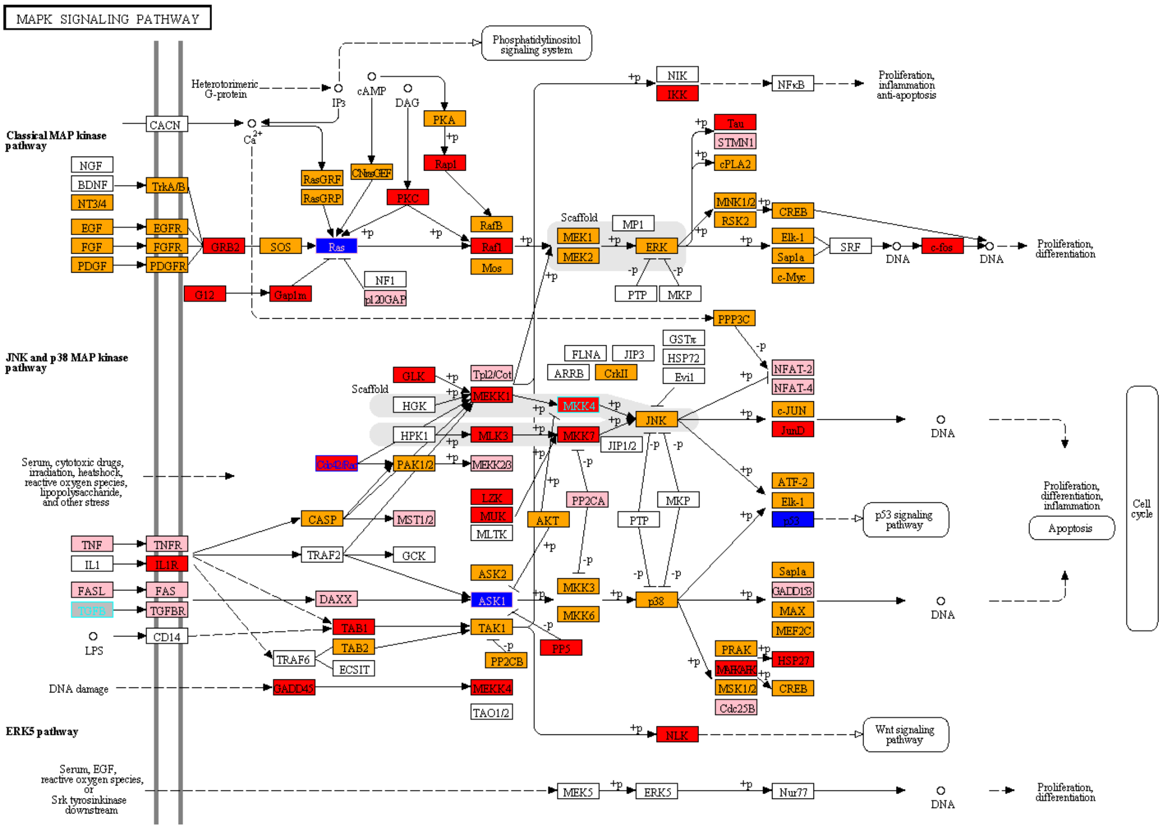


Figure 10: Example of the KEGG colour Mapper web application (source http://www.genome.jp/kegg/tool/map_pathway2.html)

Another indicative example from KEGG is the insulin pathway[6] as shown in Figure 11. If the insulin receptor (INSR) is not present, the entire pathway is shut off. Conversely, if several genes are involved in a pathway but they only appear somewhere downstream, changes in their expression levels may not affect the given pathway as much.

[6] http://www.genome.jp/kegg-bin/show_pathway?hsa04910 (last day visited 11/08/2014)

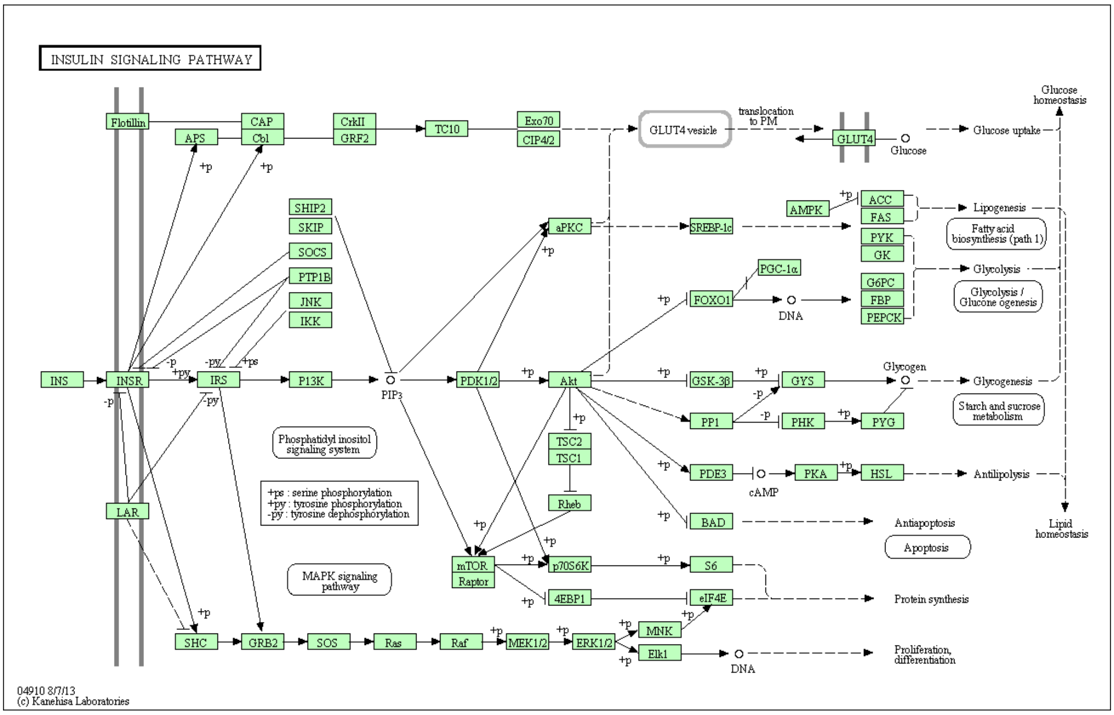
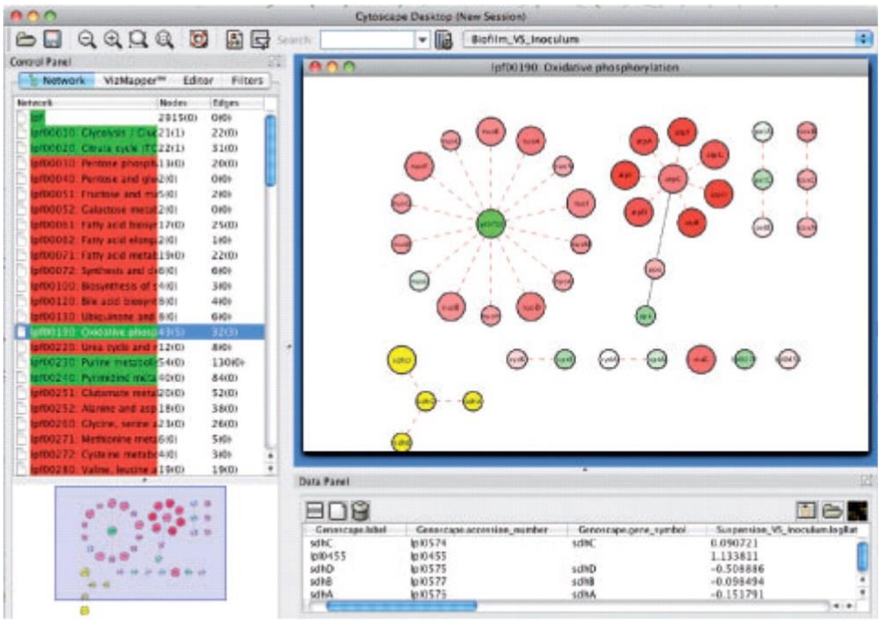


Figure 11: The KEGG insulin pathway (source http://www.genome.jp/kegg-bin/show_pathway?hsa04910)

Towards that direction a wealth of web based or standalone toolkits that take advantage of software platforms for visualizing complex networks exist. Most of the solutions rely on the cytoscape[7] Network Data Integration, Analysis and Visualization toolbox.

Genoscape [34] is an open-source Cytoscape plug-in that visually integrates gene expression data sets from GenoScript[8], a transcriptomic database and KEGG pathways into Cytoscape networks. Genoscape automatically maps most gene or gene product identifiers to KEGG identifiers, enabling the import of expression data from various sources. When importing KEGG pathways, elements are filtered in order to keep only those nodes corresponding to genes or enzymes. Using Genoscape, KEGG pathways are displayed as Cytoscape networks. Each pathway element is represented as a node. Genoscape generates a visualisation style that highlights gene expression changes and their statistical significance (Figure 12). The nodes represent genes and are coloured with a classical red/green gradient according to the expression ratio level. The size of the nodes is enlarged if the corresponding expression ratio is labelled as statistically significant.

[7] http://www.cytoscape.org/ (last day visited 11/08/2014)
[8] http://genoscript.pasteur.fr/cgi-bin/WebObjects/GenoScript (last day visited 11/08/2014)



Network	Nodes	Edges
lpf	2815(0)	0(0)
lpf00010: Glycolysis / Glu	21(1)	22(0)
lpf00020: Citrate cycle (TC	22(1)	31(0)
lpf00030: Pentose phosph	13(0)	20(0)
lpf00040: Pentose and glu	2(0)	0(0)
lpf00051: Fructose and ma	5(0)	2(0)
lpf00052: Galactose metab	2(0)	0(0)
lpf00061: Fatty acid biosyn	17(0)	25(0)
lpf00062: Fatty acid elong	2(0)	1(0)
lpf00071: Fatty acid metab	19(0)	22(0)
lpf00072: Synthesis and de	6(0)	6(0)
lpf00100: Biosynthesis of s	4(0)	3(0)
lpf00120: Bile acid biosynt	8(0)	4(0)
lpf00130: Ubiquinone and	6(0)	6(0)
lpf00190: Oxidative phos	43(5)	32(3)
lpf00220: Urea cycle and	12(0)	8(0)
lpf00230: Purine metabol	54(0)	130(0)
lpf00240: Pyrimidine meta	40(0)	84(0)
lpf00251: Glutamate meta	20(0)	52(0)
lpf00252: Alanine and asp	18(0)	38(0)
lpf00260: Glycine, serine a	23(0)	26(0)
lpf00271: Methionine meta	6(0)	5(0)
lpf00272: Cysteine metab	4(0)	3(0)
lpf00280: Valine, leucine a	19(0)	19(0)

Genoscape.label	Genoscape.accession_number	Genoscape.gene_symbol	Suspension_VS_Inoculum.logRat
sdhC	lpl0574	sdhC	0.090721
lpl0455	lpl0455		1.133811
sdhD	lpl0575	sdhD	-0.508886
sdhB	lpl0577	sdhB	-0.098494
sdhA	lpl0576	sdhA	-0.151791

Figure 12: Visualisation of GRN at GenoScape (source [34])

A similar approach incorporated Cytoscape is PiNGO [35]. PiNGO implements a simple network-based method to find genes associated with processes or pathways of interest. Input networks may be gene co-expression networks, protein or genetic interaction networks, or integrated networks. Edge weights are not taken into account. The candidate genes for each target category are listed along with P-values and associated raw counts that give a good indication of the prominence of the target category in the candidate gene's neighbourhood. Finally, the output network reveals the genes contributed to the discovery of particular candidate genes.

It appears that many publications at GSA and topology information use the Cytoscape open source visualization toolkit. Cline et al [36] proposed a protocol that explains how to use Cytoscape to analyse the results of mRNA expression profiling and other functional genomics and proteomics experiments, in the context of an interaction network obtained for genes of interest. Five major steps described: (i) obtaining a gene or protein network, (ii) displaying the network using layout algorithms, (iii) integrating with gene expression and other functional attributes, (iv) identifying putative complexes and functional modules and (v) identifying enriched Gene Ontology annotations in the network. Authors also made a comparative study of network analysis platforms that can be used for expression profiles and cellular networks.

The caBIG[9] project introduced the Differential Dependency Network (DDN) [37]. DDN is an analytical tool for detecting and visualizing statistically significant

[9] https://cabig.nci.nih.gov (last day visited 11/08/2014)

topological changes in transcriptional networks representing two biological conditions. DDN enables differential network analysis and provides an alternative way for defining network biomarkers predictive of phenotypes. DDN has been implemented as a standalone Java application to integrate network analysis and visualization seamlessly but a Cytoscape plug-in, CytoDDN, also exists.

Ibrahim et al [38] described a gene selection method, which identifies groups of strongly correlated genes that discriminate disease traits. In addition to using static predefined pathways knowledge, the method is adaptive in the sense that it involves a pathways ranking process to identify the most relevant pathways perturbed in a given pathological state and pathway topology.

A different topological approach, such as the centrality of nodes in the network or their tendency to form clusters has been implemented at the TopoGSA [39] (Topology-based Gene Set Analysis) web-application[10]. TopoGSA computes topological properties for the entire network, the uploaded gene/protein set and random sets of matched sizes. The available network topological properties are: (i) The degree of a node (gene or protein) is the average number of edges (interactions) incident to this node, (ii) The local clustering coefficient quantifies the probability that the neighbours of a node are connected, (iii) the shortest path length(SPL) for two nodes, (iv) the "*betweenness*" of a node that can be calculated from the number of shortest paths and (v) the centrality scores are given by the entries of the dominant eigenvector of the network adjacency matrix.

While the previous approaches are useful, the valuable information from GRNs such as the inherent regulatory relationships found in biological pathways among the different genes has never been incorporated in a gene set analysis methodology.

Table 1 summarizes the gene set analysis methodologies according to main features such as input/data usage, output/purpose of use and type of application and visualization functionalities. As we can see from the table none of these methodologies can identify discriminant sub-paths and all neglect the regulatory mechanisms reported in the GRNs. Furthermore, a few methodologies support visualization features and only one supports web based interface.

Table 1: List of Gene Set Analysis methodologies using pathways according to main features such as data usage, purpose of use, visualization functionalities and platform information

[10] http://bree.cs.nott.ac.uk/R-php-1/PPI (last day visited 11/08/2014)

		Use of microarray data	Use GRNs	Use sub-paths	Use pathway genes	Use topology	Use regulatory mechanisms	Identify discriminant genes	Identify discriminant pathways	Identify discriminant sub-paths	Web based	Visualization support
Gene Set Analysis	Siu et al [27]	✓	✓	X	✓	X	X	✓	X	X	X	X
	Wang et al [28]	✓	✓	X	✓	X	X	✓	X	X	X	X
	Braun et al [29]	✓	✓	X	✓	X	X	✓	✓	X	X	X
	Tai et al [30]	✓	✓	X	✓	X	X	✓	X	X	X	X
	Sfakianakis et al [31]	✓	✓	X	✓	X	X	✓	X	X	X	X
	Beltrame et al [32]	✓	✓	X	✓	X	X	✓	X	X	X	X
	KEGG color mapper	X	✓	X	X	✓	X	X	X	X	✓	✓
	Genoscape [34]	✓	✓	X	✓	✓	X	✓	X	X	X	✓
	PiNGO [35]	✓	✓	X	✓	✓	X	✓	X	X	X	✓
	Cline et al [36]	✓	✓	X	✓	✓	X	✓	✓	X	X	✓
	DDN [37]	✓	✓	X	✓	✓	X	X	✓	X	X	✓
	Ibrahim et al [38]	✓	✓	X	✓	✓	X	✓	✓	X	X	X
	TopoGSA [39]	✓	✓	X	✓	✓	X	✓	✓	X	X	X

2.3. Discriminant pathways and sub-pathways

An area that combines GRNs and microarray data, tries to identify the most discriminant GRNs for specific phenotypes. The phenotype information is extracted from microarrays and the evaluation of the most discriminant GRNs is based on the value of every gene in GRNs as it is expressed in microarray data.

The study of the function, structure and evolution of GRNs in combination with microarray gene-expression profiles and data is essential for contemporary biology research. The usual computational task involving microarray experiments is the gene selection procedure with GRNs used mainly for data annotation or GRNs reconstruction. Due to limitations in DNA microarray technology higher differential expressions of a gene do not necessarily reflect a greater likelihood of the gene being related to a disease and therefore, focusing only on the candidate genes with the highest differential expressions might not be the optimal procedure. A table which summarizes the pathway selection methodologies according to main features such as input/data usage, output/purpose of use and type of application and visualization functionalities can be found in the end of the subsection 2.3.

2.3.1. Pathway selection using microarray data

The straightforward approach for the identification of the most discriminant GRNs is to extract phenotype information microarrays and evaluate all the known GRNs (from literature or databases such as KEGG) for the identification of the most informative GRNs at the specific phenotype. The evaluation is based on the value of every gene in GRNs as it is expressed in microarray data.

Draghici et al [40] proposed a tool called Onto-Express, which automatically translate lists of differentially regulated genes into functional profiles. Onto-Express proposed a methodology for use of gene regulatory networks to find the pathways that contain the most discriminate genes (extracted from microarrays). The work is based on a combination of microarrays and gene regulatory networks (pathways) but the pathways are only used for informative purposes.

Oncomine [41] is a bioinformatics application for cancer signature identification. At version 3 of the application pathway information was added to the system for enrichment analysis of gene expressions. The extracted signature from multiple microarrays related to cancer reveal pathways that are co-ordinately over expressed in the respective cancer types.

Eu.Gene [42] is an application that tries to identify biological pathways transcriptionally affected under experimental conditions. The application can use multiple pathway databases and convert them to a common format (Ensembl Gene and Transcript IDs). Eu.Gene Analyzer implements two different statistical methods to evaluate the pathways that are most affected by differences in gene expression observed in a functional genomic experiment: the one-tailed Fisher Exact Test and Gene Set Enrichment Analysis (GSEA).

Adewale et al [43] proposed a statistical analysis of pathways using microarray data. Specifically the authors' handle the microarray data to identify pathways associated with the phenotype (e.g. time to death for breast cancer).Genes that participate (active at the microarray data) in a pathway make the pathway candidate. Then candidate pathways are tested if are significantly associated with various phenotype data and finally only the statistically significant pathways (group of genes) are selected.

Ma et al [44] proposed a methodology for the identification of gene pathways with predictive power for breast cancer prognosis. The work is based on statistical significance methods (p-value) using two quality controls: (i) to compute the predictive power of each gene within each pathway (ii) to compute the predictive power of each pathway in multiple datasets. The method works with multiple datasets, the pathway information is extracted from KEGG and it ignores the relationships between genes in a pathway (use the pathway as a group of genes).

PathBLAST [45] identifies and visually promotes pathway alignments of two different networks At PathBLAST the user specifies a short protein interaction path for query against a target protein–protein interaction network selected from a network database. PathBLAST returns a ranked list of matching paths from the target network along with a graphical view of these paths and the overlap among them. PathBLAST performs alignment of protein networks just as BLAST is used to perform rapid alignment of protein. The approach does not take into account microarray data.

GeneMANIA prediction server [46] constructs and displays an interactive functional association network constructed from a user-defined list of genes and functionally similar or shared propertied genes. Data sources used for gene similarity search include co-expression data from Gene Expression Omnibus; physical and genetic interaction data from BioGRID; predicted protein interaction database I2D; and pathway and molecular interaction data from Pathway Commons, which contains data from BioGRID, Memorial Sloan-Kettering Cancer Center, Human Protein Reference Database, HumanCyc, Systems Biology Center New York, IntAct , MINT, NCI-Nature Pathway Interaction Database and Reactome. The main drawback is GeneMANIA is that it can support only a limited set of initial genes due to the high complexity of the data sources used in the similarity search. Authors reported also an implementation of GeneMANIA as a Cytoscape plugin [47].

An approach for the identification of differentially expressed pathways has been proposed by Nacu et al [48]. The proposed methodology compute a score that measures to what extent a group of genes is differentially expressed. With a scoring function the system reveals groups of interacting genes. Two scoring methods developed and evaluated: (i) go through a limited list of predefined groups and select the ones with high scores (ii) search for high-scoring sets among all possible sets subject to some structural constraints.

All the above methods handle the gene regulatory networks only as a group/list of genes. Information about the topology of the network and the reactions/relationships between genes in a pathway is ignored.

2.3.2. Discriminant sub-pathways from MA and GRN topology

Several approaches for integrating microarray measurements with network knowledge were described in the literature and some of them proposed computational methods for detection of sub-networks that show correlated expression.

Chen et al [49] proposed a sub-pathway-based enrichment approach for identifying a drug response principal network, which takes into consideration the quantitative structures of the pathways. Authors are based on the biological pathways hint that a sub-pathway may respond more effectively or sensitively than the whole pathway. The methodology consists of the generation of a large number of relative sub-pathways (from the KEGG public database), mapping of the unfiltered expression data onto them and statistically scoring for identification of the principal component of sub-pathways that is most perturbed by two stage designs. Principal component of sub-pathways are then combined into a larger drug response network, on which topological and biological analyses are performed. The algorithm uses the NetworkAnalyzer [50] for the analysis of the topological properties of the sub-pathways. NetworkAnalyzer computes and displays a comprehensive set of topological parameters, from the network diameter to average clustering coefficients and shortest path lengths but ignores the regulatory mechanisms of the signalling pathways (activations/inhibitions).

DEGAS [51] (De Novo Discovery of Dysregulated Pathways in Human Diseases) methodology identifies connected gene sub-networks significantly enriched for genes that are dysregulated (disrupted of normal function) in specimens of a disease using correlation expressions. Given a set of expression profiles labelled as cases and another set of controls, DEGAS aims to detect sub-networks dysregulated in multiple genes in the cases, while allowing for distinct affected gene sets in each case profile.

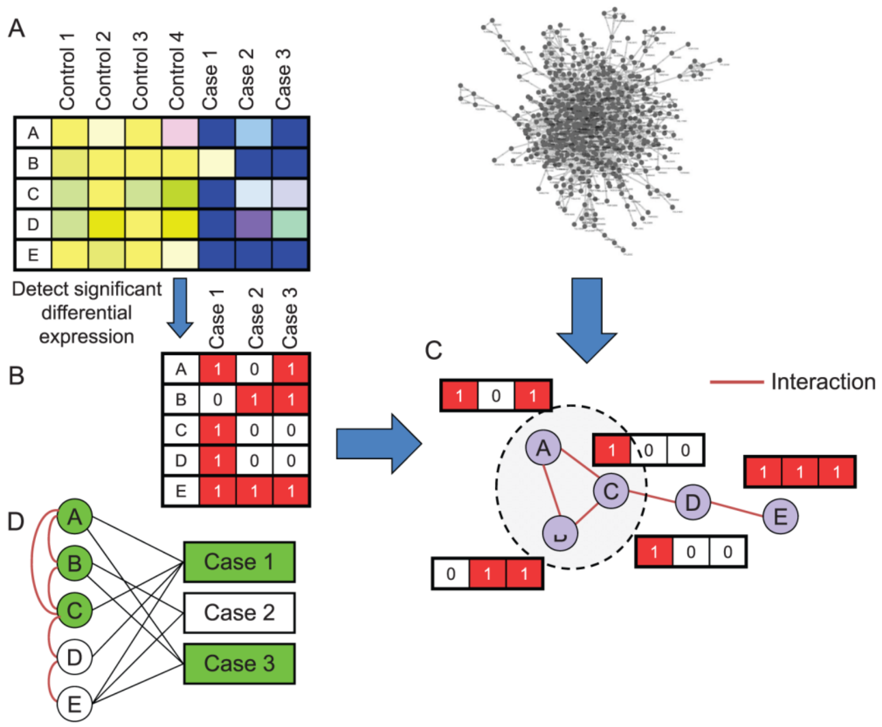


Figure 13: Identification of dysregulated pathways using DEGAS (source [51])

As shown in Figure 13 DEGAS methodology takes as input expression data of case and control cohorts (A) and a protein interaction network. The expression data are converted into a binary matrix. The output is the interaction network (C): The vector next to each protein is the dys-regulation status (0 or 1) of that gene in each case. A dysregulated pathway is a minimal sub-network in which at least k genes are dysregulated in all but l cases.

The gene regulatory relations we consider are restricted to what might be observed in a microarray experiment: a change in the expression of a regulator gene modulates the expression of a target gene mainly via protein-DNA interactions. In other words, there are genes that causally regulate other genes. A change in the expression of these genes might change dramatically the behaviour of the whole network. The identification and prediction of such changes is a challenging task in bioinformatics.

Another similar effort that actually uses the same algorithm for the identification of the dysregulated genes/cases is the KeyPathwayMiner [52]. Given a biological network and a set of case-control studies, KeyPathwayMiner efficiently extracts all maximal connected sub-networks (Figure 14). These sub-networks contain the genes that are mainly dysregulated, e.g., differentially expressed. The exact quantities for "mainly" and "most" are modelled with two easy-to-interpret parameters (K, L) that allow the user to control the number of outliers (not dysregulated genes/cases) in the solutions. KeyPathwayMiner use the Cytoscape visualization library to map the dysregulated sub-networks. Version 2.0 of KeyPathwaMiner [53] provide two more algorithms (one greedy and one optimal) to solve the formal graph problem and an improved user interface.

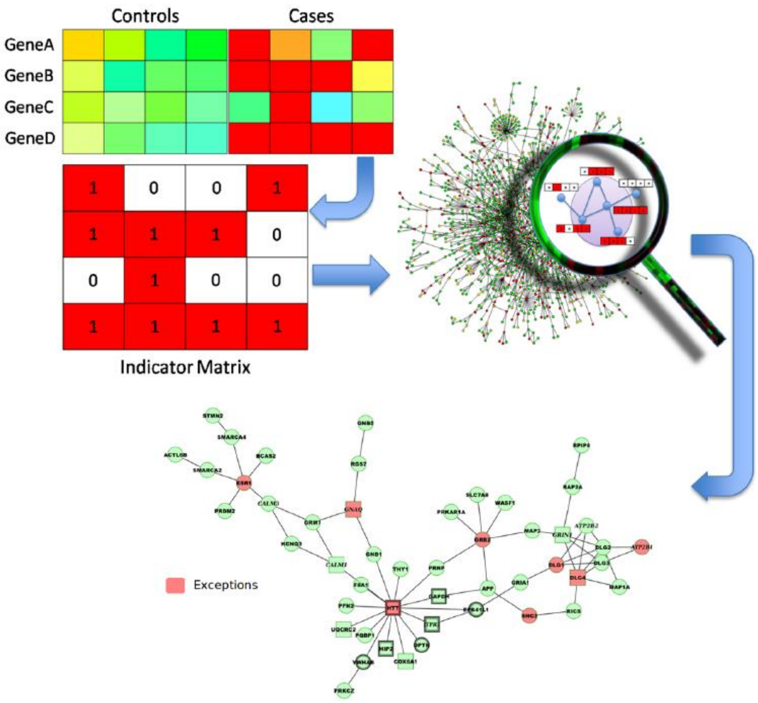


Figure 14: KeyPathwayMiner methodology (source http://keypathwayminer.mpi-inf.mpg.de)

Once again the main limitation of the KeyPathwayMiner is that interactions between two nodes (genes) are computed according to the expression values of the corresponding genes.

Ideker et al. [54] used sub-graph extraction as a technique to predict pathways from biological networks and a set of genes. The authors extended the methodology to the extraction of more complex, non-linear sub-networks in protein–protein and protein–DNA networks given yeast gene expression data. A recently work of the same team apply a protein network-based approach that identifies markers not as individual genes but as sub-networks extracted from protein interaction databases [55]. The resulting sub-networks of the methodology provide models of the molecular mechanisms underlying metastasis. Authors proved that the identified sub-networks are significantly more reproducible between different breast cancer cohorts than individual marker genes selected without network information and network-based classification achieves higher accuracy in prediction, as ascertained by selecting markers from one data set and applying them to a second independent validation data set. To integrate the expression and network data sets, authors overlaid the expression values of each gene on its corresponding protein in the network and searched for sub-networks whose activities across the patients were highly discriminative of metastasis. An overview of the sub-network identification is mapped visually at the Figure 15.

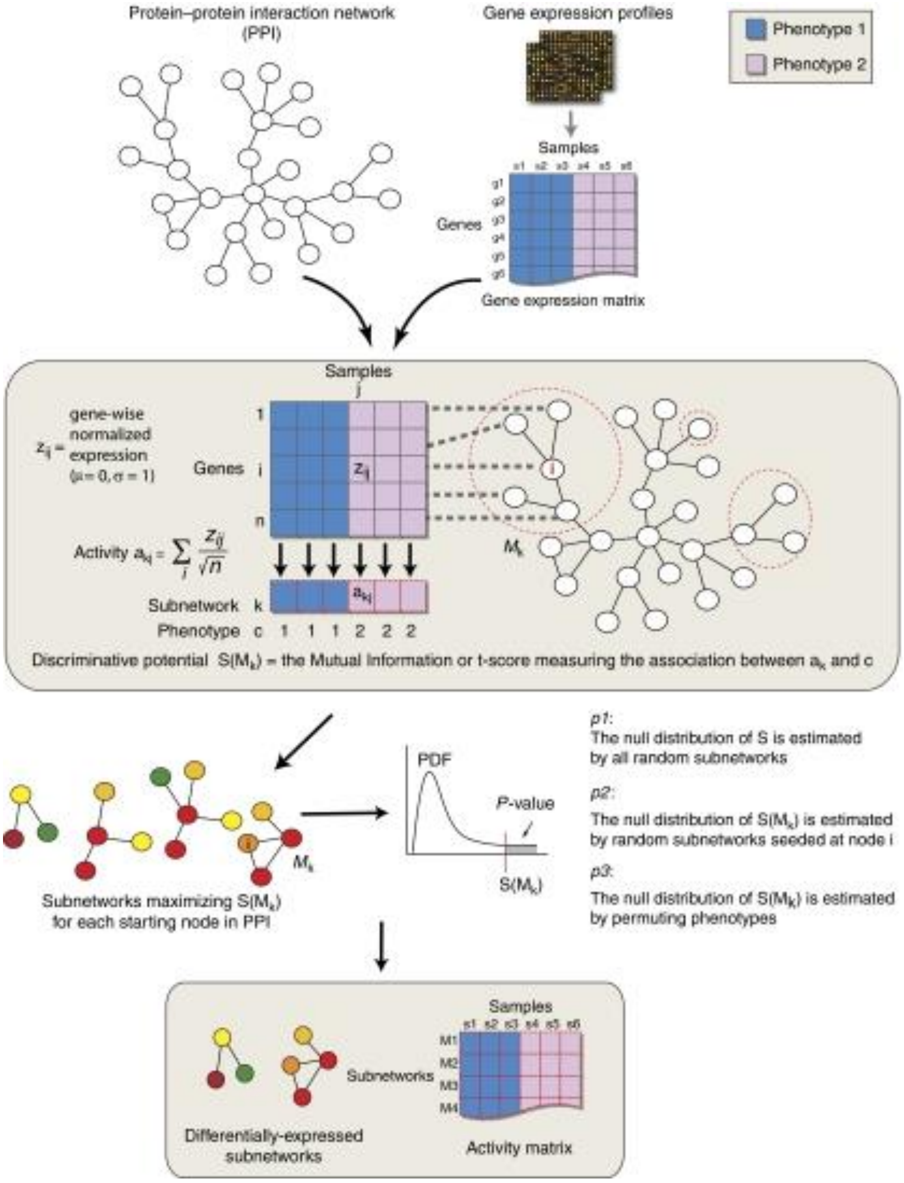

Protein–protein interaction network (PPI)
Gene expression profiles
Phenotype 1
Phenotype 2
Samples
s1 s2 s3 s4 s5 s6
Genes
g1 g2 g3 g4 g5 g6
Gene expression matrix
Samples
j
z_{ij} = gene-wise normalized expression ($\mu = 0, \sigma = 1$)
Genes
1
i
n
z_{ij}
Activity $a_{kj} = \sum_i \frac{z_{ij}}{\sqrt{n}}$
Subnetwork k
a_{kj}
Phenotype c 1 1 1 2 2 2
M_k
i
Discriminative potential $S(M_k)$ = the Mutual Information or t-score measuring the association between a_k and c
PDF
P-value
$S(M_k)$
M_k
i
Subnetworks maximizing $S(M_k)$ for each starting node in PPI
p1:
The null distribution of S is estimated by all random subnetworks
p2:
The null distribution of $S(M_k)$ is estimated by random subnetworks seeded at node i
p3:
The null distribution of $S(M_k)$ is estimated by permuting phenotypes
Samples
s1 s2 s3 s4 s5 s6
M1
M2
M3
M4
Subnetworks
Differentially-expressed subnetworks
Activity matrix

Figure 15: Sub-network identification process (source [54])

Sub-networks do not take into account initial relation of genes (from gene regulatory networks), but are considered active whenever they involve highly expressed genes. Sampling the space of possible sub-networks with simulated annealing can identify such sub-networks.

Wu and Stein [56] described a semi-supervised algorithm that first discovers modules of interacting genes (sub-pathways) involved in the disease process independently of clinical status and then identifies clinically significant modules using supervised principal component analysis. The implementation is based on top of a human protein functional interaction network constructed by combining curated and un-curated data sources. This functional interaction network covers roughly half of annotated human proteins and is highly reliable based on a variety of metrics, including confirmation of its predictions by domain experts. The network as a whole is un-weighted without regulation mechanisms and is not specific for any particular tissue or phenotype.

CLiPPER algorithm [57] implements a two-step empirical approach based on the exploitation of graph decomposition into a junction tree to reconstruct the most relevant signal path. In the first step clipper selects significant pathways according to statistical tests on the means and the concentration matrices of the graphs derived from pathway topologies. Then, it "clips" the whole pathway identifying the signal paths having the greatest association with a specific phenotype. For example, a proportional increase of the expression of the genes A and B in one of two conditions will result in significantly different mean expression between the two conditions. The correlation strength between A and B, however, does not change. In this case, we would have pathways with significant altered mean expression levels but unaltered biological interactions. CliPPER searches for pathways strongly involved in a biological process by requesting that the mean or the variance of the expression levels result significantly altered between two conditions. Clipper empirically identifies the portions of the network mostly associated to the phenotype using the structure of the junction tree as a backbone.

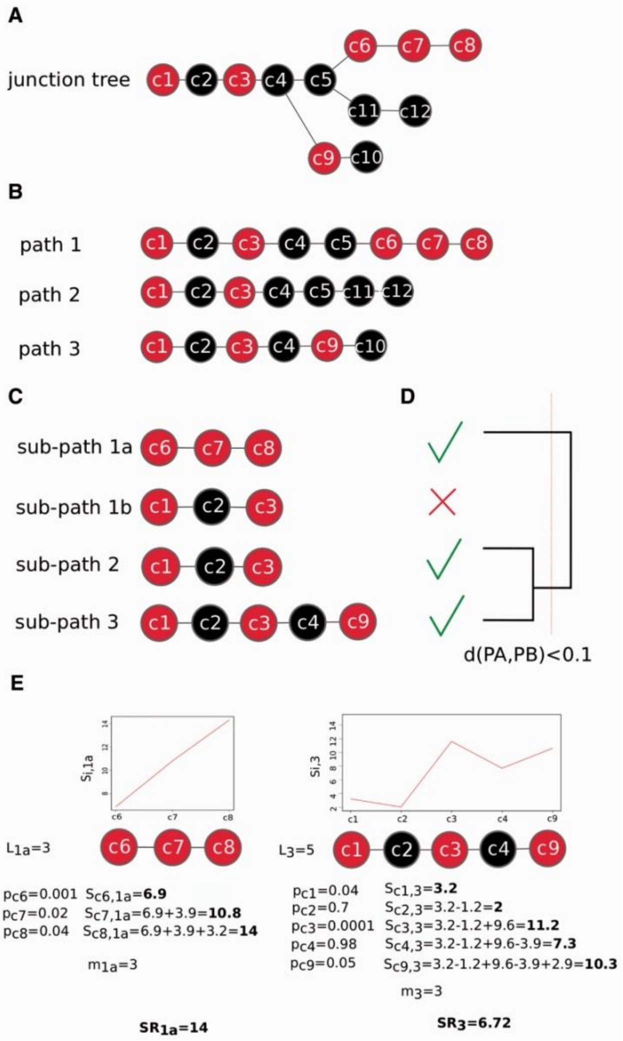


Figure 16: Clipper toy example of sub-path selection (source [57])

Figure 16 shows a toy example of clipper approach to sub-pathway selection. The construction of the junction tree with significant cliques is shown in red (part A). Identification of the paths in the tree is shown in part B, the identification of all the sub-paths within each path in part C, the selection of the best sub-path for each path and cluster analysis for sub-path collapse in part D and the final sub-path selected in part E.

Even though CliPPER uses parts of the pathway (sub-pathways) as junction tree, the sub-pathway selection method ignore the relations/regulations between genes participating in the signalling pathways.

Figure 17 shows results of CliPPER for the chronic myeloid leukaemia KEGG pathway with complexes belonging to the sub-path identified colour according to their expression.

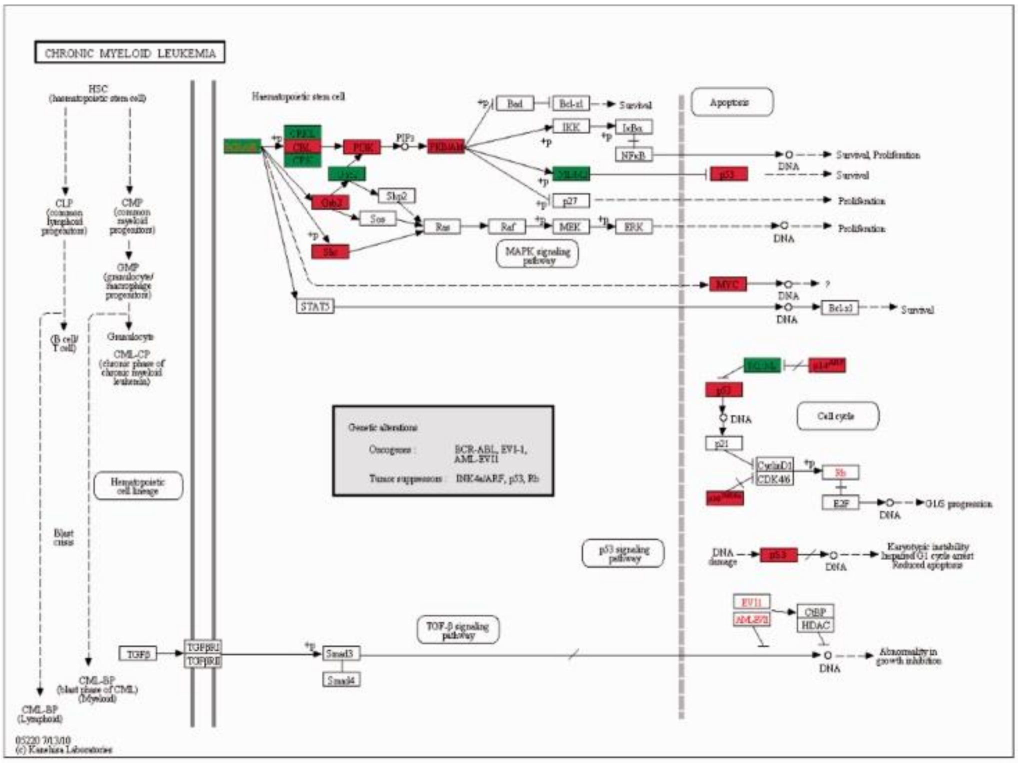
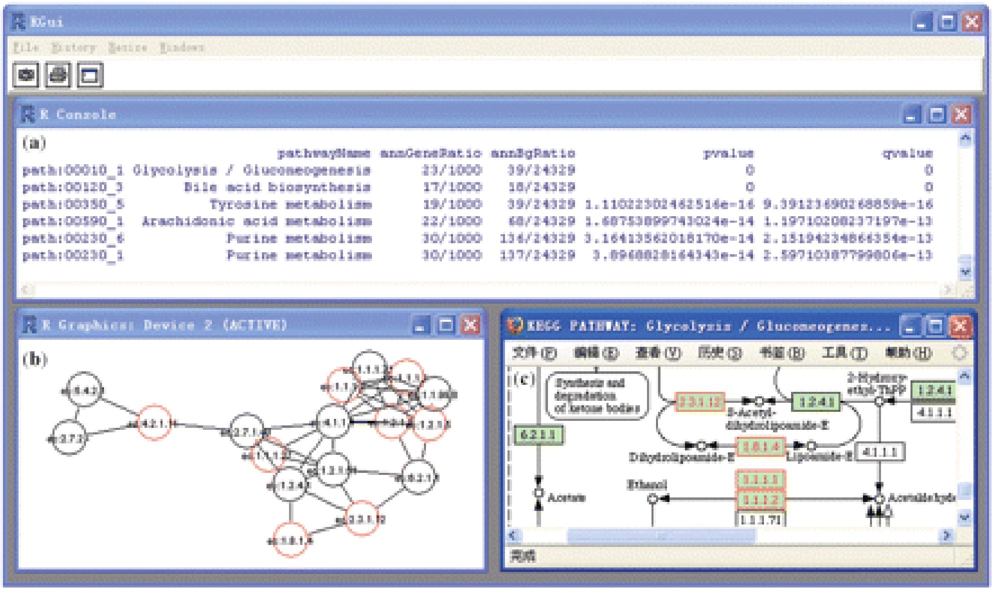


Figure 17: CliPPER results over KEGG pathway (source [57])

Kazmi et al [58] developed a meta-analysis tool for functional gene regulatory paths and sub-paths using information from microarray data. The up-regulated genes (found in microarray data) that participate in pathways are highlighted on the gene regulatory networks. The system takes advantage of the activations between genes within the pathway and tries to identify the functional paths or propose new paths. Expression values for genes that are not available from the microarray experiment are also added using a predictive algorithm.

Another software package (R based software) for identification of pathways is the SubpathwayMiner [59]. It is a pathway analysis tool relative to pathway annotation and identification, which applies pathway structure information to pathway identification. According to pathway structure information provided by KEGG, the system can detect distance similarity among enzymes in each pathway and mine each sub-pathway in which distance among all enzymes is no greater than the parameter k (a user-defined distance). SubpathwayMiner converts each metabolic pathway to an undirected graph with enzymes as nodes. Two nodes in an undirected graph are connected by an edge if there is a common compound in the enzymes corresponding reactions. As a result, the metabolic pathway is simplified when chemical compounds are omitted from the graph. Visualization of the resulting pathways is possible through linking to the KEGG website as shown in Figure 18 where (b) shows enzymes coloured red if the according enzyme is identified in the submitted sets of genes and (c) visualize a pathway through linking to the KEGG website. On the pathway map, enzymes are coloured red if the according enzyme is identified in the submitted set of genes


RGui
File History Device Windows
R Console
(a)
R Graphics: Device 2 (ACTIVE)
(b)
KEGG PATHWAY: Glycolysis / Gluconeogenes...
文件(F) 编辑(E) 查看(V) 历史(S) 书签(B) 工具(T) 帮助(H)
(c)
Synthesis and degradation of ketone bodies
2.3.1.12
1.2.4.1
6.2.1.1
1.8.1.4
4.1.1.1
1.1.1.1
1.1.1.2
1.1.1.71
Acetate
Ethanol
Dihydrolipoamide-E
Lipoamide-E
S-Acetyl-dihydrolipoamide-E
2-Hydroxy-ethyl-ThPP
完成

```
(a)
                          pathwayName annGeneRatio annBgRatio               pvalue               qvalue
path:00010_1 Glycolysis / Gluconeogenesis      23/1000   39/24329                    0                    0
path:00120_3         Bile acid biosynthesis      17/1000   18/24329                    0                    0
path:00350_5            Tyrosine metabolism      19/1000   39/24329 1.11022302462516e-16 9.39123690268859e-16
path:00590_1  Arachidonic acid metabolism      22/1000   68/24329 1.68753899743024e-14 1.19710208237197e-13
path:00230_6              Purine metabolism      30/1000  136/24329 3.16413562018170e-14 2.15194234866354e-13
path:00230_1              Purine metabolism      30/1000  137/24329  3.89688281643430e-14 2.59710387799806e-13
```

Figure 18: SubpathwayMiner environment (source [59])

The main limitation of the above proposed approaches is that all the interactions between genes within a GRN are considered to be connections in a graph (e.g. they do not take into account if an interaction is activation or inhibition) where the nodes are the genes and edges are interactions between genes.

2.3.3. Discriminant sub-paths from microarray, GRN topology and regulatory mechanisms

The most informative and promising methodology of microarrays and GRNs combination is the identification of discriminant sub-paths taking advantage of topology and regulatory mechanisms.

Geistlinger et al [16] introduced the Gene Graph Enrichment Analysis (GGEA), which exploit fundamental regulation types in a novel enrichment framework for signed and directed gene regulatory networks, to judge whether the topology of the network is well fitted by the expression data. GGEA performs three essential steps (Figure 19): first, the gene set is mapped onto the underlying regulatory network, yielding an induced sub-network. That is the affected part of the network, which consists of edges that involve members of the gene set. Second, each edge of the induced network is scored for consistency with the expression data, i.e. the signs of the expression changes of two interaction partners are evaluated for agreement with the regulation type (activation/inhibition) of the link that connects both genes. Third, the edge consistencies are summed up over the induced network, normalized and estimated for significance using a permutation procedure.

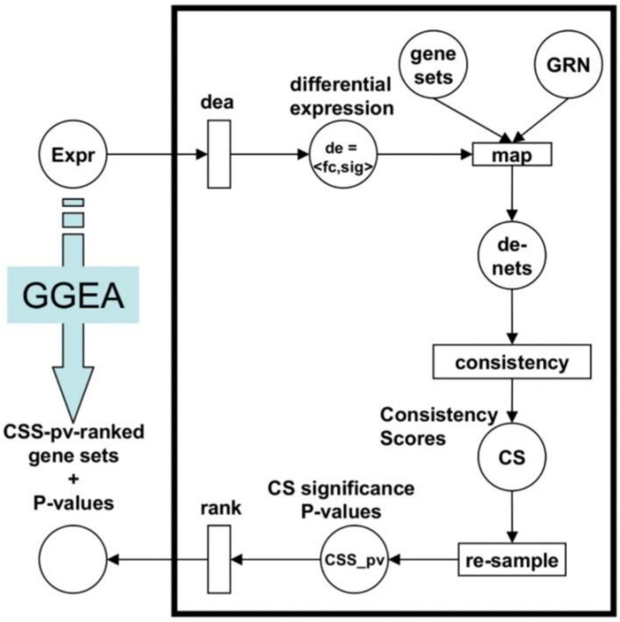


Figure 19: GGEA steps (source [16])

The GRNs are modelled as Petri Nets having features of fuzzy logic. The regulations of the GRN are required to be specified with direction and effect. In that model (Figure 20), regulator (R) and regulated target (RT) are represented via Petri Net places holding tokens of fuzzy values for both fold change (fc) and significance of fc (sig). The variety of regulatory effects occurring in the GRN are defined by specific fuzzy rules reg∈{f+,f−,f+−,f?,…} meaning activation f+, inhibition f− and dual effects f+−.

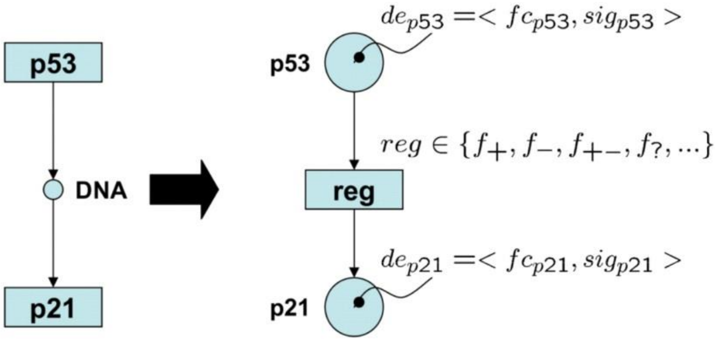

p53
DNA
p21
p53
$de_{p53} =< fc_{p53}, sig_{p53} >$
$reg \in \{f_+, f_-, f_{+-}, f_?, ...\}$
reg
$de_{p21} =< fc_{p21}, sig_{p21} >$
p21

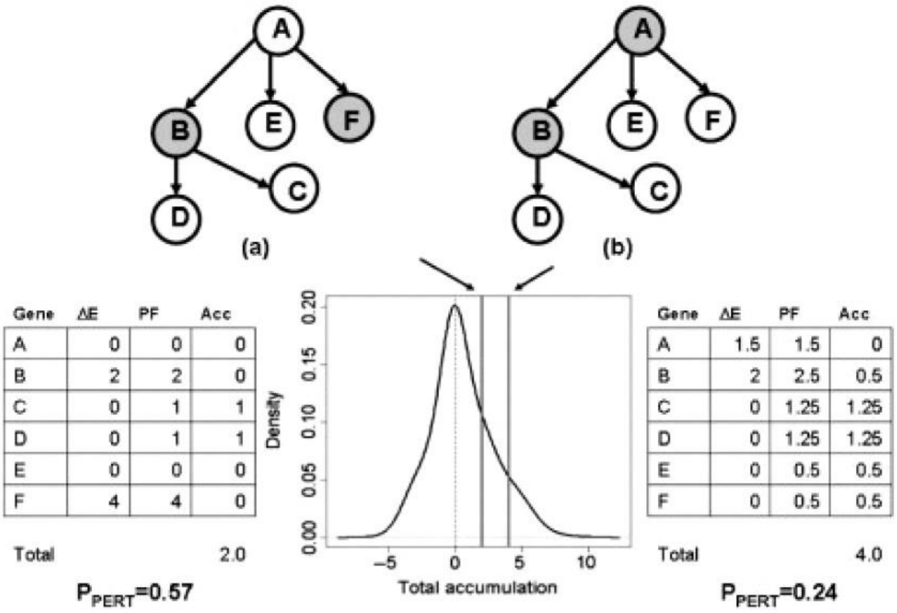
Figure 20: GGEA regulatory interactions mapping to Petri Net (source [16])

GGEA uses the regulation type of GRNs (activation/inhibition) to measure the consistency between expected (i.e. modelled) behaviour and the measured values. This approach solves the major problem of the set enrichment strategies, which is the contrary constrains between GRNs and expression data (e.g. two significantly up-regulated genes increase the enrichment of the set, even if one gene inhibits the other), but the GRN regulation information is only be used as a significance/ranking parameter in the whole pathway.

Similar to GGEA, another advanced discriminant sub-pathway identification system is the signalling pathway impact analysis (SPIA) [60]. SPIA combines the evidence obtained from the classical enrichment analysis with a novel type of evidence, which measures the actual perturbation on a given pathway under a given condition. To our knowledge this is the most advanced effort is in terms of gene interactions. The authors introduce a global probability value, P_G, which is calculated for each pathway, incorporating parameters, such as the log fold-change of the differentially expressed genes, the statistical significance of the set of pathway genes and the topology of the signalling pathway. P_G is a combined probability value of P_{NDE} and P_{PERT} that can be used to rank the pathways. P_{NDE} is the probability of observing the given number of differentially expressed genes or higher, just by chance and P_{PERT} is calculated in a bootstrapping process in which both the pathway and the number of differentially expressed genes per pathway are fixed. P_{PERT}, is calculated based on the amount of perturbation measured in each pathway and defined as:

$$PF(g_i) = \Delta E(g_i) + \sum_{j=1}^{n} \beta_{ij} \frac{PF(g_i)}{N_{ds}(g_i)}$$

Where the sign of β reflects the type of interaction: +1 for induction (activation), −1 for repression and inhibition, as described by each pathway. Note that β will have non-zero value only for the genes that directly interact with the gene g_i according to the pathway description. Each pathway is finally marked as activated (positive perturbation score = positively perturbed) or the inhibited (or negatively perturbed)



Gene	ΔE	PF	Acc
A	0	0	0
B	2	2	0
C	0	1	1
D	0	1	1
E	0	0	0
F	4	4	0
Total			2.0

P_{PERT}=0.57

Gene	ΔE	PF	Acc
A	1.5	1.5	0
B	2	2.5	0.5
C	0	1.25	1.25
D	0	1.25	1.25
E	0	0.5	0.5
F	0	0.5	0.5
Total			4.0

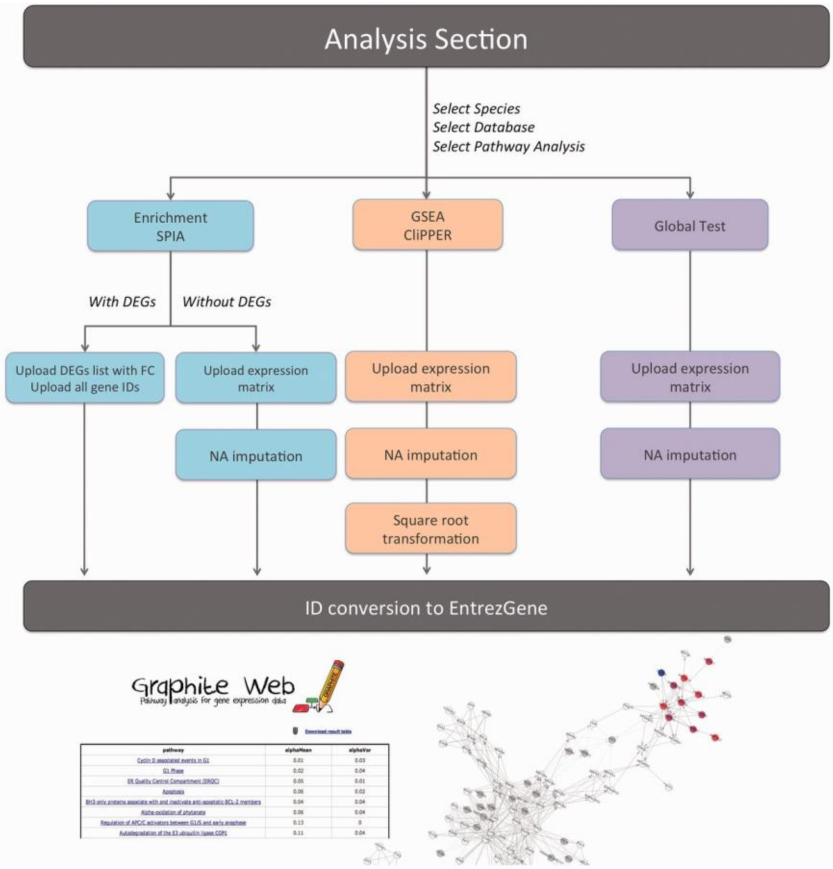
P_{PERT}=0.24

Figure 21: SPIA perturbation analysis example (source [60])

Figure 21 shows a six-gene pathway with two differentially expressed genes (shown in grey) in two different situations. One of the two differentially expressed genes is in common (gene B) while the second gene is either a leaf node (a), or the entry point in the pathway (b). In (a), gene (F) cannot perturb the activity of other genes; in (b) gene (A) has the ability to influence the activity of all the remaining genes in the pathway, as the topology of the pathway indicates. An over-representation analysis would find the two situations equally (in) significant (P_{NDE}=0.48 for a set of 20 monitored genes, out of which five are found to be DE). The perturbation evidence extracted by SPIA will give more significance to the situation in (b) (P_{PERT}=0.24), even though fold-changes in (b) are almost twice as small as those in (a) (P_{PERT}=0.57). SPIA provides information of the pathway as a whole only and does not tackle functional and non-functional parts of the pathway (sub-pathways).

Graphite Web[11] [61] is a web tool for gene set analysis exploiting pathway topology. Graphite web implements five different gene set analyses on three model organisms and two pathway databases and is freely available. Graphite web deals with microarray or RNA-seq data. It implements different multivariate gene set analyses, gene set enrichment analysis (GSEA), signalling pathway impact analysis (SPIA), CliPPER on three model organisms (human, mouse and drosophila) and two pathway databases (KEGG and Reactome). We added Graphite web in this category since it uses the SPIA methodology for signalling pathway analysis.

[11] http://graphiteweb.bio.unipd.it (last day visited 11/08/2014)



pathway	alphaMean	alphaVar
Cyclin D associated events in G1	0.01	0.03
G1 Phase	0.02	0.04
ER Quality Control Compartment (ERQC)	0.05	0.01
Apoptosis	0.06	0.02
BH3-only proteins associate with and inactivate anti-apoptotic BCL-2 members	0.04	0.04
Alpha-oxidation of phytanate	0.08	0.04
Regulation of APC/C activators between G1/S and early anaphase	0.13	0
Autodegradation of the E3 ubiquitin ligase COP1	0.11	0.04

Figure 22: Graphite Web flow of operations (source [61])

Graphite web implements a system of pathway visualization and provides an easy access to multivariate and topological pathway analyses. The combination of a pathway-specific visualization with powerful gene set analyses gives to the user the possibility to explore in great detail signalling pathways and the position of the influential genes within them.

Another method that identifies intergenic relationships within enriched biologically relevant sub-pathways is the Topology Enrichment Analysis frameworK TEAK [15]. TEAK employs a novel in-house algorithm and a tailor-made Clique Percolation Method to extract linear and nonlinear KEGG subpathways, respectively and scores subpathways using the Bayesian Information Criterion for context specific data and the Kullback-Leibler divergence for case-control data. Subpathway extraction is an important component of TEAK that extracts root to leaf linear paths or subpathways from the directed edges of the KEGG non-metabolic pathways. A root *r* has zero incoming links and a positive number of outgoing links, whereas a leaf l has a positive number of incoming links and zero outgoing links.

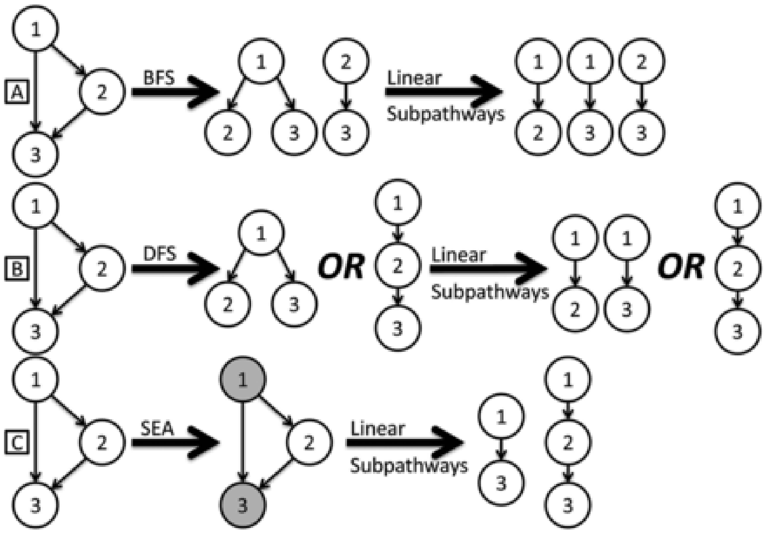


Figure 23: TEAK methodology to identify the subpathways of a network (source [15])

The subpathway algorithms and the Bayesian networks used by TEAK are only applicable to directed networks. The type of regulation between the nodes (genes) is considered to be always activation (e.g. over expression of gene A leads always to over expression of gene B if we have an A→B link). It's not clear if inhibition of genes is also treated in the same way or it is ignored. To rank the linear and nonlinear subpathways, TEAK first uses the Bayes Net Toolbox to fit a context specific Gaussian Bayesian network for each sub-pathway. Briefly, a Gaussian Bayesian network is a Bayesian network in which all of its nodes are linear Gaussians.

PATHOME [13] (pathway and transcriptome information) is another recent methodology for detecting differentially expressed biological pathways. The goal of this algorithm is to identify a set of sub-pathways that differentiate two experimental groups (for example, cancer vs non-cancer) by considering both prior knowledge about mutual regulations and experimental gene expression data.

If two adjacent entries are connected by an edge that denotes activation (arrowheaded edge), the expression correlation between the two entries is assumed to be positive; if the two entries are connected by an edge that denotes inhibition (blunt-ended edge), the expression correlation between the two entries is assumed to be negative (Figure 24). This rule is applied separately to each experimental group. In each group, PATHOME identifies the consecutive segment starting from the leaf node of each sub-pathway so that all the edges of the segment should satisfy the association rule. That leads to the determination of the segment (in the sub-pathway) that is to be statistically evaluated in the test step.

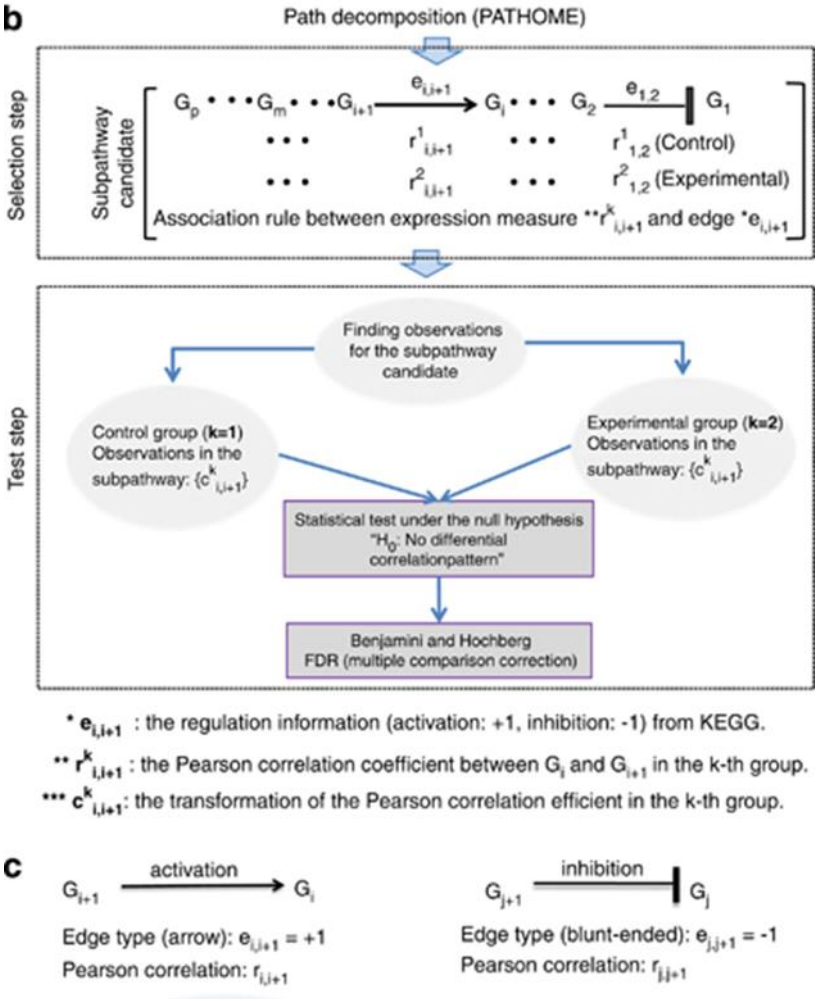

b
Path decomposition (PATHOME)
Selection step
Subpathway candidate
$G_p \cdots G_m \cdots G_{i+1} \xrightarrow{e_{i,i+1}} G_i \cdots G_2 \xrightarrow{e_{1,2}} G_1$
$r^1_{i,i+1}$ $r^1_{1,2}$ (Control)
$r^2_{i,i+1}$ $r^2_{1,2}$ (Experimental)
Association rule between expression measure $^{**}r^k_{i,i+1}$ and edge $^*e_{i,i+1}$
Test step
Finding observations for the subpathway candidate
Control group (k=1) Observations in the subpathway: $\{c^k_{i,i+1}\}$
Experimental group (k=2) Observations in the subpathway: $\{c^k_{i,i+1}\}$
Statistical test under the null hypothesis "H_0: No differential correlationpattern"
Benjamini and Hochberg FDR (multiple comparison correction)
$^*e_{i,i+1}$: the regulation information (activation: +1, inhibition: -1) from KEGG.
$^{**}r^k_{i,i+1}$: the Pearson correlation coefficient between G_i and G_{i+1} in the k-th group.
$^{***}c^k_{i,i+1}$: the transformation of the Pearson correlation efficient in the k-th group.
c
G_{i+1} activation G_i
Edge type (arrow): $e_{i,i+1} = +1$
Pearson correlation: $r_{i,i+1}$
G_{j+1} inhibition G_j
Edge type (blunt-ended): $e_{j,j+1} = -1$
Pearson correlation: $r_{j,j+1}$

Figure 24: PATHOME pathway decomposition and genes regulation mapping (source [13])

PATHOME analyses the interconnectivity between two adjacent nodes. The interconnectivity measure, the Pearson product-moment correlation coefficient, is obtained even in three samples in a group. PATHOME can be applied to a small number of samples, such as three samples in a group. Summarizing the first step, a candidate sub-pathway for the next step should satisfy the following two conditions: (i) the two experimental groups agree with the association rule between the expression correlation and the edge information for the adjacent entries along the path; and (ii) both consecutive segments for the two groups have at least four elements (three consecutive edges) in order to filter a sub-pathway with short segments.

Table 2 summarizes the discriminant pathways methodologies according to main features such as input/data usage, output/purpose of use and type of application and visualization functionalities. As we can see from the table only five methodologies (Graphite Web uses SPIA) can handle effectively the regulatory mechanisms and only three out of them can identify discriminant sub-paths in

GRNs. Furthermore, most of the methodologies lack of visualization features and support for web based platform.

Table 2: List of discriminant pathways and sub-pathways methodologies according to main features such as data usage, purpose of use, visualization functionalities and platform information

		Use of microarray data	Use GRNs	Use sub-paths	Use pathway genes	Use topology	Use regulatory mechanisms	Identify discriminant genes	Identify discriminant pathways	Identify discriminant sub-paths	Web based	Visualization support
Discriminant pathways and sub-pathways	Draghici et al [40]	✓	✓	X	✓	X	X	X	✓	X	X	X
	Oncomine [41]	✓	✓	X	✓	X	X	X	✓	X	X	X
	Eu.Gene [42]	✓	✓	X	✓	X	X	X	✓	X	X	X
	Adewale et al [43]	✓	✓	X	✓	X	X	X	✓	X	X	X
	Ma et al [44]	✓	✓	X	✓	X	X	X	✓	X	X	X
	PathBLAST [45]	✓	✓	X	✓	X	X	X	✓	X	X	✓
	GeneMANIA [46]	✓	✓	X	✓	X	X	✓	✓	X	X	✓
	Nacu et al [48]	✓	✓	X	✓	X	X	✓	✓	X	X	X
	Chen et al [49]	✓	✓	✓	✓	✓	X	X	✓	✓	X	X
	DEGAS [51]	✓	✓	✓	✓	✓	X	X	✓	✓	X	X
	KeyPathwayMiner [52]	✓	✓	✓	✓	✓	X	X	✓	✓	X	✓
	Ideker et al. [54]	✓	✓	✓	✓	✓	X	X	✓	✓	X	X
	Wu and Stein [56]	✓	✓	✓	✓	✓	X	X	✓	✓	X	X
	CLiPPER [57]	✓	✓	✓	✓	✓	X	X	✓	✓	X	✓
	Kazmi et al [58]	✓	✓	✓	✓	✓	✓*	X	✓	✓	X	✓
	SubpathwayMiner [59]	✓	✓	✓	✓	✓	X	X	✓	X	✓	✓
	Geistlinger et al [16]	✓	✓	✓	✓	✓	✓	X	✓	X	X	X
	SPIA [60]	✓	✓	✓	✓	✓	✓	X	✓	✓	X	X
	Graphite Web [61]	✓	✓	✓	✓	✓	✓	✓	✓	X	✓	✓
	TEAK [15]	✓	✓	✓	✓	✓	✓	✓	✓	✓	X	X
	PATHOME [13]	✓	✓	✓	✓	✓	✓	✓	✓	✓	X	X

* takes advantage only of the activations between genes

2.4. Outcome from the literature

The study of the function, structure and evolution of GRNs in combination with microarray gene-expression profiles is essential for contemporary biology research. Due to limitations in DNA microarray technology - due to the different platforms utilised, to the different experimental protocols and mainly to small sample sizes, higher differential expressions of a gene do not necessarily reflect a greater likelihood of the gene being related to a disease and therefore, focusing only on the candidate genes with the highest differential expressions might not be the optimal procedure [9], [10].

Based on our literature research we identified and propose taxonomy of the methodologies that combine gene-expression data and GRNs in order to identify and assess discriminant pathway and sub-pathways (Figure 25).

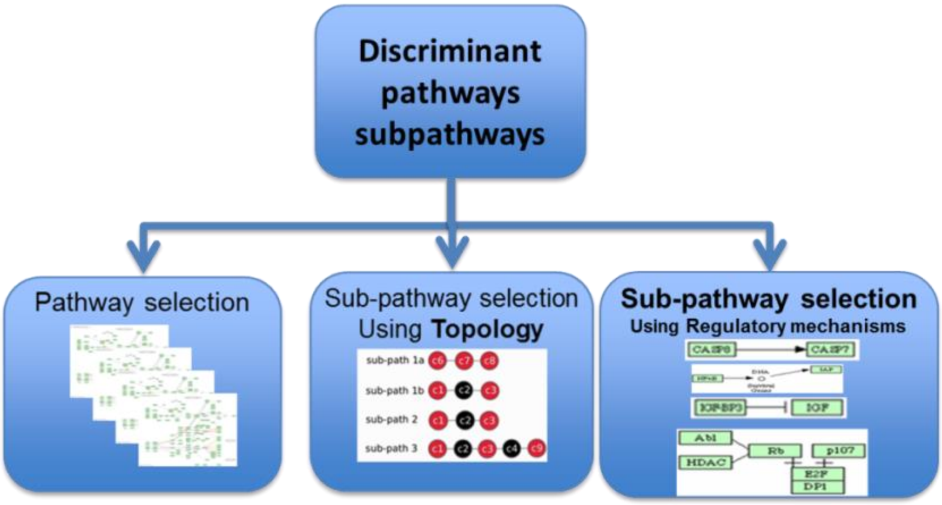


Figure 25: Taxonomy of discriminant pathways and sub-pathways. Three main categories: Pathway selection, sub-pathway selection using topology and sub-pathway selection using regulatory mechanisms

A general observation concerns the different levels of knowledge extraction from the GRNs employed by the different methods.

- The first category naming "*pathway selection*" focuses on the identification of differentially expressed pathways using microarray data. Within this approach information about the topology, the existing sub-paths, as well as the reactions/relationships between genes in a pathway is ignored.
- The second category "*sub-pathway selection using topology*" goes one step further and tries to identify discriminant pathways or sub-pathways. Within this approach identification and selection of the most discriminant paths ignore the present gene relations/regulations.

- The last and most informative category is the "*sub-pathway selection using regulatory mechanisms*". This approach takes advantage of the GRN topology as well as the type of GRN gene relations (e.g. activation or inhibition).

The last category – being in its infancy, exhibits the fewer methodologies so far, but it takes the most out of GRNs and gene-expression data compared to the other two and is a promising alternative for the identification of the regulatory mechanisms that underlie and putatively govern various phenotypes.

The sub-paths selection using the underlying GRN gene regulatory interactions approach solves the major problem of the set enrichment strategies that refers to the conflicting constrains between GRNs and gene-expression data. A typical example of the conflicting constrains is reflected in the situation when two significantly up-regulated genes increase the enrichment of the set in microarray expression data, even if the first gene inhibits the other in a GRN.

There exists a limited number of systems that utilize knowledge from known GRNs, namely GGEA [16], SPIA [60], TEAK [15] and DEAP [62]. However, these systems cannot visualize efficiently the results, a fact that does not facilitate inspection of results and limits the exploratory potential by the users. Some gene set enrichment analysis methodologies and tools utilize pathway visualization approaches to overcome this problem. Since these are based on a gene-oriented approach are still unable to handle differentially expressed pathways or even differentially expressed sub-paths.

In chapter 3 we introduce our proposed methodology for the identification of differentially expressed functional paths or sub-paths within a gene regulatory network (GRN) using microarray data analysis. The analysis takes advantage of interactions among genes (e.g. activation, inhibition) as nodes of a graph network, which are derived from expression data.

3.Methodology

Deciphering and manifestation of functioning and regulation of genes represents a necessary condition toward the effective incorporation of genomic data in everyday clinical practice. Two of the most significant forms of molecular data come from microarray gene expression sources and gene interactions sources – as encoded in Gene Regulatory Networks.

Existing GRN databases provide us with widely utilized networks of proved molecular validity. The most known are networks that describe important cellular processes such as cell-cycle, apoptosis, signalling and regulation of important growth factors. Online public repositories contain a variety of information that includes not only the network per se but links and rich annotations for the respective nodes (genes) and edges (reactions). MinePath utilizes the KEGG pathways [63] repository. KEGG provides a format representation standardized by its own mark-up description language (KGML[12]).

Figure 26 outlines the flow of operations in the MinePath methodology. MinePath presents a novel perception of GRNs and gene expression data. Initially we locate all functional paths encoded in GRNs and we try to assess which of them are compatible with the gene-expression values of samples that belong to different clinical categories (diseases and phenotypes). The differential power of the selected paths is computed and their biological relevance is assessed. The approach is applied on a set of microarray studies with the target of revealing putative regulatory mechanisms that govern the treatment responses of specific phenotypes.

[12] http://www.genome.jp/kegg/xml/ (last day visited 11/08/2014)

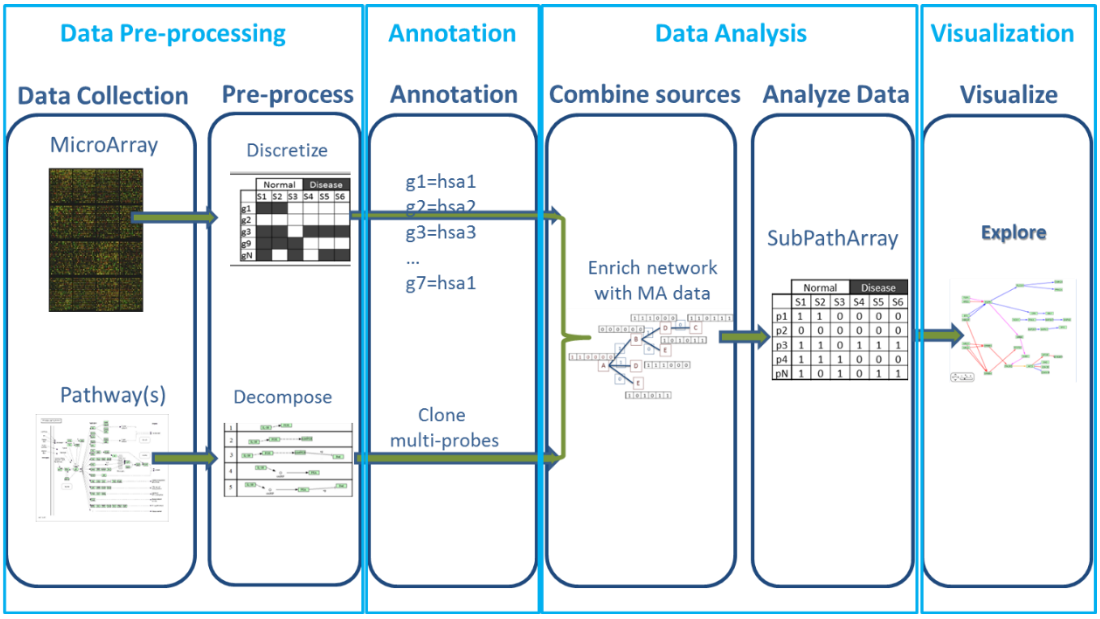


Figure 26: MinePath flow of operations. Four main processes starting from data pre-processing to annotation, then data analysis and finally to the visualization of the results

GRN and gene-expression data matching aims to differentiate GRN paths and identify the most prominent functional sub-paths for the given samples. In other words, the quest is for the sub-paths that exhibit high matching scores for one phenotypic class and low matching scores for the other. This is a paradigm shift from the mining of differential genes to the mining of GRN functional sub-paths. The whole algorithmic process for the identification of phenotype differential sub-paths is inherently simple.

The method unfolds into four modular steps:

I. ***Data pre-processing***: On the one hand, gene expression values are discretized into two states with values 1 and 0 for up-regulated and down-regulated genes, respectively, so that a binary gene-expression sample matrix is formed. On the other hand, each target GRN is decomposed into its constituent sub-paths, e.g., the path A → B —| C is decomposed into three sub-paths, A → B, B —| C and A → B —| C (note that all sub-paths, as well as the overlapping ones, are identified, formed and stored). The pre-processing step for gene expression data discretization will be discussed in section 3.1.1 and the gene regulatory networks decomposition will be described in section 3.1.2.

II. ***Identification of functional sub-paths:*** Each sub-path is interpreted on the basis of its functional active-state and is represented by a binary ordered-vector with active states. For example, sub-path A → B —| C is considered active when A↑ and B↑ (up-regulated) and C↓ (down-regulated), resulting into its active-state ordered vector <1,1,0> for the corresponding genes. Section 3.2 describes in depth the identification and formation of active sub-paths and the respective data annotation procedure.

III. ***Data Analysis (data mining):*** The binary ordered-vector of each sub-path is aligned and matched against all (discretized) binary gene-expression sample profiles. A sub-path is considered to match a sample if and only if all the corresponding genes in the sub-path exhibit the same active-state in the sample, i.e., genes A, B are up-regulated and gene C is down-regulated, resulting into the corresponding sample ordered-vector <1,1,0>, which matches the sub-path vector. In addition, a binary sub-path expression matrix is formed with rows the sub-paths, columns the input samples and cell-values 1, 0 for the respective sub-path being active for the corresponding sample or not. In other words, the sub-paths are taking the place of sample descriptor features, and are utilized for the construction of sub-path based phenotype prediction models. More details about the data mining procedures, the filtering and the selection of the best sub-paths can be found in section 3.3.

IV. ***Visualization:*** Finally the differential power of each sub-path is computed and appropriate parameterized metrics are implemented (users may adjust them to his/her exploratory needs). The highly ranked (best matching) sub-paths are kept according to user-defined thresholds. Subsequently each sub-path is characterized about its phenotype inclination; sub-paths with positive differential power values are characterized as inclined to phenotype 1 and those with negative power as phenotype 2. The system also identifies the sub-paths that are always active in both phenotypes. More details about the innovative visualization of active gene–to–gene regulatory relations that differentiate between the target phenotypes are presented in section 3.4.

The following sections (3.1, 3.2, 3.3 and 3.4) describe the core steps of the MinePath methodology and the web based user interface. Section 3.5 provides the implementation details for the realization of the MinePath platform and in section 3.6 we introduce implemented extensions of the platform.

3.1. Data pre-processing

Data pre-processing is an important step in the data mining process. Real-world data is often incomplete, inconsistent and/or lacking in certain behaviours or trends and is likely to contain many errors. Data pre-processing is a proven method of resolving such issues especially in the genomics domain where we also face the "*curse of dimensionality*" phenomenon (as discussed in the introduction), where the convergence of any estimator to the true value of a smooth function defined on a space of high dimension is very slow. Furthermore, microarrays are challenging for machine learning methods, since the respective datasets typically have a very large number of features and small number of instances. Learning algorithms are thus confronted with the phenomenon and need to address it in order to be effective.

3.1.1. Microarrays and gene expression data

Microarray technology aims to identify the genes that are expressed in particular cells of an organism at particular time or, at particular conditions (e.g., disease-states or, disease-types). A microarray is typically a glass (or some other material) slide, on to which DNA molecules are attached at fixed locations (spots). There may be tens of thousands of spots on an array, each containing a huge number of identical DNA molecules (or fragments of identical molecules), of lengths from twenty to hundreds of nucleotides. The spots are either printed on the microarrays by a robot, or synthesized by photo-lithography (similarly as in computer chip productions) or by ink-jet printing.

Figure 27 shows the general schema of a microarray experimental set-up. After hybridization and scanning the total mRNA from the samples in two different conditions is extracted and labelled. The final product is a microarray image (the '.tiff' format is followed). Each spot on the array image is identified, its intensity measured and compared to the background (the image quantization process, conducted by dedicated image analysis software). To obtain the final gene-expression matrix from spot quantization, all the quantities related to some gene are combined and the entire matrix is scaled to make different arrays comparable. In the resulted gene-expression matrix, rows represent genes, columns represent samples and each cell contains a number characterizing the expression level of the particular gene in the particular sample.

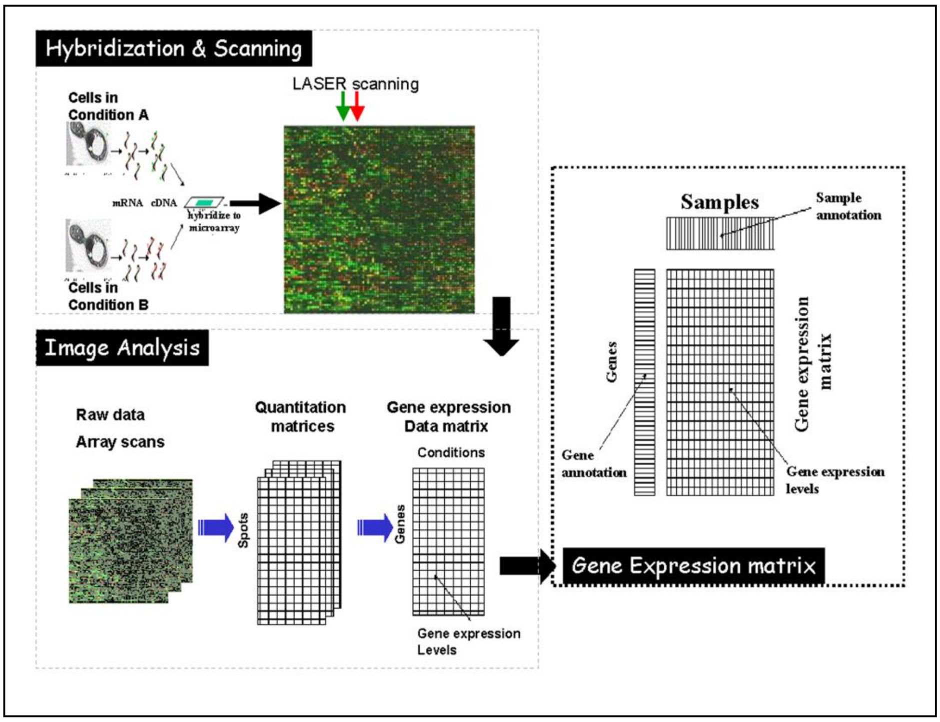


Figure 27: Experimental set-up of Gene Expression Data. The process starts from the hybridization (up left) to image analysis (down left) and the result is the gene expression matrix (right part of the figure)

3.1.1.1. Discretization of gene expression data

In many gene-expression profiling studies the researchers decide to visualize the potential clustering of the genes (or, the samples), as well as the final selected set of genes in a discretized manner. It is known that the predictive accuracy of classifiers improves when gene expression data is discretized [64]. This procedure transforms the expression values of each gene into two or more discrete values making easier the characterization of each gene as "expressed" (or else over expressed, up regulated) or "not expressed" (or else under-expressed, down regulated) for a given sample. Apart from the easier data interpretation, discretization offers some additional benefits as the elimination of the strong influence that causes the outliers coming from incomplete experimental setup. This can lead to more qualitative data analysis [65]. Many extensive studies exist for the discretization of gene expression data such as [66] and [67].

MinePath utilizes discretization of the gene-expression continuous values into the core of the gene-selection process. Discretization of a given gene's expression values means that each value is assigned to an interval of numbers that represents the expression-level of the gene in the given samples. A variable set of such intervals may be utilized and assigned to naturally interpretable values e.g., *low, high*. Given the situation that, in most of the cases, we are confronted with the problem of selecting genes that discriminates between two classes (i.e., disease-states) we believe that it is convenient to follow a two-interval discretization of gene-expression patterns. Below we give a general statement of the discretization problem when two classes are present, followed by an algorithmic process that heuristically solves it. Therefore, expression value represented with 0 indicates a non-expressed or under-expressed gene, whereas value of 1 indicate overexpressed gene. These values are being derived using the following process (shown in Figure 28) in the heart of which resides an information theoretic ranking formula:

i. The expression levels of gene ***A*** over the total number of samples are sorted in descending order.
ii. The midpoints between each two consecutive values are calculated
iii. For each midpoint, the samples are clustered into two subgroups, ***H*** and ***L***.
iv. For each midpoint, an information gain formula is applied, which computes the entropy [68] of the system in respect to its division into subgroups. ***IG(μ***$_{\kappa}$***)*** is the Information Gain of the system for midpoint ***μ***$_{\kappa}$. ***E(L)*** is the total entropy of the system taking into account their prior assignment into classes (ex. case - control), whereas ***E(L/μ***$_{\kappa}$***) = E(H***$_{\kappa}$***,L***$_{\kappa}$***)*** is the entropy of the system taking into account its division into subgroups around midpoint ***μ***$_{\kappa}$.
v. Finally, the midpoint that results in the highest information gain is selected as the best one able to discriminate against the two subgroups and all the samples in the ***H*** group are considered to be overexpressed getting a value of **1**, whereas the ones in the ***L*** group are the non-expressed/under-expressed, getting a value of **0**.

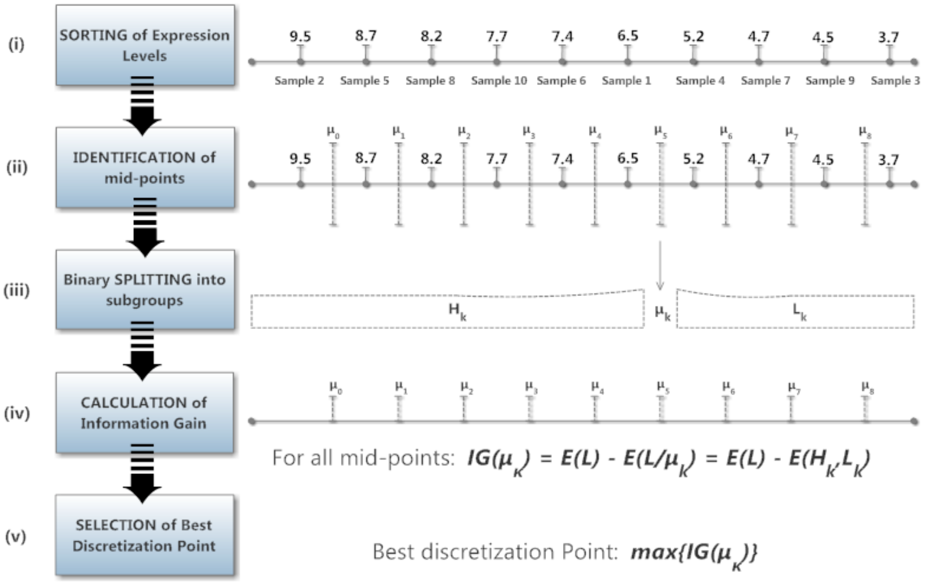
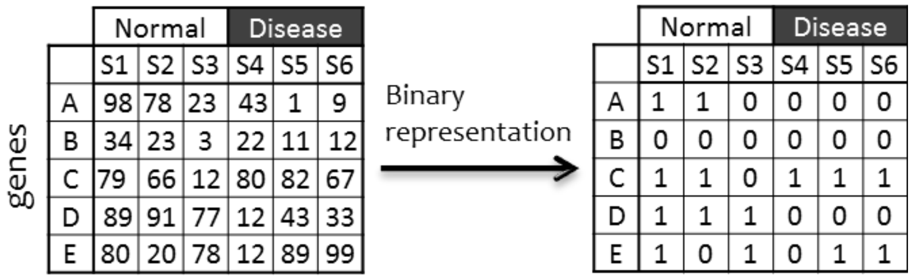

(i)
SORTING of Expression Levels
9.5 8.7 8.2 7.7 7.4 6.5 5.2 4.7 4.5 3.7
Sample 2 Sample 5 Sample 8 Sample 10 Sample 6 Sample 1 Sample 4 Sample 7 Sample 9 Sample 3
(ii)
IDENTIFICATION of mid-points
μ0 μ1 μ2 μ3 μ4 μ5 μ6 μ7 μ8
9.5 8.7 8.2 7.7 7.4 6.5 5.2 4.7 4.5 3.7
(iii)
Binary SPLITTING into subgroups
Hk μk Lk
(iv)
CALCULATION of Information Gain
μ0 μ1 μ2 μ3 μ4 μ5 μ6 μ7 μ8
For all mid-points: IG(μK) = E(L) - E(L/μK) = E(L) - E(HK,LK)
(v)
SELECTION of Best Discretization Point
Best discretization Point: max{IG(μK)}

Figure 28: The Gene Discretization process. The algorithm sorts the expression values of a gene, then identifies the mid-points, splits into sub-groups, calculates the information gain and selects the best split point.

This discretization process is applied to each gene separately and the final dataset is a matrix of discretized gene expression values. A similar approach has been used before in other expression profiling studies [69] [70]. Figure 29 shows an indicative example of a “dummy” microarray with 5 genes (rows) and 6 samples (columns) categorized into two classes, normal and diseased. To the left of the figure we can see the absolute or normalized values of our “dummy” microarray and to the right we have the discretized matrix when we applied the proposed methodology.


genes
Binary representation

	Normal			Disease		
	S1	S2	S3	S4	S5	S6
A	98	78	23	43	1	9
B	34	23	3	22	11	12
C	79	66	12	80	82	67
D	89	91	77	12	43	33
E	80	20	78	12	89	99

	Normal			Disease		
	S1	S2	S3	S4	S5	S6
A	1	1	0	0	0	0
B	0	0	0	0	0	0
C	1	1	0	1	1	1
D	1	1	1	0	0	0
E	1	0	1	0	1	1

Figure 29: Microarray discretization, an indicative example. To the left the gene expression matrix and to the right the discretised gene expression matrix

3.1.2. Gene regulatory networks

The origin of concurrent knowledge about GRNs does not come from any concrete theoretic framework. GRNs are inferred from the biological literature on a given system and represent a distillation of the collective knowledge about a set of related biochemical reactions.

However, although incomplete, this knowledge covers almost every biology function such as metabolism, genetic/environmental information processing, cellular processes, human diseases and drug development, while it is constantly under refinement and enrichment. Online sources of GRN data include KEGG[13], STRING[14] [71], BioCarta[15] [72], ReActome[16] [73], BioPax[17] [74], Pathway Commons[18] [75], just to name few.

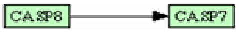
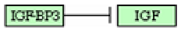
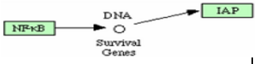
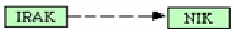
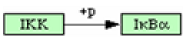
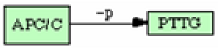
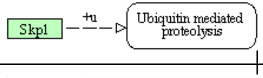
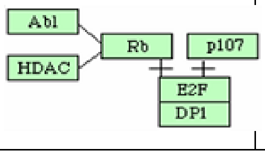
We chose to incorporate KEGG data for our analysis. Since its first introduction in 1995, KEGG DB for pathways has been widely used as a reference knowledge base for understanding biological pathways and functions of cellular processes.

[13] http://www.genome.jp/kegg/ (last day visited 11/08/2014)
[14] http://string-db.org/ (last day visited 11/08/2014)
[15] http://www.biocarta.com/ (last day visited 11/08/2014)
[16] http://www.reactome.org/ (last day visited 11/08/2014)
[17] http://www.biopax.org/ (last day visited 11/08/2014)
[18] http://www.pathwaycommons.org/ (last day visited 11/08/2014)

The knowledge from KEGG has proven of great value by numerous works in a wide range of fields [76].

Although it has been shown that KEGG has some errors [77], these are not so prominent and can be counterbalanced by the simplicity, the variety and the standard ontology that KEGG provides. Through KEGG public database, pathways can be downloaded in KGML format. KGML (stands for KEGG Markup Language) is an exchange format of KEGG graph objects including GRNs. The GRN is described through standard graph annotation. Nodes can be either genes, groups of genes, compounds or other networks. Edges can be one of the gene relations known from the biology theory (activation, inhibition, expression, indirect, phosphorylation, diphosphorylation, ubiquination, association and dissociation). Each gene relation has a different semantic that depicts the precise biology phenomenon that happens during the regulation of the specific network (Table 3).

Table 3: The types of gene interactions and the corresponding gene truth tables; Column relation represents the biological relations in GRNs; Symbol: the KEGG symbol for the relation; Graph representation: an example from KEGG; Truth table: mathematical table used in logic; Semantic: the representation of the relation in pseudocode.

<table>
<tr><th>Relation</th><th>Symbol</th><th>Graph representation in KEGG (examples)</th><th>Truth table</th><th>Semantic</th></tr>
<tr><td>Activation</td><td>$A \rightarrow B$</td><td> </td><td><table><tr><td></td><td></td><td colspan="2">B</td></tr><tr><td></td><td></td><td>ON</td><td>OFF</td></tr><tr><td>A</td><td>ON</td><td>✓</td><td>×</td></tr><tr><td></td><td>OFF</td><td>×</td><td>×</td></tr></table></td><td>B is ON iff A is ON</td></tr>
<tr><td>Inhibition</td><td>A --| B</td><td> </td><td><table><tr><td></td><td></td><td colspan="2">B</td></tr><tr><td></td><td></td><td>ON</td><td>OFF</td></tr><tr><td>A</td><td>ON</td><td>×</td><td>✓</td></tr><tr><td></td><td>OFF</td><td>✓</td><td>×</td></tr></table></td><td>B is OFF iff A is ON <strong><em>OR</em></strong> B is ON iff A is OFF</td></tr>
<tr><td>Expression</td><td>$A \xrightarrow{E} B$</td><td> </td><td colspan="2">Same as activation</td></tr>
<tr><td>Indirect</td><td>$A \xrightarrow{I} B$</td><td> </td><td colspan="2">Same as activation</td></tr>
<tr><td>Phosphorylation</td><td>$A \xrightarrow{+p} B$</td><td> </td><td colspan="2" rowspan="2">In KGML file is stated either as activation or as inhibition</td></tr>
<tr><td>Diphosphorylation</td><td>$A \xrightarrow{-p} B$</td><td> </td></tr>
<tr><td>Ubiquination</td><td>$A \xrightarrow{+u} B$</td><td> </td><td colspan="2">Same as inhibition</td></tr>
<tr><td>Association</td><td>A---B</td><td rowspan="2"> </td><td rowspan="2"><table><tr><td></td><td></td><td colspan="2">B</td></tr><tr><td></td><td></td><td>ON</td><td>OFF</td></tr><tr><td>A</td><td>ON</td><td>✓</td><td>✓</td></tr><tr><td></td><td>OFF</td><td>✓</td><td>✓</td></tr></table></td><td rowspan="2">Physical bonding (nonfunctional)</td></tr>
<tr><td>Dissociation</td><td>A-|-B</td></tr>
</table>

More details about the mapping of the relations within MinePath are described in section 3.2.2 and in

Appendix I (KEGG pathways).

3.1.2.1. Pathway decomposition

MinePath relies on a novel approach for GRN processing that takes into account all possible functional interactions of the network. The different interactions correspond to the different sub-paths that can be followed during the regulation of a target gene.

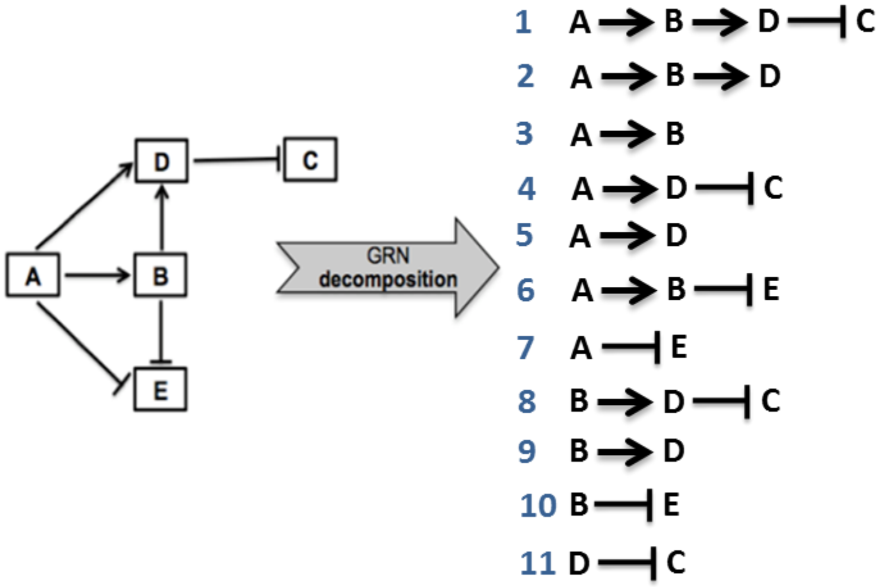


Figure 30: Functional-path decomposition: Left: A target part of an artificial GRN; Right: The eleven decomposed functional sub-paths.

GRNs are downloaded from the KEGG repository. With an XML parser (based on the specifications of KEGG's KGML representation of GRNs) we obtain all the internal network semantics. Even though we use a powerful and open source graph theory library for the processing and the decomposition of the gene regulatory networks, called Cytoscape[19] [78] we had to implement our own parser for the transformation of KGML files to XGMML (format supported by Cytoscape). Solutions like the kgmlreader[20], a Cytoscape plugin for importing KGML files to Cytoscape, could not be used because during the transformation valuable information could be lost (e.g. some edges at metabolic pathways do not have directionality and errors at transforming specific pathways). A description of KGML with the KGML entries and all the possible values can be found at the KEGG Markup Language[21].

[19] http://www.cytoscape.org/ (last day visited 11/08/2014)
[20] https://code.google.com/p/kgmlreader/ (last day visited 11/08/2014)
[21] http://www.kegg.jp/kegg/xml/docs/ (last day visited 11/08/2014)

In a subsequent step, all possible GRN sub-paths are extracted as exemplified in Figure 30. Each sub-path is uniquely annotated as functional according to Kauffman's principles [79] that follow a binary setting: each gene in a functional sub-path can be either 'ON' or 'OFF'. Following the principles reported in [80] the following functional gene regulatory semantics apply.

1. The network is a directed graph with genes (inputs and outputs) is the graph nodes and their directed connecting edges to represent the causal (regulatory) links between them.
2. Each node can be in one of the two states 'ON' or 'OFF'. These states correspond to the gene being expressed (i.e., the respective substance being present) or not expressed, respectively.
3. Time is viewed as proceeding in discrete steps; at each step the new state of a node is a Boolean function of the prior states of the nodes with arrows pointing towards it. Since the directed edge connecting two genes defines explicitly their regulation we can set all possible state-values that a gene may take in a functional sub-path. Thus, each extracted sub-path contains not only the relevant sub-graph but the state-values of the involved genes as well. A sub-path is functional if it is 'active' during the GRN regulation process; in other words we assume that all genes in a sub-path are functionally active.

Furthermore, we extended the MinePath algorithm and can optionally (using a parameter as input) export and take into account the starting and ending points of each sub-path as a new sub-path. This extension proposed by the molecular biology group (Dr. Dimitris Kafetzopoulos) from the Foundation of Research and Technology Hellas (FORTH) Institute of Molecular Biology and Biotechnology (IMBB) and it is based on the limited knowledge (incompleteness) encoded in GRNs. GRNs provide us information about specific sub-pathways between two genes but it is unknown if other pathways/roots connect these two genes. Such an approach could reveal new roots in the GRNs and bypass the limited knowledge of the connection between two genes. Following our example in Figure 30 this parameter of the algorithm will add the following sub-paths: A--|C and B--|C (these extra sub-paths do not appear in Figure 30).

3.1.2.2. Binary representation of regulatory edges

We encode the GRNs as Cytoscape networks using binary representation for the regulatory edges connecting the gene nodes. Cytoscape is freely distributed under the open-source GNU Lesser General Public License, which allows any use of the software, including feature extension by programming[22]. In Cytoscape nodes representing biological entities, such as proteins or genes, are connected with edges representing pairwise interactions, such as experimentally determined protein–protein interactions. Nodes and edges can have associated data attributes describing properties of the protein or interaction.

The main reactions in a gene regulatory network are inhibition, activation, association and disassociation. Table 3 and Figure 31 describe the mapping of GRN network to Cytoscape network with edges encoded in binary format according to the GRN reactions. Expression and indirect reactions are expressed as activations; ubiquination is expressed as inhibition; and phosphorylation/ diphosphorylation reactions are either activation or inhibition (stated in KGML the file).

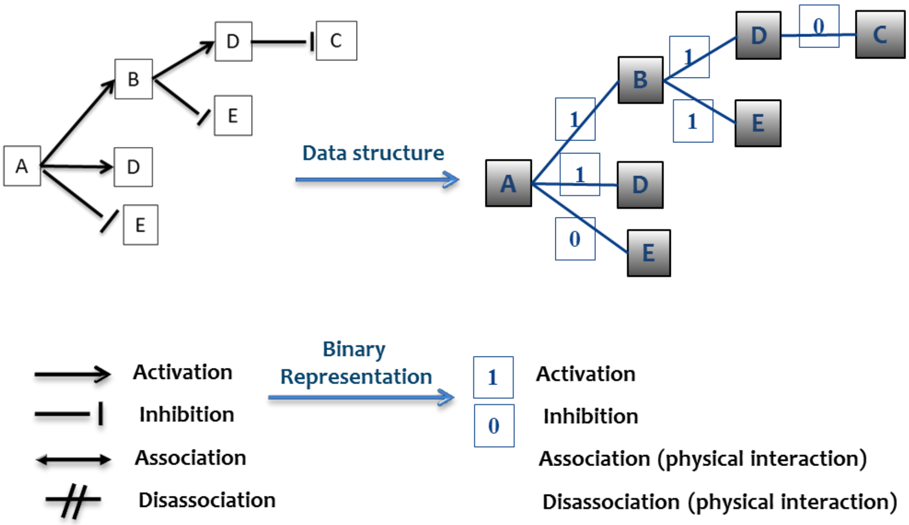


Figure 31: Encoding of GRN reactions to binary edge representation. Activation is represented as an edge with label 1, inhibition as an edge with label 0 and associations/disassociations remain in the graph representation as non-directed interactions which represent a physical interaction between two genes.

Using the pathway decomposition we can retrieve functional paths from a variety of different GRNs (cell-cycle, apoptosis, etc.) and may combine different molecular pathways and networks. Furthermore the binary representation of the network in conjunction with the binary representation of the gene expression data gives us a robust and scalable data structure that can be queried and analysed using machine learning techniques in real time.

3.2. Functional sub-paths and data annotation

MinePath exploits microarray experiments and respective gene-expression data for which the research scientist expects (suspects) that the targeted GRNs play an important role. For example the cell-cycle and apoptosis GRNs play an im-

[22] http://www.gnu.org/licenses/lgpl.html (last day visited 11/08/2014)

portant role in tumour genesis and cancer progression. With an operation that matches gene-expression profiles with sub-paths, the valid and most prominent GRN functional sub-paths are identified. These paths uncover and present potential underlying gene regulatory mechanisms that govern the gene-expression profile of the samples under investigation. Such a discovery may guide the fine classification of samples as well as the re-classification of diseases, based on the most prominent molecular evidence.

The samples of a binary transformed (discretized) gene-expression matrix are matched against targeted molecular pathways and respective GRN functional paths (retrieved form the pathway decomposition).

3.2.1. Probe Sets to Genes

For MinePath the appropriate mapping between the genes identifiers used in the gene expression data to the corresponding KEGG identifiers is needed. Both the GRNs and the gene expression data have to use the same ids. GRNs use gene ids while gene expression platforms use probes. A probe is a specific segment of single-strand DNA that is complementary to a desired gene. For example, if the gene of interest contains the sequence AATGGCACA, then the probe will contain the complementary sequence TTACCGTGT. When added to the appropriate solution, the probe will match and then bind to the gene of interest.

Due to the large number of databases and associated IDs, the conversion of gene identifiers is one of the initial and central steps in many workflows related to genomic data analysis. In the literature and the web we can find several freely available ID conversion tools. Although each tool has distinct features and strengths, as reviewed by Khatri et al [81], they all adopt a common core strategy to systematically map a large number of interesting genes in a list to the associated biological annotation. One of the first online annotation tools in the genomics is the Database for Annotation, Visualization and Integrated Discovery (DAVID[23]) tool [82]. Other online tools that annotate probes to gene IDs are Bablomics[24], DRAGON[25], GeneCruise[26] and AILUN[27] just to name a few. A generic figure highlighting the relations among identifiers is shown in the Figure 32. As we can see the KEGG ids can be annotated through the Entrez Gene IDs.

[23] http://david.abcc.ncifcrf.gov/ (last day visited 11/08/2014)
[24] http://babelomics.bioinfo.cipf.es/ (last day visited 11/08/2014)
[25] http://pevsnerlab.kennedykrieger.org/annotate.htm (last day visited 11/08/2014)
[26] http://genecruiser.broadinstitute.org/genecruiser3/ (last day visited 11/08/2014)
[27] http://ailun.stanford.edu/ (last day visited 11/08/2014)

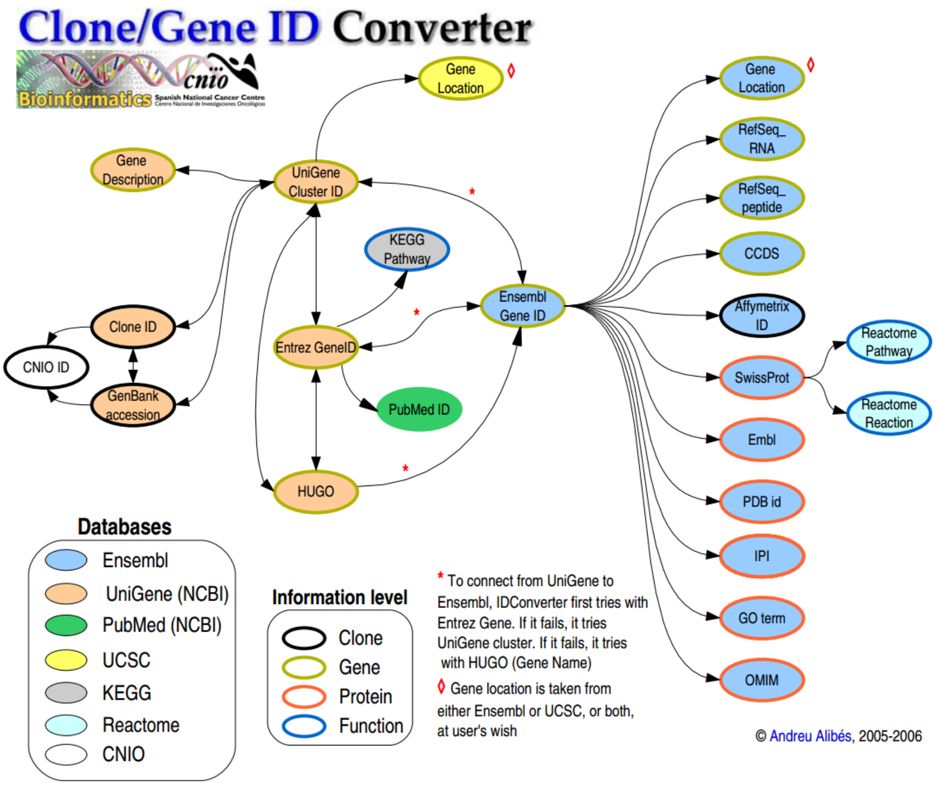
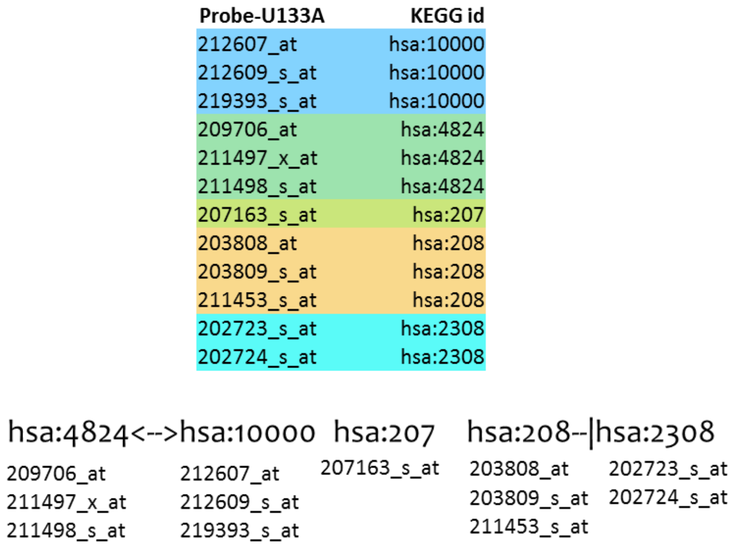


Figure 32: Annotation, relations among Gene identifiers (source http://idconverter.bioinfo.cnio.es/IDConverter.pdf)

For MinePath we use the Bablomics web platform as an offline pre-processing step when the gene-expression data come from platforms that do not support annotation to KEGG IDs or to Entrez IDs. For instance Affymetrix GeneChips[28] provide annotation files for the probes as Entez IDs.

The mapping from a gene nomenclature and thesaurus to another rises the many to one issue where many probes are assigned to the same KEGG gene ID. An indicative example is shown in the upper part of Figure 33 where the gene hsa:1000 is mapped to three Affymetrix probes from the U133A platform and the same holds for the hsa:4824 and hsa:208.

[28] http://www.affymetrix.com/catalog/131455/AFFY/Human-Genome-U133-Plus-2.0-Array#1_3 (last day visited 11/08/2014)



Probe-U133A	KEGG id
212607_at	hsa:10000
212609_s_at	hsa:10000
219393_s_at	hsa:10000
209706_at	hsa:4824
211497_x_at	hsa:4824
211498_s_at	hsa:4824
207163_s_at	hsa:207
203808_at	hsa:208
203809_s_at	hsa:208
211453_s_at	hsa:208
202723_s_at	hsa:2308
202724_s_at	hsa:2308

hsa:4824<-->hsa:10000 hsa:207 hsa:208--|hsa:2308

209706_at	212607_at	207163_s_at	203808_at	202723_s_at
211497_x_at	212609_s_at		203809_s_at	202724_s_at
211498_s_at	219393_s_at		211453_s_at	

Figure 33: Probes to Gene IDs (many to one); on top the mapping of KEGG ids to U133A probe-sets, each colour is one gene assigned to many probes; at the bottom one sub-path and under each gene the corresponding probes

In general, the multiple probes targeting the same gene does not (should not) show different expression levels'. So, taking into account the expression status of just one of the probes is enough. Since we cannot assure the consistency between the different microarray platforms, MinePath provides two options to cope the one to many (probe to gene) issue:

1. ***Max Probe***: This is the default option that checks the multiple probes for the gene and places a logic OR for the assessment of the gene's value. This is actually the selection of the value of the probe with the highest intensity out of all the probes that map to the same gene.
2. ***Probes clones***: The user may optionally set at MinePath to produce all the possible combinations of sub-paths based on probes and not on gene ids. We call this option "probes clones".

Robinson et al [83] proposed that for genes with multiple probe-sets, isoform specific expression changes may be a more appropriate means of interpreting standard microarray expression data than the current one gene = one probe-set paradigm. Going back to the example of Figure 33 we see, at the lower part, a sub-path with two gene interactions and five genes. Under each gene we have the genes mapped to probes for the U133A platform. While in the default option of MinePath this is a single sub-path if we initiate the "probe clones" option MinePath will generate 3*3*1*3*2 = 54 sub-paths with all the possible probe combinations. Table 4 shows the number of sub-paths for 14 KEGG pathways in the default and the "probes clones" options. The "probes clones" have been computed for the U133A Affymetrix probes.

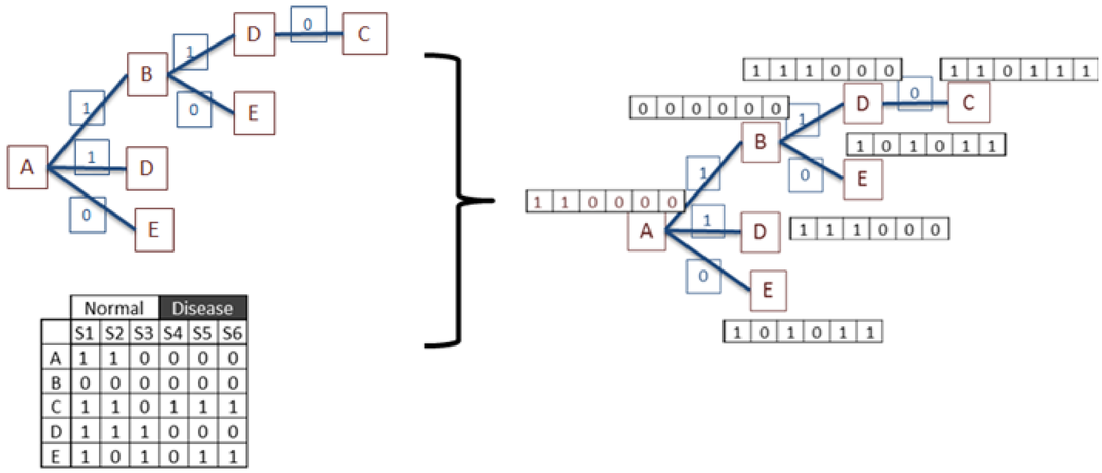
Pathway	Description	Genes in U133A plat.	Sub-paths	Sub-Paths after clones
hsa04010	MAPK signalling	481	1291	21109
hsa04012	ErbB signalling	164	486	4277
hsa04020	Calcium	335	157	189
hsa04110	Cell cycle	231	161	437
hsa04115	p53 signalling	123	277	1939
hsa04150	mTOR signalling	91	65	365
hsa04210	Apoptosis	157	145	1505
hsa04310	Wnt signalling	256	277	371
hsa04350	TGF-beta signalling	140	57	79
hsa04370	VEGF signalling	129	61	187
hsa04510	Focal adhesion	404	420	1275
hsa04520	Adherens junction	179	442	10873
hsa04912	GnRH signalling	205	145	1488
hsa05200	Pathways in cancer	634	988	16014

Table 4: Number of genes and sub-paths for 14 KEGG pathways with and without "probes cloning"

Even though the complexity of the system grows exponentially when we take into account the "probes clones" the system is capable to compute the differential sub-paths without significant delays.

3.2.2. Matching Gene Expression Data and GRNs

We aim to identify the sub-paths that exhibit high matching scores for one of phenotypic class and low matching scores for the others. This is a paradigm shift from the mining of differential genes to the mining of GRN functional sub-paths. The algorithm for differential sub-path identification is inherently simple. We enrich the binary representation of the GRN network with the (binary) data from the discretized microarray data (Figure 34).



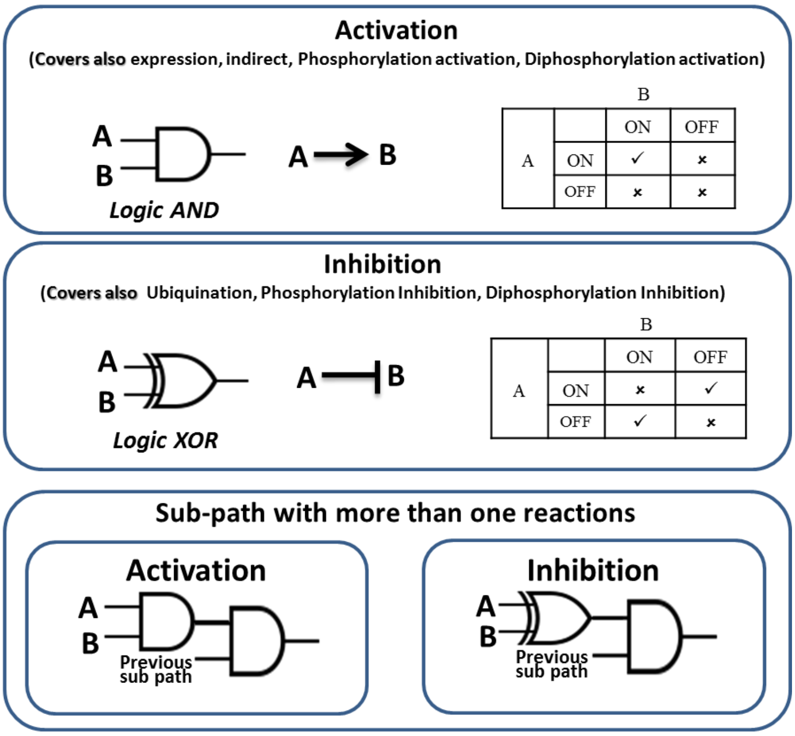
	Normal			Disease		
	S1	S2	S3	S4	S5	S6
A	1	1	0	0	0	0
B	0	0	0	0	0	0
C	1	1	0	1	1	1
D	1	1	1	0	0	0
E	1	0	1	0	1	1

Figure 34: Combining microarray binary data with GRN network; to the left the binary graph representation of the GRN and the discretized microarray data; to the right the discretised gene expression data mapped on the binary graph of the GRN

With such a setup the entire data are in binary format and are stored in a directed graph with binary representation of the relations between the genes (nodes). The candidate sub-paths can be easily extracted from the graph using basic Boolean operations [84] for optimization.

- The activation between two genes A and B can be mapped as a logical AND into their respective microarray data.
- An inhibition between two genes (e.g. A--|B) can be mapped using the logical operation XOR at the microarray data of the target gene (in our case gene B). Figure 35 shows the mapping of the activation and inhibition gene interactions. Table 3 shows the complete mapping of all the supported (by KEGG) gene interactions to the two basic (activation/inhibition) states. A common misunderstanding is that inhibition is functional only when the source gene is up-regulated, e.g. A--| B, A is up-regulated then B is down-regulated. Inhibition is also function when A is down-regulated and B is up-regulated. In the literature we can find such examples [85].
- Association and disassociation are special cases of a gene regulatory network since they does not represent a specific regulatory mechanism between two genes or two group of genes, but a condition in which specific genotypes are associated with other factors, such as specific diseases [86]. In most of the cases, genetic association studies aim to detect association between one or more genetic polymorphisms and a trait, which might be some quantitative characteristic or a discrete attribute or disease. For that reason, MinePath identifies and visualizes the associations and disassociations independently of the gene expression values.

- Computing sub-paths with more than one reaction: When the candidate sub-path has more than one reaction we have to take into account: (i) the last reaction (between the final and pre-final gene or group of genes) and (ii) the resulting binary representation of the previous sub-path (sub-path without the last reaction). The last reaction is combined with logic AND with the previous sub-path to compute the final sub-path vector as shown in Figure 35.


Activation
(Covers also expression, indirect, Phosphorylation activation, Diphosphorylation activation)
A
B
Logic AND
A → B
Inhibition
(Covers also Ubiquination, Phosphorylation Inhibition, Diphosphorylation Inhibition)
A
B
Logic XOR
A ⊣ B
Sub-path with more than one reactions
Activation
A
B
Previous sub path
Inhibition
A
B
Previous sub path

		B	
		ON	OFF
A	ON	✓	×
	OFF	×	×

		B	
		ON	OFF
A	ON	×	✓
	OFF	✓	×

Where:

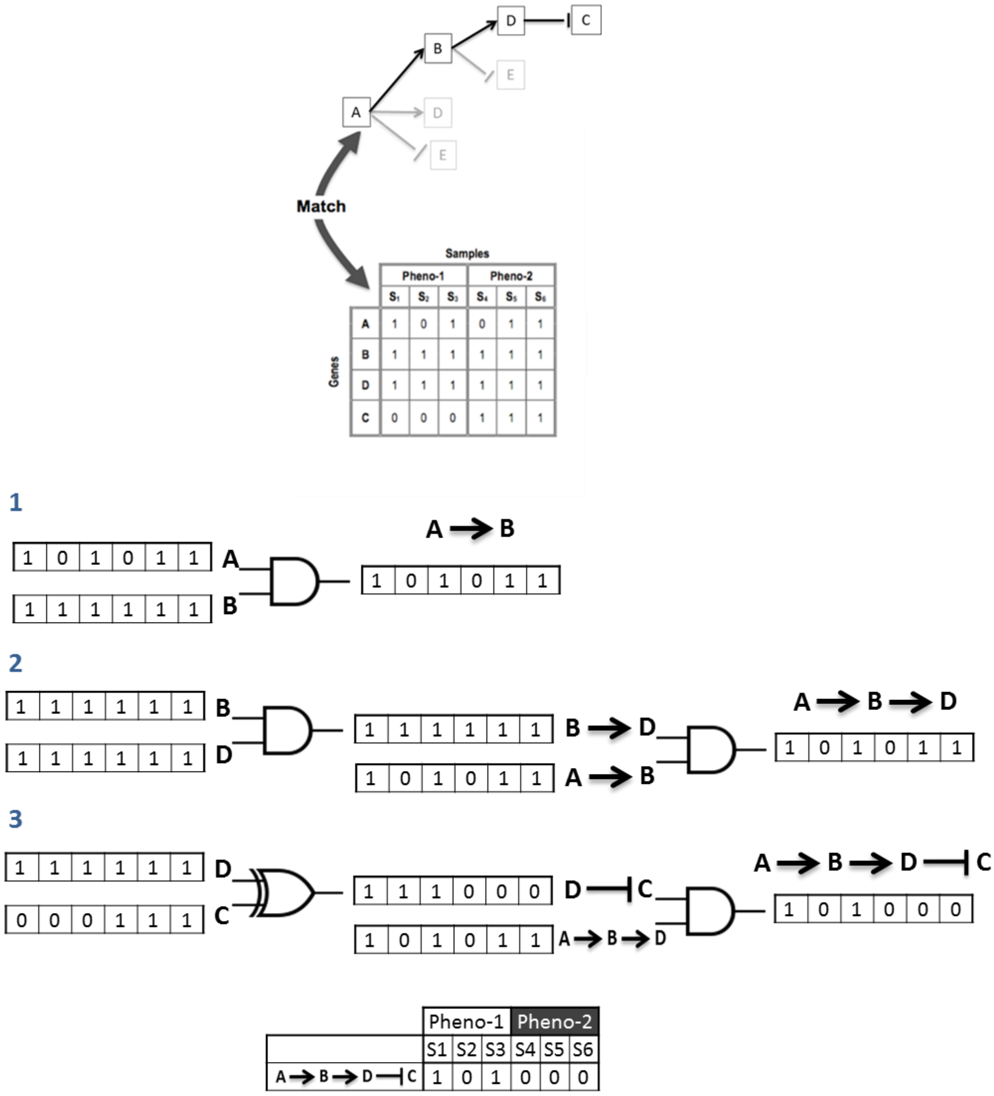
A: source Gene(s)
B: target Gene(s)

Figure 35: Mapping gene interactions using logic gates. Activation mapped as logic AND, inhibition as logic XOR while sub-paths with more than one reaction require the combination of previous sub-path and the last relation using a logic AND

All the possible sub-paths are known using the methodology of the Pathway decomposition (section 3.1.2.1) and the binary tree data structure gives us the needed information (binary representation per gene). With the help of the logical operations the creation of the matrix with the candidate sub-paths per sample can be produced in a fast and optimized way.

Of course the sub-paths do not contain only one gene to gene relation. In most of the cases sub-paths are a chain of reactions (activations or inhibitions) linking many genes or gene groups. The binary operations as described previously are used to map sub-paths that contain more than one relation too. In the case of sub-paths that include multiple activations e.g. A→B→C, MinePath initially computes the A→B (using a logic AND) and the B→C (using a logic AND), then the resulting binary representations are merged using a logic AND, which will give the final binary representation of A→B→C. The same holds for sub-paths that include more than one inhibition in a row, but in that case we use the logic XOR.

A simple example of the Boolean algebra for the identification of the candidate sub-paths is given in Figure 36. As we can see the activation A→B is calculated using the logic gate AND. The same holds for A→B→D where we compute the result of A→B in conjunction (using again a logic AND) with the binary representation of B→D. Then the resulting vector of A→B→D and the vector (binary representation) of D⊣C are combined using logic gate AND to create the vector for our sub-path A→B→D⊣C. To compute the D⊣C we use logic XOR at the binary representation of D and C genes, as described in Figure 35.



	Pheno-1			Pheno-2		
	S_1	S_2	S_3	S_4	S_5	S_6
A	1	0	1	0	1	1
B	1	1	1	1	1	1
D	1	1	1	1	1	1
C	0	0	0	1	1	1

	Pheno-1			Pheno-2		
	S1	S2	S3	S4	S5	S6
A→B→D⊣C	1	0	1	0	0	0

Figure 36: Boolean algebra for the differentially expressed sub-paths, calculation of the A→B→D⊣C. On top the discretized gene expressions and the sub-path; in the middle step by step the calculations for the sub-path; bottom the results of the sub-path for the specific samples

After the decomposition of each pathway into its functional components, each sub-path is matched against the respective samples' gene-expression profiles of the respective microarray studies. The result is an array of sub-paths with binary values for every sample in the form of a discretized microarray.

3.3. Analysis (data mining)

As already exemplified, GRN and gene-expression data matching aims to differentiate GRN sub-paths and identify the most prominent functional sub-paths for the given samples.

The data annotation step of MinePath (section 3.2) produces a binary matrix containing information about the sub-paths (active or not) for the specific samples. This transformation does not aim to reduce the dimensionality issue of microarrays (tens of thousands of genes for tens of samples). In fact the produced matrix (sub-paths & samples) contains more features than the initial gene expression dataset (genes & samples). Let's take an indicative example of a well know microarray platform the Affymetrix U133A[29]. This is a relatively small, in terms of probes, chip supporting 22.283 probes. Using the annotation files provided by Affymetrix we identify 20967 genes in the form of Entrez IDs. The sub-paths that we identify when we decompose all (224 in total) the human (hsa) GRNs from KEGG are more than 50.000. So initially we had a matrix (the gene expression data) with ~22.000 genes per sample and after the transformation we get a matrix with more than 50.000 sub-paths per sample.

Following sections describe the methodologies for the filtering/ranking of the sub-paths and the validation procedure that is based on well-known algorithms from the machine learning area.

3.3.1. Sub-paths selection

Having a dataset with tens of thousands of features (sub-paths per sample) is apparent that a researcher would try to identify the "best" or in our case the most discriminant features (sub-paths). MinePath uses feature selection methodologies for the specific step.

In the literature, we can find a plenitude of feature selection methods, most of them rising as a need to analyse data of very high dimension [87]. This step tries to select the features that best discriminate between the different phenotypes (disease states). The problem is well-known in the machine learning community

[29] http://www.affymetrix.com/estore/browse/products.jsp?productId=131536#1_1 (last day visited 11/08/2014)

as the problem of feature-selection (with its dual ‘feature-elimination’) [88] and various ‘wrapper-based’ [89], or, ‘filtering’ [90], approaches have been proposed.

Traditionally, in machine learning research the number of features, $\boldsymbol{m}$, is quite smaller than the number, $\boldsymbol{k}$, of cases (samples in the case of gene-expression studies) that is, $\boldsymbol{m << k}$. In contrast, gene-expression studies refer to a huge number of features and quite few samples. In most domains the number of sub-paths is in the range of 2.000 – 200.000 (depending on the gene expression platform) and the number of samples in the range of 50 – 200, that is $\boldsymbol{k << m}$. In a situation like that it is questionable if a ‘wrapper’ based feature-selection approach could help, because of its high-computational cost. That is why we follow a ‘filtering’ approach.

The feature selection algorithms for gene expression data, target to identify the most discriminant genes for specific phenotypes. One could see many similarities in the gene selection and sub-paths selection objective. The main difference comes from the handling of the non-expressed sub-paths, which in our case are informative and can be interpreted as non-functional roots in the GRN for a specific phenotype. That type of knowledge is informative and valuable for sub-paths contrary to gene selection approaches where an under-expressed gene means that it is not activated and most of the algorithms ignore it.

For the purposes of MinePath we have implement two different filtering/ranking methodologies, (i) the discriminant ranking and (ii) the polarity ranking. The discriminant is a methodology introduced initially for gene expression data [69] while the polarity has been implemented for MinePath. The user has the option to select these filtering methods and by default the system uses the polarity. The following sections introduce these two methodologies.

3.3.2. Discriminant power

The discriminant power feature selection implementation is based on a ranked-ordering approach. For each sub-path we count the number of samples that it holds or not. Assume the two phenotypic classes ***P*** (positive), ***N*** (negative). The following quantities are computed:

- $\boldsymbol{H_P}$ = number of ***P*** samples that the sub-path holds.
- $\boldsymbol{L_P}$ = number of ***P*** samples that the sub-path does not hold.
- $\boldsymbol{H_N}$ = number of ***N*** samples that the sub-path holds.
- $\boldsymbol{L_N}$ = number of ***N*** samples that the sub-path does not hold.

Formula (1) computes the ***discriminant rank*** for each sub-path (r_{sb}) that measures the power of the sub-path to distinguish between the two classes:

(1) $$r_{sb} = (H_P \times L_N) - (H_N \times L_P)$$

A complete positive sub-path holds for all ***P*** cases and does not hold for any N case i.e., $H_P = P$, $L_N = N$, $L_P = H_N = 0$ and $\boldsymbol{r_{sb}}$ takes its maximum *positive* value *PxN.* In this case the sub-path is considered as descriptive for, or is associated with or, is inclined to class *P*. The sub-path remains completely distinguishing in the inverse case where, $L_P = P$, $H_N = N$, $H_P = L_N = 0$, only that now $\boldsymbol{r_{sb}}$ takes its maximum *negative* value. In this case the sub-path is associated with class *N*. In other words the sub-path ranking formula encompasses and expresses a *differentiation* characteristic that represents the descriptive power of the sub-path with respect to the present phenotypic classes. So, ordering the positive ranks in descending order and the negative ranks in ascending order we may identify the most discriminant sub-path with respect to phenotypic classes *P* and *N*.

3.3.3. Polarity

Since MinePath handles sub-paths instead of genes, a special ranking system able to take into account the absence of a sub-path to the opposite class is needed. As we have already mention the information that a sub-path is non-active (or non-functional) in a specific phenotype is crucial and most of the ranking algorithms devoted to gene selection does not take into account such functionality.

The polarity ranking has been implemented specifically for MinePath and is a two-step filtering procedure. Let's take again the same mapping for the computed quantities:

- $\boldsymbol{H_P}$ = number of ***P*** samples that the sub-path holds.
- $\boldsymbol{L_P}$ = number of ***P*** samples that the sub-path does not hold.
- $\boldsymbol{H_N}$ = number of ***N*** samples that the sub-path holds.
- $\boldsymbol{L_N}$ = number of ***N*** samples that the sub-path does not hold.

Formula (2) computes the ***polarity rank*** for each gene (r_{sb}) that measures the power of the sub-path to distinguish between the two classes:

(2) $$r_{sb} = \frac{(H_P - H_N)}{(H_P + H_N)}$$

The formula provides positive values for sub-paths, which are more informative for class ***P*** and negative values for class ***N***. In addition we apply two extra filters for the polarity ranked sub-paths, even if they get high polarity rank.

- ***First filter for the polarity ranking:*** For the positive ranked sub-paths (derived from formula 2) we keep only the sub-paths that have polarity ranking over the average polarity of the positive sub-paths.

$$\begin{array}{ll} If & r_{sb} > 0\ Average(r_{sb} positive) \\ Else\ if & r_{sb} < 0\ Average(r_{sb} negative) \end{array}$$

- ***Second filter for the polarity ranking:***

$$abs(r_{sb}) \geq \frac{(H_P + H_N)}{(H_P + H_N + L_P + L_N)}$$

With the second filter the system discards highly ranked sub-paths that have quite a few functional cases in the opposite phenotype (e.g. phenotype 2) even if the sub-path is fully functional for the represented phenotype (e.g. phenotype 1).

An indicative example of the polarity filtering is shown in the following figure.

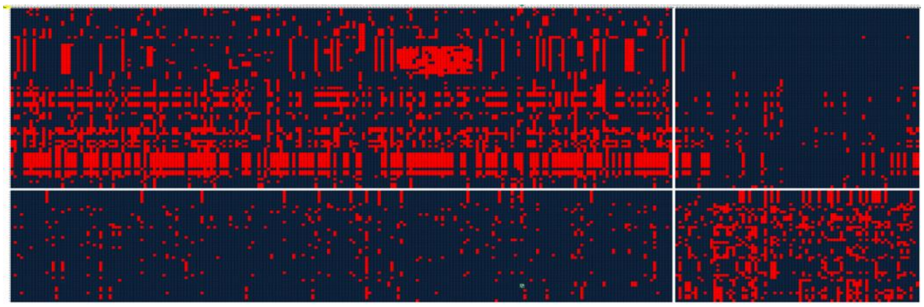


Figure 37: Polarity filtering example. Red represent functional sub-paths and blue non-functional sub-paths, the vertical white line distinguishes the two phenotypes (columns are samples) while the horizontal the best sub-paths for the two phenotypes (rows are sub-paths)

Furthermore, MinePath supported two more variations of the polarity ranking and gives the option to the user to select the best (according to the dataset) ranking method. The two options are:

- ***Relative polarity***: A variation of the polarity ranking formula where instead of absolute counts for the over expressed sub-paths for the phenotypic classes P, N (***H_P, H_N***) we use the percentages of the over expressed per class. For the relative polarity ranking the algorithm scores each sub-path with the following formula:

$$\textbf{(3)} \qquad r_{sb} = \frac{\frac{H_P}{H_P + L_P} - \frac{H_N}{H_N + L_N}}{\frac{H_P}{H_P + L_P} - \frac{H_N}{H_N + L_N}}$$

Formula 3 is valuable for datasets with unbalanced number of samples per class. Instead of the number of over expressed sub-paths (formula 2) we use the percentage of the over expressed sub-paths per class. This formula assures that the unbalanced datasets are not biases to the class containing more samples.

- **Boost true positive**: Another variation of the ranking algorithm, which is applicable only to the polarity filter, is the boost true positive option. The polarity or the relative polarity ranking formula are multiplied by the percentage of the over expressed sub-paths in the respective class. If the ranking formula (polarity or relative polarity) has positive value (the sub-path is associated with class P) then we multiply the score with the percentage of over expressed sub-paths in class P. The same holds for class N. The formula for the boost true positive option is computed as follows:

$$\textbf{(4)} \qquad If \quad r_{sb} > 0 \quad r_{sb} = r_{sb} \times \frac{(H_P)}{(H_P + L_P)}$$

$$else \qquad r_{sb} = r_{sb} \times \frac{(H_N)}{(H_N + L_N)}$$

Where r_{sb} can be the score from formula 2 or formula 3.

The boost true positive variation gives a low ranking "penalty" to sub-paths, which are activated in a small number of samples for the one class and in none or almost none samples in the other. This variation of the polarity will assure that the almost always non-functional sub-paths will be rejected.

3.3.4. Selection of best common sub-paths

Best common sub-paths are the sub-paths that appear to be functional for both phenotypes. Such sub-paths has no informative value in other domains, e.g. when we are handling gene expressions, since a gene that is always up-regulated cannot positively contribute in any research question. In the case of pathways, the sub-paths, which are always activated may fill-in the gap (functional interaction) between two sub-paths and reveal a complete functional and biologically valuable route. Figure 38 highlights the need for the best common sub-paths. The vertical dashed lines distinguish the outer from the inner cell and the grey dashed arrows (down right of the figure) show biological procedures that the paths initiate. Part A (upper part) of figure visualizes the ErBb signalling pathway, where the red lines are sub-paths functional for phenotype-1 and blue lines are functional for phenotype-2. Part B (lower part) of figure, visualizes the same pathway along with the common functional sub-paths (orange relations). As we can see in part B of the figure there are pathways from the extra-cellular (AREG, NRG1 and NRG2), which lead to protein synthesis and metabolism for phenotype-1 and for phenotype-2 from BTC and HBEGF to cell survival and cell cycle progression. These pathways share one or more sub-paths (e.g. the

PIK3R5→AKT3), which are functional for almost all the samples and link the gap between the outer genes to biological procedures.

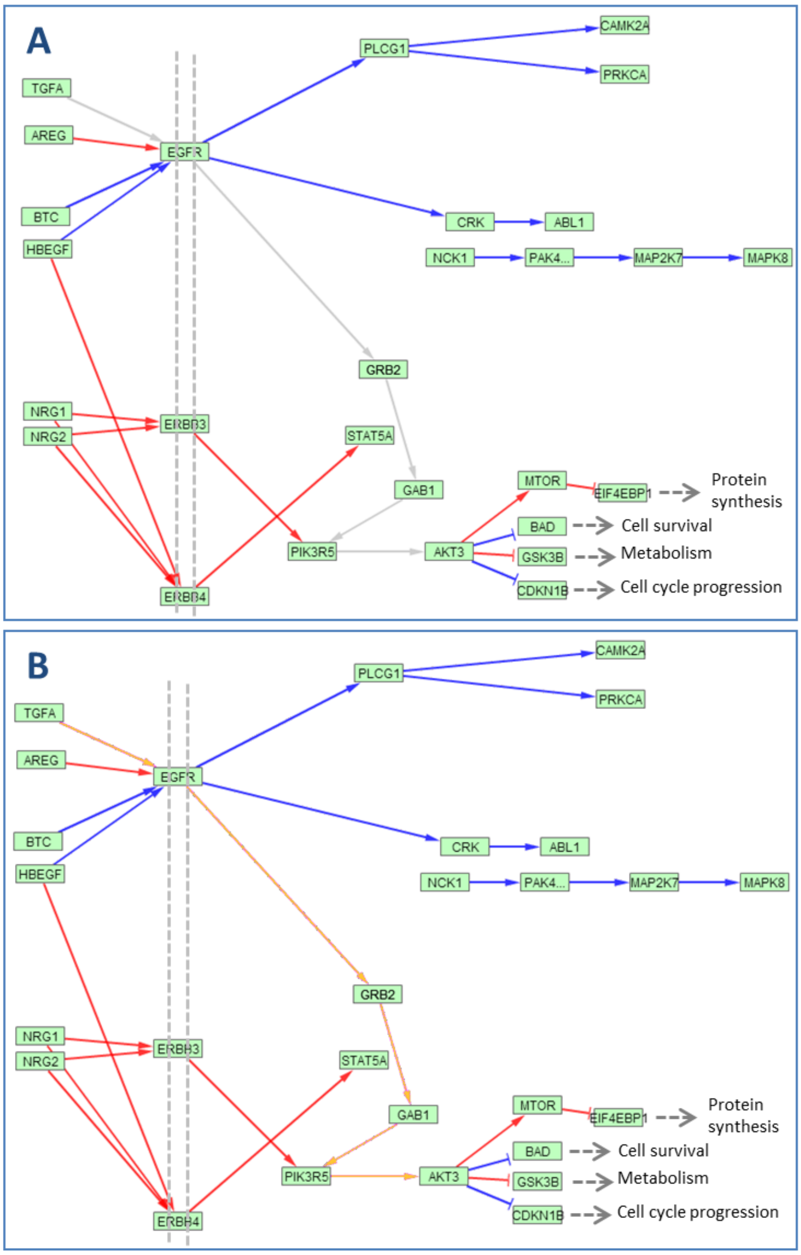


Figure 38: The need for the always functional sub-paths. On top (part A) the pathway with the differentially expressed sub-paths (red for phenotype1, blue for phenotype2), at the bottom (part B) the same pathway with the common sub-paths (orange), The root from the extra-cellular to the biological functions is clear with the common sub-paths

3.3.5. Validation

The main innovations introduced by MinePath come from the matching of the different biological data sources (gene regulatory networks and gene expression data) and system’s visualization capabilities. MinePath provides also mechanisms that validate the best sub-paths against the different phenotypes using well-known algorithms and validation procedures from the area of machine learning. For that reason the open source java library of Weka [91] has been integrated into the system.

Given a set of training samples, each assigned to one of two phenotypic categories, a training algorithm builds a predictive model able to classify new samples into one phenotype or the other. Validation is performed building a non-probabilistic binary linear classifier using randomized 10 fold cross validation procedure. 10 fold means that we divide the data into 10 subsets of (approximately) equal size. We train the classifier 10 times, and measure the respective 'out-of-sample' accuracy performance. Then we measure the accuracy, which is the proportion of true results both true positives and true negatives in the population. The overall accuracy is the measured as the mean of the accuracies achieved in the 10 runs.

MinePath supports three well known machine learning algorithms:

- Decision tree learning (C4.5 [92] software Weka J48). The C4.5 algorithm builds a decision tree from the top, identifying each time the most discriminative variable.
- Naïve Bays [93] software Weka). A simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features.
- Support Vector Machines (Linear kernel support vector machines [94] software Weka SMO) are supervised learning models with associated learning algorithms that analyse data and recognize patterns, used for classification and regression analysis.

By default MinePath computes, stores and shows 10-fold cross-validation results, but additional modelling experiments could be conducted and evaluated (e.g., following a train vs. independent test experimentation mode).

Furthermore a special implementation of the MinePath methodology towards the devise of models that predict disrupted pathways from miRNA's was also implemented. More details about the miRNA case study can be found at the extensions section and the experiments section.

MinePath uses binary data structures and Boolean algebra for the calculations, so that it is capable of operating in real time even on large datasets with hundreds of pathways. As a stress test, all KEGG human ('hsa') pathways (224 in total) were used over an artificial dataset (called '4 ER datasets') that contains gene-expression profiles of 914 samples from 4 different microarray datasets (samples are assigned ER positive or ER negative and all come from the Affymetrix U133A platform). MinePath computed and identified the most discriminant sub-paths in about 2.5'.

3.4. Visualization

In the literature there exist a limited number of systems that utilize knowledge from known GRNs, namely GGEA, SPIA, TEAK and DEAP. However, these systems suffer from insufficient visualization features, a fact that does not facilitate inspection of results and limits the exploratory potential by the users. Some gene set enrichment analysis methodologies and tools utilize pathway visualization approaches to overcome this problem. Since they are based on a gene-oriented approach, they are still unable to handle differentially expressed pathways or even differentially expressed sub-paths.

Solutions such as the KEGG Atlas/Mapper [95], WebGestalt [96], NetworkTrial [97] or even Graphite Web [98] visualize just the pathway genes using some colour scale or colour-coding schema. This problem is apparent even for small pathways such as the inhibition relation A —| B (A inhibits B; A, B represent genes) which could be considered as active in two cases: when A is up-regulated and B down-regulated or when A is down-regulated and B up-regulated. For such different cases, different colours should be assigned to the genes. The situation becomes even more complicated when one has to visualize the phenotype inclination of an interaction, e.g., an inhibition being active for one phenotype and not for another.

Contrary to similar efforts, which visualize the state of genes in a GRN, MinePath identifies and visualizes the differentially expressed GRN sub-paths. In addition, MinePath supports active interaction and re-adjustment of the visualized network and is equipped with special operational features enabling the reduction of GRN's complexity.

One of the key innovations of MinePath rest in its visualization capabilities and especially, in the visualization of active gene to gene regulatory relations that differentiate between the target phenotypes. To the best of our knowledge, MinePath is the only tool that visualizes differentially expressed relations instead of just differential genes. The colour coding of the relations in MinePath is as follows:

- '**Red**' is used to encode sub-path relations that are active for phenotype 1 (Class 1)
- '**Blue**' for relations that are active for phenotype 2 (Class 2)

- ‘**Magenta’** for relations holding for both phenotypes
- ‘**Orange’** for relations that are “always-active”
- “**Yellow**” for the association/disassociation relations
- ‘**Grey**’ for inactive relations.

MinePath also supports active interaction and immediate visualization of pathways when the user sets new thresholds for the best or always active sub-paths. It further supports the option to hide/show the overlapping relations and the association-dissociation relations (in yellow) in the pathway. In all cases, the KEGG layout topology is preserved. In addition, MinePath is equipped with special functionality enabling the reduction of network’s complexity (deletion of genes, relations and/or parts of the network) and re-orientation of its topology. A detailed description of the user interface can be found in section 3.5.2.

3.5. Implementation

3.5.1. Standalone tool

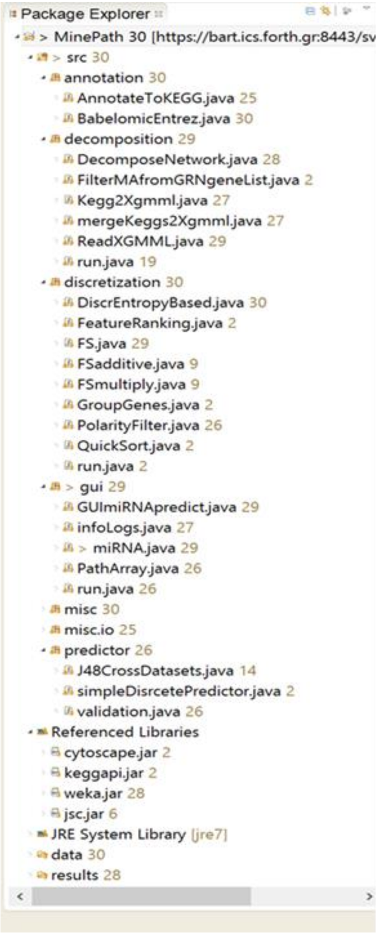
MinePath is a Java based program taking advantage of various libraries.

The structure of the source code can be found in the Figure 39. As we can see the main java packages of MinePath are:

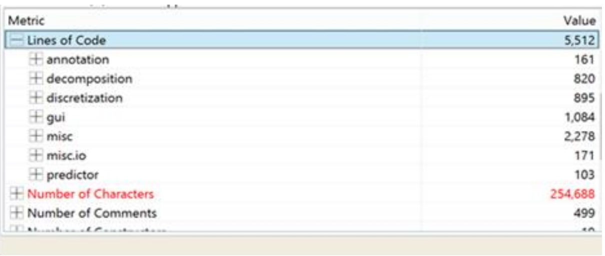
- ***Annotation***: Contains classes related to annotation of genes e.g. from a specific platform to KEGG ids or Entrez Ids.
- ***Decomposition***: Contain classes for the handling and representation of the pathways to our binary graph based data structure.
- ***Discretization***: Contains classes for the discretization of the gene expression data and various filtering classes for the ranking of the extracted sub-paths.
- ***Gui***: The main classes for the invocation of MinePath as standalone tool along with special invocation classes for specific scenarios like the miRNA.
- ***Misc***: This package contains many general purpose classes, which help in various steps the core functionality of MinePath. Same holds for ***misc.io***, which contains classes dedicated to read from and write to intermediate or output files.
- ***Predictor***: The predictor package contains files for the generation of the miRNA prediction models for the one sample prediction scenario.

As we have already mention, MinePath uses open source java libraries for the handling of the graphs and the validation of the best sub-paths based on well-known machine learning algorithms. The libraries used are:

- **Cytoscape**: an open source software platform for visualizing molecular interaction networks and biological pathways and integrating these networks with annotations, gene expression profiles and other state data. Although Cytoscape was originally designed for biological research, now it is a general platform for complex network analysis and visualization. Cytoscape core distribution provides a basic set of features for data integration, analysis and visualization.
- **Weka**: a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules and visualization. It is also well-suited for developing new machine learning schemes.



```
Package Explorer
> MinePath 30 [https://bart.ics.forth.gr:8443/sv
  > src 30
    annotation 30
      AnnotateToKEGG.java 25
      BabelomicEntrez.java 30
    decomposition 29
      DecomposeNetwork.java 28
      FilterMAfromGRNgeneList.java 2
      Kegg2Xgmml.java 27
      mergeKeggs2Xgmml.java 27
      ReadXGMML.java 29
      run.java 19
    discretization 30
      DiscrEntropyBased.java 30
      FeatureRanking.java 2
      FS.java 29
      FSadditive.java 9
      FSmultiply.java 9
      GroupGenes.java 2
      PolarityFilter.java 26
      QuickSort.java 2
      run.java 2
    > gui 29
      GUImiRNApredict.java 29
      infoLogs.java 27
      > miRNA.java 29
      PathArray.java 26
      run.java 26
    misc 30
    misc.io 25
    predictor 26
      J48CrossDatasets.java 14
      simpleDisrcetePredictor.java 2
      validation.java 26
  Referenced Libraries
    cytoscape.jar 2
    keggapi.jar 2
    weka.jar 28
    jsc.jar 6
  JRE System Library [jre7]
  data 30
  results 28
```



Metric	Value
Lines of Code	5,512
annotation	161
decomposition	820
discretization	895
gui	1,084
misc	2,278
misc.io	171
predictor	103
Number of Characters	254,688
Number of Comments	499

Figure 39: Structure & statistics of the source Code. To the left the java packages and the classes, to the right the lines of code per java package

Until now the lines of code of MinePath are more than 5.500. The lines of code per java package are shown in the right part of Figure 39.

The usage of command line MinePath is as follows:

```
MinePath <MicroArray FileName full path> <Pathways Folder Path>

Optional (parameters must appear after microarray & pathways paths):
       -upBoth N (% selected up regulated sub-paths for both classes)     De-
fault 80
       -addSD N (Add SD at the Threshold. Positive value (e.g. 0.5 or 1 or 2)
makes threshold stricter. Negative value makes threshold more elastic)    De-
fault 0
       -addStartEnd (add as subpath the first - last genes of every big (over
2 reactions) sub-path)            Default false
       -i (ignore Paths with only 1 reaction)         Default false
       -ignoreInverseInhibition (ignore inverse inhibition Down --| Up) De-
fault false (use it)
       -filterSimpleRelative (R1-R2)/(R1+R2)                              De-
fault = Polarity filter with relative values
       -filterAbsolute (use absolute values for filtering)  Default = Polar-
ity filter with relative values
       -discr (use discriminant filter)                                   De-
fault = Polarity filter with relative values
       #N (best select for discriminant filter)                           De-
fault = 100
       #-classifier (10 fold cross validation of best subpaths)           1 =
C4.5 decision tree (Default)
              2 NaiveBayes
              3 Support Vector Machines
              0 none

Usage for Kegg conversion to XGMML:
MinePath -kegg2xgmml <Pathways Folder Path>
```

MinePath, either as standalone tool or as the web based platform, provides a wealth of output files. We do not provide only the results and accuracies of the MinePath methodology but we also provide the generated sub-paths to samples matrix, which could be used by statisticians or bioinformaticians to mine the dataset in terms of sub-paths.

Following we describe the output files of the MinePath.

The ***validation.txt*** file is generated only if the user has select to validate the data using one of the 3 available options (10 fold cross validation using decision tree, support vector machines or Bayesian networks). The file provides detailed accuracy and confusion matrix. An example of the validation output file follows.

```
Validation of Best sub-pathways.

Algorithm: Support Vector Machine (10-fold cross validation).


Correctly Classified Instances        11         100 %

Incorrectly Classified Instances      0           0 %

Kappa statistic                       1
```

```
Mean absolute error                    0
Root mean squared error                0
Relative absolute error                0 %
Root relative squared error            0 %
Total Number of Instances               11
=== Detailed Accuracy By Class ===
                         TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                         1       0       1       1       1       1       JKAT1_MUT
                         1       0       1       1       1       1       JKAT1_CTRL
Weighted Avg.        1       0       1       1       1       1
=== Confusion Matrix ===
 A      b   <-- classified as
 3      0 | A = JKAT1_MUT
 0     8 | B = JKAT1_CTRL
```

The ***pathwayStats.xml*** file contains information related to statistics for the dataset such as the min, max, mean and standard deviation for the best sub-paths at each class. This information can be found in the *Experiment_information* tag of the xml. The file also contains statistics for each pathway participated in the experiment and provides information related to the number of genes, the number of sub-paths, and number of sub-paths for each class and for the common sub-paths, percentages and scores. We provide three different score formulas, which can be used to rank the selected pathways (option of the web application of MinePath). The three scores are:

I. Pathway power (***pwA***): is the sum of the significant sub-paths in the pathway (including the common sub-paths) divided by the number of the total sub-paths of the pathway.

II. Pathway discriminant power (***pwDS***): is the number of the significant sub-paths for the two classes divided by the number of the total sub-paths of the pathway.

III. The pathway score (***Score***) is calculated using the formula *Score = pwA * pwDS*

Information for each pathway exists in the xml tags *Pathway.* An example of the pathwayStats.xml follows.

```
<MinePath>
        <Experiment_information>
                <DataSet>GSE18239.txt</DataSet>
                <class1>JKAT1_MUT</class1>
                <class1Mean>0.3703267027541743</class1Mean>
                <class1Std>0.23195214689559882</class1Std>
                <class1Threshold>0.3703267027541743</class1Threshold>
                <class1max>1.0</class1max>
                <class2>JKAT1_CTRL</class2>
                <class2Mean>-0.2775619311429376</class2Mean>
                <class2Std>0.19363961897309606</class2Std>
                <class2Threshold>-0.2775619311429376</class2Threshold>
                <class2min>-0.10000000000000002</class2min>
                <class2max>-1.0</class2max>
                <commonThreshold>0.8</commonThreshold>
        </Experiment_information>
        <Pathway>
                <name>hsa04010.xgmml</name>
                <title>MAPK signalling pathway - Homo sapiens (human)</title>
                <numOfGenes>249</numOfGenes>
                <numOfSubPaths>761</numOfSubPaths>
                <score>0.369</score>
                <pwA>0.515</pwA>
                <pwDS>0.717</pwDS>
                <numOfSubPathsClass1>143</numOfSubPathsClass1>
                <numOfSubPathsOverThrClass1>41</numOfSubPathsOverThrClass1>
                <persOfSubPathsOverThrClass1>5</persOfSubPathsOverThrClass1>
                <numOfSubPathsClass2>588</numOfSubPathsClass2>
                <numOfSubPathsOverThrClass2>240</numOfSubPathsOverThrClass2>
                <persOfSubPathsOverThrClass2>31</persOfSubPathsOverThrClass2>
                <numOfSubPathsCommon>9</numOfSubPathsCommon>
        </Pathway>
        <Pathway>
                <name>hsa04012.xgmml</name>
                <title>ErbB signalling pathway - Homo sapiens (human)</title>
                <numOfGenes>87</numOfGenes>
                <numOfSubPaths>267</numOfSubPaths>
                <score>0.416</score>
                <pwA>0.674</pwA>
                <pwDS>0.617</pwDS>
                <numOfSubPathsClass1>124</numOfSubPathsClass1>
                <numOfSubPathsOverThrClass1>56</numOfSubPathsOverThrClass1>
                <persOfSubPathsOverThrClass1>20</persOfSubPathsOverThrClass1>
                <numOfSubPathsClass2>125</numOfSubPathsClass2>
                <numOfSubPathsOverThrClass2>55</numOfSubPathsOverThrClass2>
                <persOfSubPathsOverThrClass2>20</persOfSubPathsOverThrClass2>
                <numOfSubPathsCommon>1</numOfSubPathsCommon>
        </Pathway>
</ MinePath>
```

The file with the extension *-colours.txt* contains all the participating genes in the selected sub-path with a specific colour coding and the gene ids as Entrez Id for use with the KEGG colour mapper[30]. One can provide this list to the KEGG colour mapper and by selecting *Search against:* **hsa** can view the participating genes (genes from the best sub-paths) in a specific colour coding. That was the initial attempt for visualization in MinePath. Soon enough it was apparent that such tools, which visualize the status of a gene in GRN are not suitable for MinePath since we have to produce a very big and confusing colour mapping for the two phenotypes and the different types of relations in the GRN. The problem comes from the limitation that such tools are based on a gene-oriented approach and are unable to handle differentially expressed pathways or even differentially expressed sub-paths. The situation becomes even more complicated when one has to visualize the phenotype inclination of an interaction, e.g., an inhibition being active for one phenotype and not for another.

The most informative and valuable for further exploratory analysis are the matrices that MinePath produces in tab delimited or arff format. The file with extension:

- *-all-pathways.txt* provide the full matrix (sub-paths vs samples) of the dataset.
- *-all-pathways.txt.arff* is the same matrix in arff format for auto load in the Weka standalone application
- *Ranked-all-pathways.txt* provides again the full matrix with the ranking system of the MinePath (discriminant or polarity metric)
- *Ranked-all-pathways.txt-Best.txt* provides only the best (according to the ranking) sub-paths.
- *Ranked-all-pathways.PlusBest.txt* provides only the best (according to the ranking) sub-paths that characterize the first class of our dataset.
- *Ranked-all-pathways.MinusBest.txt* provides only the best (according to the ranking) sub-paths that characterize the second class of our dataset.
- *Ranked-all-pathways.OrangeBest.txt* provides only the best sub-paths that are always function (in both class) of our dataset.

In the *perPathway* folder the user can find the sub-paths to samples matrix for each pathway individually in tab delimited format with *.txt* extension or arff format (*.arrf* file). The system also provides a *.json* file per pathway, which provides information (and is the input) for the visualization feature of the MinePath web application.

[30] http://www.genome.jp/kegg/tool/map_pathway2.html (last day visited 11/08/2014)

3.5.2. Web based MinePath

The final product of MinePath (www.minepath.org) is a web-based platform that implements the methodology for the identification and visualization of differentially active paths or sub-paths within a gene regulatory network (GRN), using gene-expression data. The platform takes advantage of the regulatory mechanisms and the topology of GRNs, including the direction and the type of the involved interactions (activation/expression, inhibition).

Its core algorithm determines differentially expressed pathway sub-paths and relations instead of just differential genes. These sub-paths present evidential molecular mechanisms that govern the disease itself, its sub-type, state or other targeted disease phenotypes. In this form, MinePath introduces a new and efficient representation of the differentially expressed sub-paths over a Web-based human-computer interface. Furthermore, MinePath supports live interaction, immediate visualization of regulatory relations and it is equipped with special topological and network-adjustment functionalities.

The MinePath web-server is implemented as a Web 2.0 application. It relies on the frontend-backend software design using AJAX calls for the communication. The layout, appearance and interface of the front-end are based on the open source version of Ext-JS[31] library and pure JavaScript. For visualization and interaction of the differential GRN sub-paths the Cytoscape Web[32] library has been deployed and expanded. The backend of MinePath is a java-based application and takes advantage of the Weka[33] API for the implementation and evaluation of phenotype prediction models.

Use of MinePath is relatively simple and straightforward. The user selects or uploads a microarray dataset, then selects the gene regulatory networks to explore and run MinePath. The system will compute in real time the differentially expressed functional paths or sub-paths of the selected pathways. Then the user selects the pathway to explore and the system visualizes the differentially expressed regulatory mechanisms (relations) and sub-pathways. The complete list of operations (in steps) follows.

[31] www.sencha.com/products/extjs/

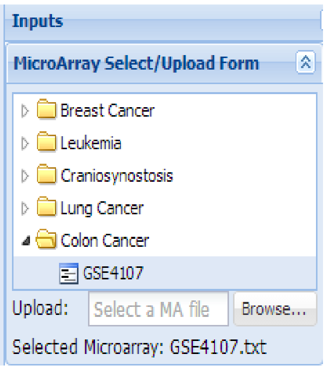
[32] http://cytoscapeweb.cytoscape.org (last day visited 11/08/2014)

[33] www.cs.waikato.ac.nz/ml/weka (last day visited 11/08/2014)

1. **Select** input data and parameters
    - 1.1. Select gene expression dataset
    - 1.2. Select the gene regulatory networks to be analyzed (by default 14 cancer related pathways are pre-selected)
    - 1.3. Select thresholds for differentially expressed functional sub-paths, the subpath ranking method and the validation algorithm (default values are preset)
2. **Run** MinePath
    - 2.1. View and download results (best sub-paths arff & tab delimited files)
    - 2.2. View ranked pathways and select which to visualize.
3. **Visualize/explore** the selected pathway using the web-based interface
    - 3.1. From the controls panel (left panel of the visualization)
        - 3.1.1. Set/change dynamically threshold for class1 (phenotype 1)
        - 3.1.2. Set/change dynamically threshold for class2 (phenotype 2)
        - 3.1.3. Set/change dynamically threshold for always active sub-pathways
        - 3.1.4. Show/Hide associations and dissociations
    - 3.2. In the viewer using right click
        - 3.2.1. Remove inactive genes
        - 3.2.2. Remove inactive relations
        - 3.2.3. Remove selected genes/relations.
        - 3.2.4. Change the layout (random) topology of the network

3.5.2.1.Select input data and parameters

3.5.2.1.1. Select or upload gene expression dataset



Inputs

MicroArray Select/Upload Form

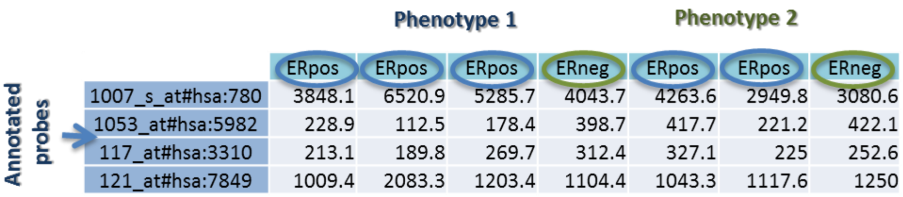
- Breast Cancer
- Leukemia
- Craniosynostosis
- Lung Cancer
- Colon Cancer
    - GSE4107

Upload: Select a MA file Browse...

Selected Microarray: GSE4107.txt

MinePath uses microarray experiments and respective gene-expression data for which we expect (suspect) the targeted GRNs play an important role. MinePath (currently) provides 12 public gene expression datasets from the Gene Expression Omnibus (GEO) database. The user can select one of the 12 annotated datasets or upload his/her own dataset. The uploaded dataset is private, viewable just by the uploaded (the uploaded data are deleted as soon as the processing of MinePath ends). The uploaded dataset should be in the form of a tab delimited txt file where the rows are the annotated gene names and the columns are the sample values. The annotated gene names must have the format **<probe>#<KeggID>** (keggID is identical to the corresponding Entez ID with prefix for each species, e.g. for human "***hsa:***"). For example if you have the probe "**1007_s_at**" for human, which maps to the "**780**" Entez ID, then the annotated gene name for MinePath must be

"**1007_s_at#hsa:780**". The first row contains the phenotype of each sample. At the web site of MinePath annotation lists for the most common microarray can be found, downloaded and used from the following links (U133A, U133B, U133plus). The phenotype must be one word without white spaces and the system expects 2 phenotypes (2 different classes). MinePath supports nominal (dot as decimal separator) or binary (0,1) values. A sample of an input data-file is shown at the figure below.



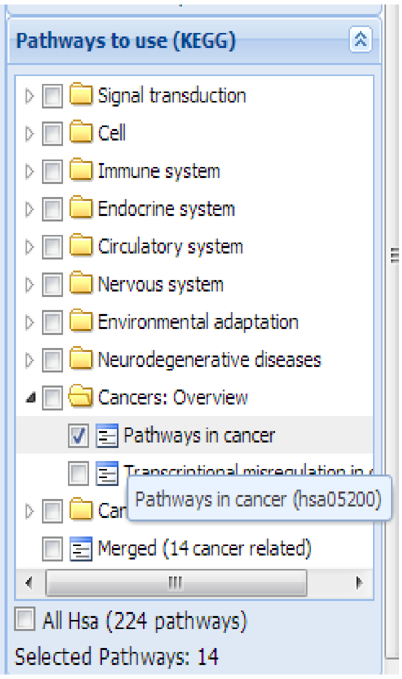
Phenotype 1 | Phenotype 2

Annotated probes

	ERpos	ERpos	ERpos	ERneg	ERpos	ERpos	ERneg
1007_s_at#hsa:780	3848.1	6520.9	5285.7	4043.7	4263.6	2949.8	3080.6
1053_at#hsa:5982	228.9	112.5	178.4	398.7	417.7	221.2	422.1
117_at#hsa:3310	213.1	189.8	269.7	312.4	327.1	225	252.6
121_at#hsa:7849	1009.4	2083.3	1203.4	1104.4	1043.3	1117.6	1250

3.5.2.1.2. Select pathways

Current version of MinePath supports all the human (hsa) related gene regulatory networks from the KEGG database.



Pathways to use (KEGG)

- □ Signal transduction
- □ Cell
- □ Immune system
- □ Endocrine system
- □ Circulatory system
- □ Nervous system
- □ Environmental adaptation
- □ Neurodegenerative diseases
- □ Cancers: Overview
  - ☑ Pathways in cancer
  - □ Transcriptional misregulation in c
- □ Car
- □ Merged (14 cancer related)

Pathways in cancer (hsa05200)

□ All Hsa (224 pathways)

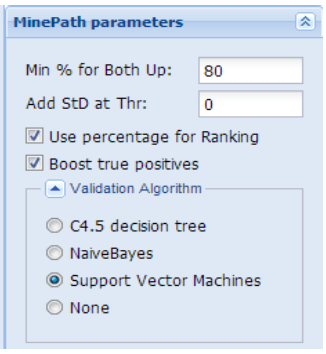
Selected Pathways: 14

By default the system has preselected 14 hsa cancer-related pathways. The preselected pathways are: ECM-receptor interaction (hsa04512), Cytocin-cytocin receptor interaction (hsa04060), Adherens junction (hsa04520), Wnt signalling (hsa04310), Focal adhesion (hsa04510), Jak-STAT signalling (hsa04630), ErbB signalling (hsa04012), MAPK signalling (hsa04010), mTOR signalling (hsa04150), VEGF signalling (hsa04370), Apoptosis (hsa04210), p53 signalling (hsa04115), Cell cycle (hsa04110) and TGF-β signalling (hsa04350). All these pathways are engaged with the 'Pathways in Cancer' integrated pathway of KEGG (hsa05200).

The user can add or remove any pathways by selecting/unselecting from the "Pathways to use" tree view menu, or use all the hsa pathways by selecting the "All hsa (224 pathways)" option.

We have also created an artificial pathway, which is the merged pathway of the 14 cancer related pathways and can be found in the pathways tree as "Merged (14 cancer related)". For more information please refer to the Extensions section.

3.5.2.2. Run MinePath



MinePath parameters

Min % for Both Up: 80

Add StD at Thr: 0

☑ Use percentage for Ranking

☑ Boost true positives

Validation Algorithm

○ C4.5 decision tree

○ NaiveBayes

● Support Vector Machines

○ None

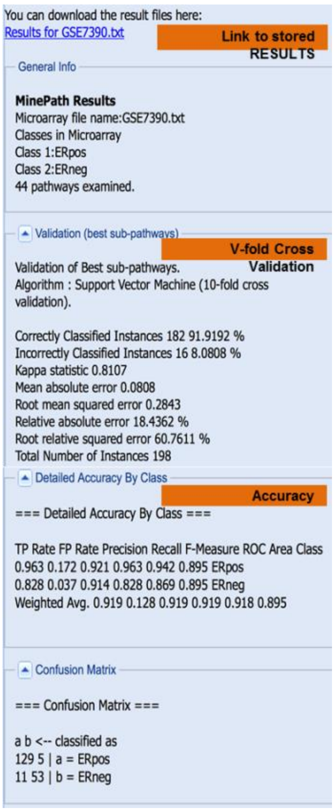
The user can also optionally set some parameters regarding the minimum thresholds of the two sub-paths phenotypes, the minimum threshold for the always active sub-paths, variations for the ranking algorithm and select validation algorithm as shown at the figure to the left.

3.5.2.3. Selected sub-paths validation

MinePath validates the best (over a threshold) sub-paths using 10 fold cross validation methodology over the selected validation algorithm that can be:

- Decision Tree
- Naïve Bayes or
- Support Vector Machines (default option).

The phenotype information is extracted from microarrays and all the selected GRNs are evaluated for the identification of the most informative GRNs at the specific phenotype. The efficient ranking of sub-paths provides the most differentiating and prominent GRN functional sub-paths for the respective target phenotypes. These sub-paths present evidential molecular mechanisms that govern the disease itself, its type, its state or other targeted disease phenotypes (e.g., positive or negative response to specific drug treatment). The results are shown to the user, as soon as the algorithm finishes, along with the option to download the result files (as shown at the figure to the left).



You can download the result files here:
Results for GSE7390.txt

General Info

```
MinePath Results
Microarray file name:GSE7390.txt
Classes in Microarray
Class 1:ERpos
Class 2:ERneg
44 pathways examined.
```

Validation (best sub-pathways)

```
Validation of Best sub-pathways.
Algorithm : Support Vector Machine (10-fold cross
validation).

Correctly Classified Instances 182 91.9192 %
Incorrectly Classified Instances 16 8.0808 %
Kappa statistic 0.8107
Mean absolute error 0.0808
Root mean squared error 0.2843
Relative absolute error 18.4362 %
Root relative squared error 60.7611 %
Total Number of Instances 198
```

Detailed Accuracy By Class

```
=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure ROC Area Class
0.963 0.172 0.921 0.963 0.942 0.895 ERpos
0.828 0.037 0.914 0.828 0.869 0.895 ERneg
Weighted Avg. 0.919 0.128 0.919 0.919 0.918 0.895
```

Confusion Matrix

```
=== Confusion Matrix ===

a b <-- classified as
129 5 | a = ERpos
11 53 | b = ERneg
```

At the downloadable results the user can find overall statistics, Weka (.arff) files for all the sub-paths along with binary values per sample, the best overall pathways sub-paths and best sub-paths per pathway - formed to enable the application of a variety of mining tasks; induction of different predictors (e.g., decision-trees, SVMs etc.); and application of different prediction performance experiments (e.g., on independent datasets). By default the web based MinePath validates the results using 10-fold cross validation.

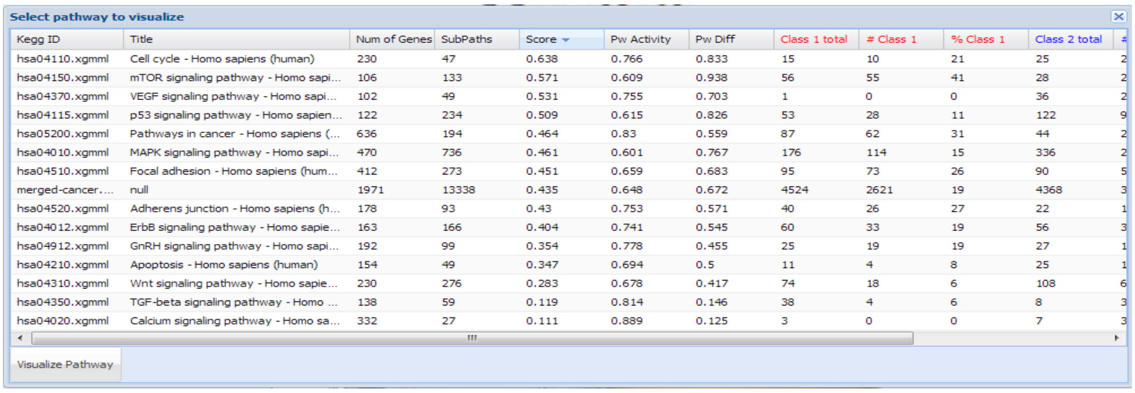
Validation of independent datasets is as simple as follows. Run Minepath for test and train datasets (must be from the same microarray platform and with the same phenotypes). Download the results and use the ***best sub-paths*** .arff file of train dataset ***as train*** in weka. Then select the ***all sub-paths*** .arff file of the downloaded test dataset ***as test*** at weka.

3.5.2.4. GRNs statistics

At the next step, MinePath shows the list of the involved pathways ranked along with statistics that helps to select which pathway to visualize. The statistics per pathway are:

- number of genes
- number of sub-pathways
- the MinePath score
- coverage score
- differential power
- statistics for sub-paths in class 1 (phenotype 1)
- statistics for sub-paths in class 2 (phenotype 2)
- number of always active sub-paths

An example is shown at the figure below.

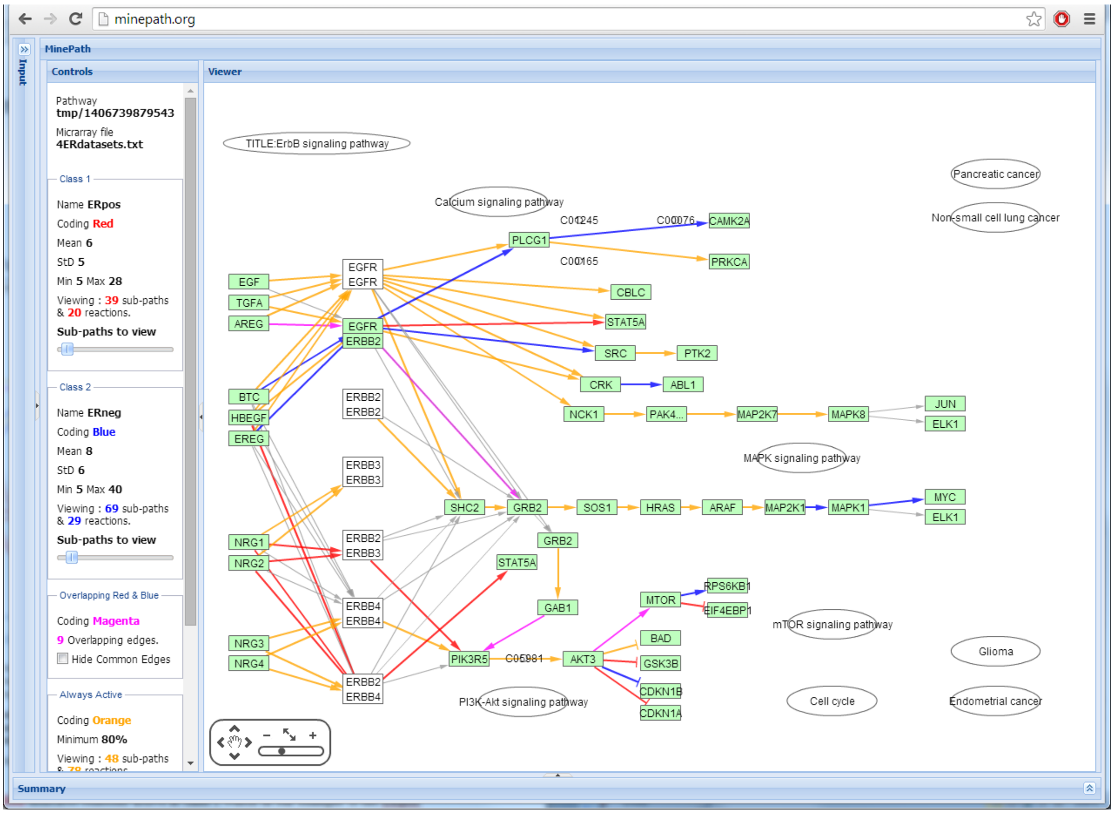


Kegg ID	Title	Num of Genes	SubPaths	Score	Pw Activity	Pw Diff	Class 1 total	# Class 1	% Class 1	Class 2 total	#
hsa04110.xgmml	Cell cycle - Homo sapiens (human)	230	47	0.638	0.766	0.833	15	10	21	25	2
hsa04150.xgmml	mTOR signaling pathway - Homo sapi...	106	133	0.571	0.609	0.938	56	55	41	28	2
hsa04370.xgmml	VEGF signaling pathway - Homo sapi...	102	49	0.531	0.755	0.703	1	0	0	36	2
hsa04115.xgmml	p53 signaling pathway - Homo sapien...	122	234	0.509	0.615	0.826	53	28	11	122	9
hsa05200.xgmml	Pathways in cancer - Homo sapiens (...	636	194	0.464	0.83	0.559	87	62	31	44	2
hsa04010.xgmml	MAPK signaling pathway - Homo sapi...	470	736	0.461	0.601	0.767	176	114	15	336	2
hsa04510.xgmml	Focal adhesion - Homo sapiens (hum...	412	273	0.451	0.659	0.683	95	73	26	90	5
merged-cancer....	null	1971	13338	0.435	0.648	0.672	4524	2621	19	4368	3
hsa04520.xgmml	Adherens junction - Homo sapiens (h...	178	93	0.43	0.753	0.571	40	26	27	22	1
hsa04012.xgmml	ErbB signaling pathway - Homo sapie...	163	166	0.404	0.741	0.545	60	33	19	56	3
hsa04912.xgmml	GnRH signaling pathway - Homo sapi...	192	99	0.354	0.778	0.455	25	19	19	27	1
hsa04210.xgmml	Apoptosis - Homo sapiens (human)	154	49	0.347	0.694	0.5	11	4	8	25	1
hsa04310.xgmml	Wnt signaling pathway - Homo sapie...	230	276	0.283	0.678	0.417	74	18	6	108	6
hsa04350.xgmml	TGF-beta signaling pathway - Homo ...	138	59	0.119	0.814	0.146	38	4	6	8	3
hsa04020.xgmml	Calcium signaling pathway - Homo sa...	332	27	0.111	0.889	0.125	3	0	0	7	3

Figure 40: Selection of pathway to visualize, the GUI provides statistics for each pathway such as number of genes, number of sub-pathways and various scores. The user can also short the results based on any of these categories

3.5.2.5. Visualize/explore

The user can select any of the pathways to explore/visualize. An example of the ErbB pathway for the ‘4ERdatasets’ dataset (a set of four independent discretized and then merged gene-expression studies targeting the ER phenotypic status of the respective patients, from the four studies are naming GSE2034, GSE2990, GSE3494 and GSE7390) using the 14 preselected pathways is shown at the figure below. We also deleted the inactive gene interactions (in the viewer - right click feature).


minepath.org

MinePath

Controls

Pathway
tmp/1406739879543
Micrarray file
4ERdatasets.txt

Class 1
Name ERpos
Coding Red
Mean 6
StD 5
Min 5 Max 28
Viewing : 39 sub-paths & 20 reactions.
Sub-paths to view

Class 2
Name ERneg
Coding Blue
Mean 8
StD 6
Min 5 Max 40
Viewing : 69 sub-paths & 29 reactions.
Sub-paths to view

Overlapping Red & Blue
Coding Magenta
9 Overlapping edges.
□ Hide Common Edges

Always Active
Coding Orange
Minimum 80%
Viewing : 48 sub-paths & 78 reactions

Summary

Figure 41: Indicative example of MinePath visualization. To the left are the controls of the MinePath visualization tool. To the right is the viewer where we see the pathway (with the KEGG topology) and red edges represent functional sub-paths for phenotype1 (in this case ER+), blue for phenotype 2 (ER-), orange always active sub-paths, magenta overlapping functional sub-paths and grey non functional sub-paths

→ ACTIVATION/EXPRESSION for class-1(ERpos)
→ ACTIVATION/EXPRESSION for class-2(ERneg)
→ ACTIVATION/EXPRESSION for both (ERpos, ERneg)

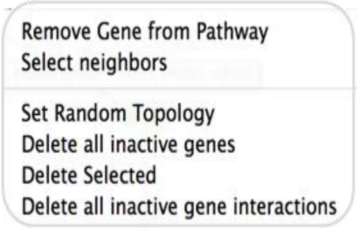
⊣ INHIBITION for class-1 (ERpos)
⊣ INHIBITION for class-2 (ERneg)
⊣ INHIBITION for both (ERpos, ERneg)

→ ACTIVATION/EXPRESSION active in both classes

The graph preserves the KEGG layout topology. It is enriched with the expressed regulatory mechanisms (relations) between genes that differentiate between the two phenotypes:

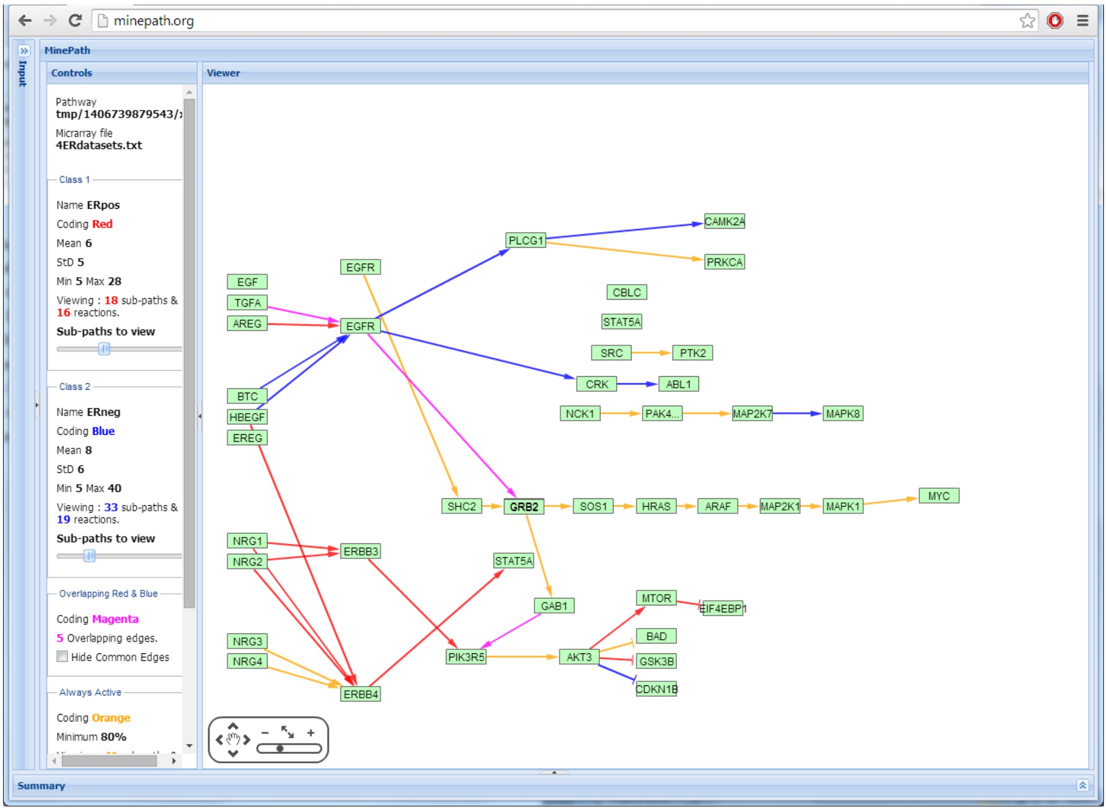
- Red indicates relations active at class 1, which in our example is the ERpos
- Blue indicates relations active at class 2 (ERneg)
- Magenta indicates overlapping relations in the two classes
- Orange for sub-paths that are always active.

Contrary to other pathway visualization tools, MinePath calculates and visualizes differentially expressed relations instead of just differential genes. Furthermore MinePath supports active interaction and immediate visualization when the end user sets new thresholds for the two phenotypes or for the always active sub-paths, as well as to hide/show the overlapping relations and hide/show the association-dissociations of the pathway from ***the control panel*** (left part of MinePath viewer).



In addition, MinePath is equipped with special functionality that enables the reduction of network's complexity (deletion of genes, relations and/or parts of the network), as well as re-orientation of its topology. The functionality is available with *a right click (in the viewer)*.

Using the aforementioned example and exploring the specific pathway we can stress the thresholds to retain '*strong*' sub-paths per phenotype (class): Using **13** as threshold for class 1 (ERpos) results to **18** sub-paths and again **13** for class 2 (ERneg) results to **33** sub-paths; We also use **95%** for all always active sub-paths, which results to **45**. Then, by right clicking at the viewer we delete all inactive genes, delete all inactive gene interactions and merge the 2 GRB2 gene-rectangles (GRB2 appears 2 times in ErbB due to the topology of KEGG). The resulting (reduced) pathway will become as the one in the following figure (moving around gene-rectangles and relation-edges we made its layout 'prettier').



Pathway
tmp/1406739879543/

Micrarray file
4ERdatasets.txt

Class 1

Name ERpos
Coding Red
Mean 6
StD 5
Min 5 Max 28
Viewing : 18 sub-paths & 16 reactions.
Sub-paths to view

Class 2

Name ERneg
Coding Blue
Mean 8
StD 6
Min 5 Max 40
Viewing : 33 sub-paths & 19 reactions.
Sub-paths to view

Overlapping Red & Blue

Coding Magenta
5 Overlapping edges.
□ Hide Common Edges

Always Active

Coding Orange
Minimum 80%

Figure 42: Using MinePath controls over a GRN. Thresholds 13 for class 1 (ERpos), 13 for class 2 (ERneg), 95% for always active sub-paths and deleted all inactive genes and gene interactions

Inspecting the reduced network, it is clear that there is a pathway starting from NRG (1 and 2) and ends at inhibiting the GSK3B and EIF4EBP1 for ERpos phenotype; and a pathway starting from TGFA or BTC or HBEFG that ends-up at inhibiting BAD and CDKN1B for ERneg phenotype. You can see that these sub-pathways share common parts, which are active at both phenotypes (ERpos and ERneg). More details for the established clinico-genomic information and knowledge that supports the finding need for pan-erbb inhibitors can be found in the experiments section.

Armed with the aforementioned features, MinePath serves the users' exploratory needs to reveal the regulatory mechanisms that underlie and putatively govern the expression of target phenotypes.

3.6. Extensions

MinePath has been implemented to be modular and to be easily extended to support more algorithms (e.g. discretization algorithms, filtering algorithms and validation algorithms) and different clinical scenarios or research questions.

3.6.1. miRNAs to disrupted sub-paths

Such a need came from a European Union funded research project called P-Medicine[34]. MinePath was demonstrated to the consortium, feedback was very good and the project coordinator (prof. Norbert Graf) asked if we could use and extend MinePath to identify disrupted sub-paths from GRNs using only miRNA data.

To our knowledge such a tool, which will be able to identify disrupted sub-paths from miRNA data, does not exist in the literature. Similar tools such as the GeneTrial[35] or the mirPath[36] use ORA and measure the disruption of the pathway as whole and not specific sub-paths in the pathway.

The research question for the miRNA extension was:

- *To find disrupted pathways in nephroblastoma using miRNA expression data.*

miRNA data from nephroblastoma serve as the source of disrupted metabolic pathways. These data needs to be normalized and then correlated to pathway

[34] http://p-medicine.eu/ (last day visited 11/08/2014)

[35] http://genetrail.bioinf.uni-sb.de (last day visited 11/08/2014)

[36] http://diana.imis.athena-innovation.gr/DianaTools/index.php?r=mirpath/index (last day visited 11/08/2014)

data coming from the KEEG pathway database. MinePath will analyse the tumour of disrupted metabolic pathways. By correlation to clinical data of patients, individual pathway disruptions or main disruptions for a cohort of patients with nephroblastoma will be produced as a result. The tool should be made in a general way that by describing the databases and the interfaces the tool will get domain independent.

The need to use miRNAs instead of gene expression data in such a scenario is essential in the clinical practice since miRNA exams can be produced fast, with blood sample the first day of the patient in the hospital. With such a tool we could possibly get personalized insights in the clinical routine since we could categorize the new patient to responsive or non-responsive of a possible treatment, prior the treatment.

For the miRNA scenario we assume that all the KEGG pathways are fully functional. Disrupted pathways will be the pathways that are not "active" according to the specific cohort (microRNAs for nephroblastoma or gene expressions for ALL).

The idea is that miRNAs have known targeted genes, which means that an up-regulated miRNA can target (down-regulate) one or more genes. As we mentioned earlier we assume that all the KEGG pathways are functional, which means that all the sub-paths are functional. We identify the genes that have been down-regulated due to the targeting of miRNAs and we consider the rest genes as up-regulated. An example of the mapping of miRNAs to targeted genes is shown in Figure 43.

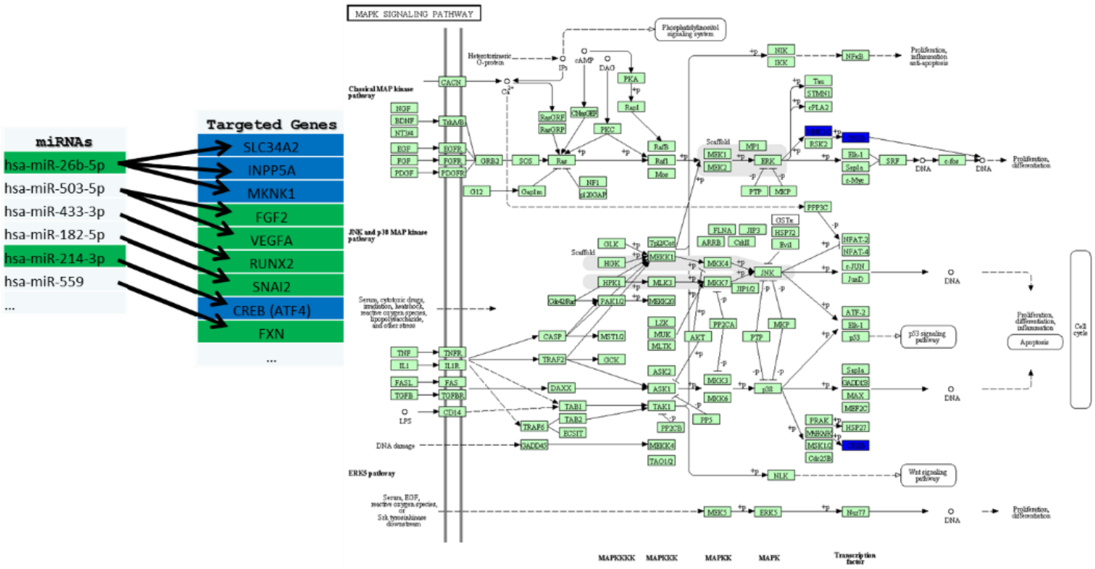


Figure 43: miRNA example. To the left are miRNAs with targeted gene (blue) and to the right the effect of the targeted genes in a specific pathway. Green represents the active (up-regulated) genes and blue the targeted (down-regulated) genes.

The pathway analysis scenario has two steps. The first step is to create and train a model able to predict outcomes for new samples. MinePath uses two classes approach to identify differentially expressed pathways & sub-pathways and has been extended to support miRNA expression data. The reference cohort for the creation of the model is based on the hsa (human) KEGG pathways and the GSE38419 public miRNA dataset.

MirTarBase[37] database has been used to identify targeted genes from the miRNAs. The clinical variable of GSE38419 dataset is the characterization to a wilm's tumour patient or to a healthy person and the model has been trained to predict one of these two classes. Steps for the first part of the scenario are shown in Figure 44. Then the prediction model is registered to the p-medicine workbench as a new biomedical resource for possible use.

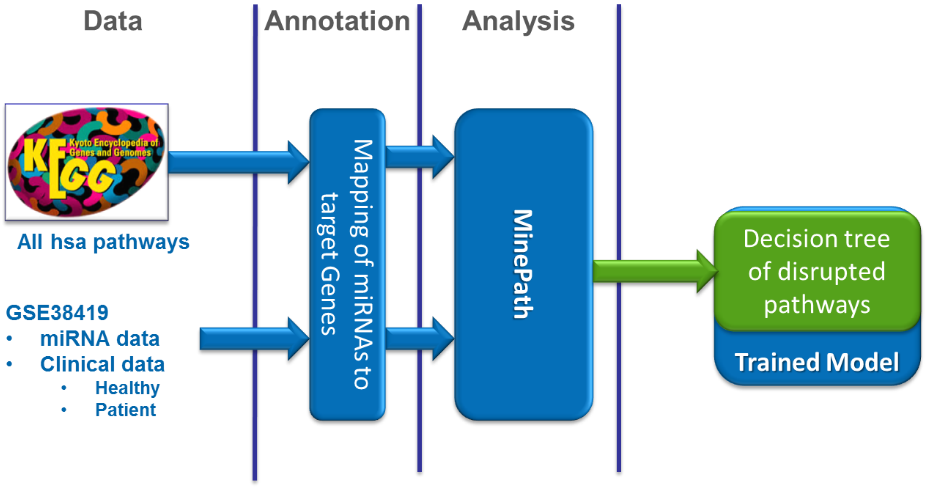


Figure 44: Flow of operations for the training step of the miRNA pathway analysis model. From left to right: Initially we collect the data, we identify the target genes from the miRNAs, we analyse using MinePath and finally we train the model using the disrupted sub-paths (from MinePath).

The second part is to predict if a new patient is characterized (according to his/her miRNA expression data) to wilm's tumour patient or to healthy person.

When a new patient, who is candidate for wilm's tumour, arrives in the hospital the clinician requests for a miRNA exam and searches in the p-medicine workbench for tools able to predict the disease based on disrupted pathways from miRNA expression data. The pathway analysis tool is identified as a candidate tool and the doctor downloads the tool (Figure 45).

[37] http://mirtarbase.mbc.nctu.edu.tw/ (last day visited 11/08/2014)

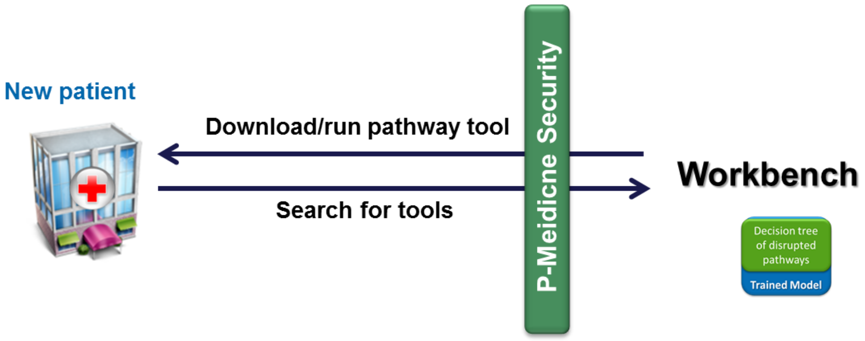


Figure 45: Download (from p-medicine workbench) and use of pathway analysis model in the clinical domain

The doctor gives as input to the tool the miRNA expression data of the patient and the tool normalizes/discretizes the genomic data according to the reference cohort (from step 1). Then the Mirtarbase database is used to identify targeted genes from the miRNAs. MinePath extracts the disrupted sub-paths for the specific patient and feeds the prediction model (created at step 1) to identify if the sample belongs to wilm's tumour patient or to a healthy person according to his/her miRNA expression data.

For the feasibility study of the miRNA extension we used the public dataset (GSE38419) from GEO. The dataset contains:

- Clinical data: Healthy and wilm's tumour patients
- Genomic data: miRNAs (848) per sample/patient

Details about the validation of the results can be found in the Experiments section.

3.6.2. Merging gene regulatory networks

A common operation on graphs is merging, that is, combining different graphs together. It is inspired by the fact that many KEGG pathways embed other pathways, for example MAPK signalling pathway embeds 6 pathways including Wnt signalling pathway. This extra functionality provides the possibility to merge them into one graph for further analysis. This is an extra of-line functionality that can be used only from the standalone tool of MinePath.

Using this extra functionality we created an artificial pathway, which is the merged pathway of the 14 cancer related pathways and can be found in the pathways tree of the web based platform as "Merged (14 cancer related)". The pathways that have been merged are shown in Table 5.

Table 5: Pathways engaged within the 'Pathways in Cancer' KEGG (hsa05200)

	KEGG Id	Pathway description
1	has04310	Wnt signalling
2	hsa04010	MAPK signalling
3	hsa04012	ErbB signalling
4	hsa04060	Cytocin-cytocin receptor interaction
5	hsa04110	Cell cycle
6	hsa04115	p53 signalling
7	hsa04150	mTOR signalling
8	hsa04210	Apoptosis
9	hsa04350	TGF-β signalling
10	hsa04370	VEGF signalling
11	hsa04510	Focal adhesion
12	hsa04512	ECM-receptor interaction
13	hsa04520	Adherens junction
14	hsa04630	Jak-STAT signalling

The merged pathway contains more than 2.500 sub-paths. An indicative example (screenshot from the MinePath web site) is shown in the following figure.

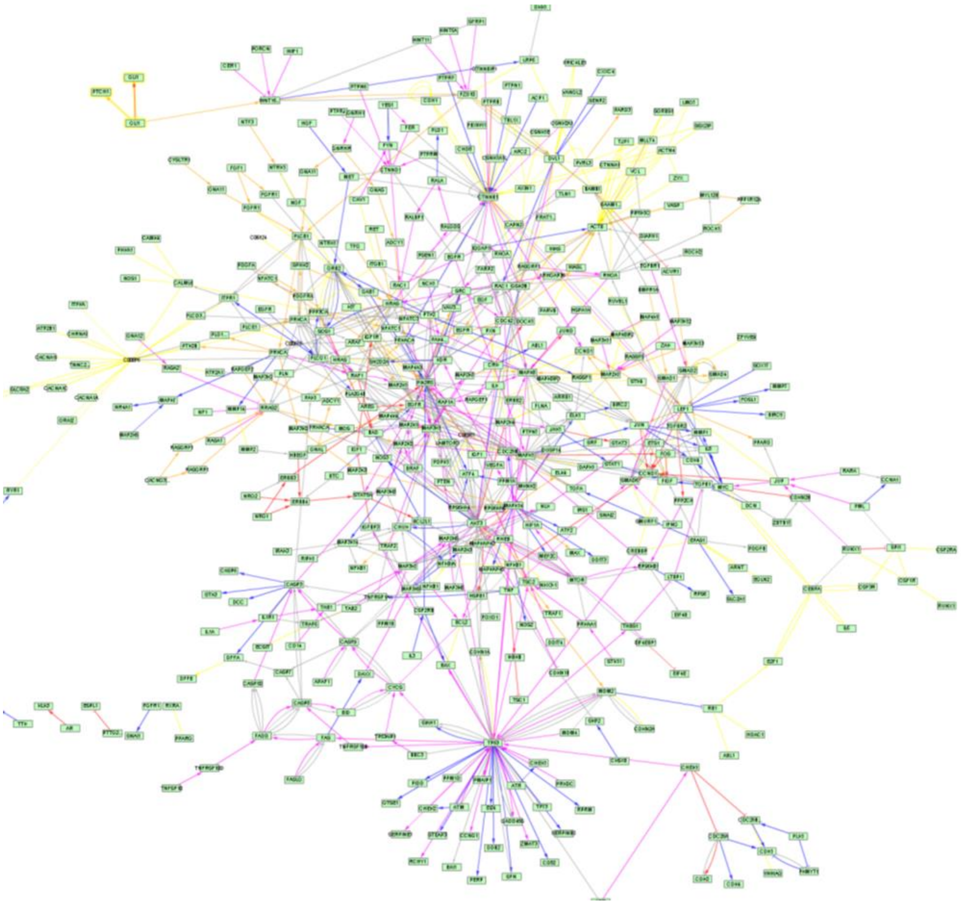


Figure 46: The merged pathway (14 cancer related pathways)

4. Experiments

In this chapter we discuss indicative results from MinePath using some of the available datasets from the web site of the tool.

The experiments prove the validity of MinePath methodology and highlight the value of the web based user interface in the quest of biological interpretation of the pathway analysis results.

4.1. MinePath comparison study

Though a comparative benchmark is hard to find, due to a missing gold standard that classifies detected sub-paths as right or wrong in the context of the investigated expression data, we tried to identify biological evidence in the literature and focussed on the specificity of the findings and the sensitivity of the method used. From the four known methodologies which cope with the regulatory mechanisms, only the glioma experiment used by GGEA is publicly available.

4.1.1. Glioma, comparison with GGEA

Glioblastoma or Glioma is the most common and malignant primary intracranial human neoplasm. GGEA [16], one of the four pathway analysis algorithms which takes advantage of the regulatory mechanisms, observed large agreement in the result lists of significant pathways with FiDePa [99] method. 17 pathways listed in the FiDePa result also occur in the top 25 of the GGEA ranking over public datasets for glioma. According to the authors, the positive control glioma is better ranked (and has higher significance) by GGEA. Further, several unspecific and disease unrelated pathways detected by FiDePa are discarded by GGEA and replaced by specific, cancer-related pathways (e.g. renal cell carcinoma, endometrial cancer). For the top rank, GGEA (Pathways in Cancer; not detected by FiDePa) gives a clear disease-related hint, while FiDePa (MAPK signalling pathway) reports a general signalling process.

We applied MinePath to the glioma dataset that has been investigated before with the method GGEA [16]. The reference dataset is a merging of two different studies using as classes the glioma cases from the GSE4271 [100] (100 samples) versus the control cases from the GSE1133 [101] (158 samples). For consistency evaluation, we used the regulatory interactions occurring in human non-metabolic KEGG pathways (gene regulatory and signaling pathways) similarly to the experiment from GGEA. Specifically we used all the human KEGG pathways which fall under the signal transduction, cell, immune system, endocrine system, nervous system and cancer related categories (in total 76).

MinePath identified the most discriminant sub-paths (1915 in total) which were evaluated based on the support vector machines algorithm from the Weka software and performed in both 10-fold and leave-one-out cross validation 100% accuracies. Detailed results for the leave-one-out cross validation can be found in Table 6.

Table 6: LOOCV & 10-fold CV results of best sub-paths from the Glioma dataset (Weka SVM)

Accuracy	100%
Precision	1
Recall	1

Confusion Matrix		
	Glioma	Control
Glioma	100	0
Control	0	158

FiDePa and GGEA report only the significant pathways while, MinePath identifies discriminant sub-paths. Based on the outcome of these two methodologies and the results of MinePath we can see from Table 7 that most of the best pathways from GGEA have been identified as highly discriminant using MinePath (pathways with ranking over 0.8). We observe large agreement in the result lists of the three methods. Furthermore we can see that MinePath ranked Glioma pathway as highly discriminant (score 1) while using FiDePa is ranked in 20th position and using GGEA in 12th position. MinePath identified cell cycle and Adipocytokine signaling pathway as highly discriminant pathways which is in accordance with FiDePa but not with GGEA. Cell cycle in most cancer cell types, including glioma, is a critical mechanism of development, progression, and resistance to treatment [102]. Pathways with the highest discriminant power in MinePath are Glioma, Neurotrophin signaling, Pancreatic cancer, Renal cell carcinoma, Chronic myeloid leukaemia, Insulin signaling and Adherens junction. The results of MinePath show high similarities with GGEA and FiDePa methodologies.

Table 7: Comparison of pathway analysis results from MinePath, GGEA and FiDePa methodologies in Glioma dataset. In grey pathways under the threshold of MinePath

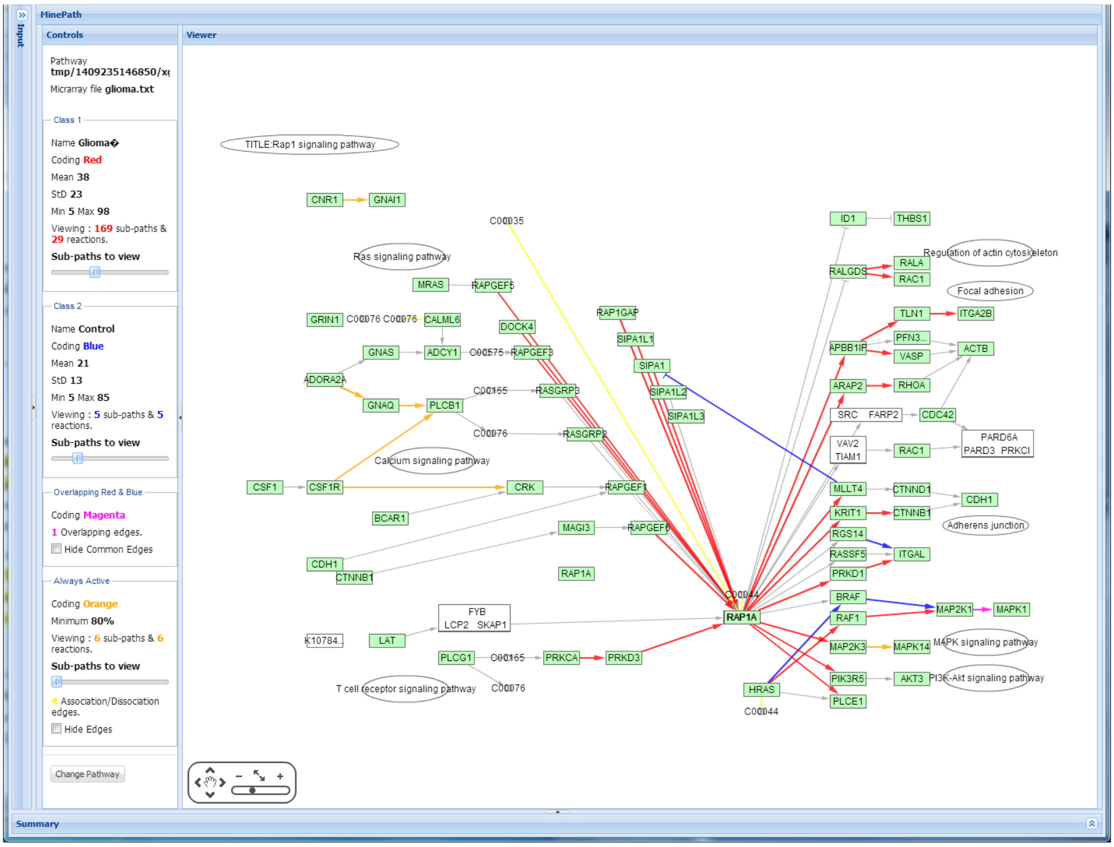
Pathway	MinePath score (pw diff)	ORA *P* (GGEA)	Rank (FiDePa)
Neurotrophin signalling	1	5.5E-15	–
Pancreatic cancer	1	3.8E-14	12
Renal cell carcinoma	1	1.3E-13	–
Chronic myeloid leukaemia	1	6.3E-13	8
Glioma	1	5.1E-12	20
Insulin signalling	1	3.2E-11	18
Adherens junction	1	4.9E-11	6

MAPK signalling	0.977	0.0000044	1
Cell cycle	0.966	---	19
Adipocytokine signaling pathway	0.964	---	14
Toll-like receptor signalling	0.962	1.2E-09	10
Acute myeloid leukaemia	0.957	0.00000039	–
Apoptosis	0.955	0.04	3
Leucocyte transendothelial migration	0.952	3.9E-11	24
Nature killer cell mediated cytotoxicity	0.938	6.5E-11	2
Pathways in cancer	0.93	1.8E-24	–
T cell receptor signalling	0.926	1.2E-17	7
ErbB signalling	0.926	8.9E-13	–
mTOR signalling	0.92	0.0000012	15
B cell receptor signalling	0.917	4.2E-12	17
Colorectal cancer	0.875	1.1E-14	11
Focal adhesion	0.855	1.4E-18	5
Wnt signalling	0.851	1.2E-10	–
GnRH signalling	0.829	6.5E-11	16
VEGF signalling	0.8	1.5E-13	22
Non-small cell lung cancer	0.8	0.00000034	–
Fc epsilon RI signalling	0.44	4.1E-13	9
Endometrial Cancer	---	0.00000016	–

Going one step further, the most discriminant sub-path based on the MinePath ranking (also functional for glioma – meaning that is active in most of the glioma samples and inactive in most of the control cases) is the NF-kB→HIF-1a in the HIF-1 signaling pathway. HIF-1 signaling pathway plays a critical part in tumor proliferation due to its role in hypoxia [103] and it is known that the hypoxic environment is created because of the extreme energy demands of the rapidly dividing cells when a tumor develops and grows.

Mendez et al [104] proved, in glioma cells, that HIF-1α protein plays a role in the survival and self-renewal potential of cancer stem cells. Authors identified genes that might further elucidate the role of HIF-1α in tumor migration, invasion and stem cell biology, making HIF-1α gene a very important gene for glioma.

Another interesting outcome comes from the Ras1 signaling pathway where we can see in Figure 47 that is mainly functional for the glioma samples (glioma functional sub-paths are shown in the figure with red). While it is known that alterations in the rap1 signaling pathway are common in human gliomas [105], it is not clear how the Rap1A hub gene is altered. A methodology like MinePath and its visualization capabilities could assist in the quest of such research questions.



Pathway **tmp/1409235146850/x**

Micrarray file **glioma.txt**

Class 1

Name **Glioma�**

Coding **Red**

Mean **38**

StD **23**

Min **5** Max **98**

Viewing : **169** sub-paths & **29** reactions.

Sub-paths to view

Class 2

Name **Control**

Coding **Blue**

Mean **21**

StD **13**

Min **5** Max **85**

Viewing : **5** sub-paths & **5** reactions.

Sub-paths to view

Overlapping Red & Blue

Coding **Magenta**

1 Overlapping edges.

□ Hide Common Edges

Always Active

Coding **Orange**

Minimum **80%**

Viewing : **6** sub-paths & **6** reactions.

Sub-paths to view

4 Association/Dissociation edges.

□ Hide Edges

Figure 47: Ras1 signalling pathway for glioma dataset. Red edges represent glioma sub-paths, blue control sub-paths and orange common sub-paths.

4.1.2. Gastric cancer, comparison with PATHOME

The comparison sample groups in the gene expression data set GSE13861[38] were 65 primary gastric adenocarcinoma frozen tissue samples and 19 normal appearing gastric tissue samples. Gastric cancer is the second leading cause of cancer-related death in the world, and prognosis is difficult to predict for individual patients. Most of gastric cancer patients receive similar treatments, typically surgery followed by chemotherapy because there are no reliable biomarkers to optimize therapy [106]. PATHOME was compared with two GSA tools, the GSEA and DAVID using the gastric cancer dataset and having as reference standard for cancer related pathways the pathways reported at the review of Vogelstein & Kinzler [107].

PATHOME used a lower significance cut off compared with that of the GSEA and DAVID methods and detected more differential cancer-related pathways. Actually PATHOME identified 8 out of the 19 cancer related pathways, DAVID and GSEA identified 1 out of the 19 each and MinePath identified 11 out of the 19. For MinePath we used the 19 pathways with the most gastric cancer functional sub-

[38] http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE13861 (last day visited 11/08/2014)

paths while in PATHOME the 27 most significant pathways were used. From the cancer related pathways reported in the reference standard, MinePath identified 3 out of the 9 while the other PATHOME and GSEA identified only one and DAVID none.

Table 8: Comparison table of PATHOME, DAVID, GSEA and MinePath for gastric cancer gene expression data using a reference standard for cancer pathways. Note: X (not detected), 0 (detected).

Reference Standard**	KEGG Pathway	Title	PATHOME*	DAVID	GSEA	MinePath
HIF1	hsa04150	mTOR signaling	X	X	X	0
	hsa05200	Pathways in cancer	0	X	X	0
	hsa05211	Renal cell carcinoma	X	X	X	X
P53	hsa04115	P53 signaling	X	X	X	X
RB(cell cycle)	hsa04110	Cell cycle	X	X	0	X
Apoptosis	hsa04210	Apoptosis	X	X	X	X
GLI	hsa04340	Hedgehog signaling	X	X	X	X
APC	hsa04310	Wnt signaling	0	X	X	0
RTK	hsa04012	ERBB signaling	X	X	X	X
	hsa05200	Pathways in cancer	0	X	X	0
SMAD	hsa04350	TGF-βsignaling	X	X	X	0
PI3K	hsa04012	ERBB signaling	X	X	X	X
	hsa05200	Pathways in cancer	0	X	X	0
	hsa04150	mTOR signaling	X	X	X	0
	hsa04010	MAPK signaling	0	X	X	0
	hsa04910	Insulin signaling	0	X	X	0
	hsa04510	Focal adhesion	0	0	X	0
	hsa04062	Chemokine signaling	0	X	X	0
	hsa04370	VEGF signaling	X	X	X	X
	19	Hits	8	1	1	**11**
		Selected	27	15	17	19

PATHOME reported significant sub-paths relating to WNT signalling, MAPK signalling, insulin signalling, focal adhesion and chemokine signalling (Table 8). Among these identified pathways, selected the WNT pathway as identified uniquely by PATHOME for further cell line and animal studies for accuracy validation. We have to mention that MinePath also identified WNT pathway as significant to gastric cancer.

4.2. MinePath Biological Validation - Craniosynostosis

Craniosynostosis is a birth defect in which one or more of the joints between the bones of the baby's skull close prematurely before the baby's brain is fully formed. Seven bones make up the skull of a newborn which are separated by spaces called sutures. The sutures meet at the fontanelles, the soft spots at the front and the back of a baby's skull. In order for the brain to grow, the sutures usually remain open and gradually grow together to form the adult skull. Craniosynostosis is the premature fusion (closing) of one or more of the sutures of a baby's skull.

In most cases of craniosyntosis there are no other birth defects, known as non-syndromic while the development with other birth defects, is known as syndromic. Non-syndromic craniosynostosis is the most common form of the condition, accounting for more than 80% of all cases [108]. Non-syndromic craniosynostosis usually occurs in a non-inherited fashion (not passed on from either parent), only involves fusion of one suture and are classified according to which suture is fused including unicoronal synostosis, metopic synostosis, sagittal synostosis and lambdoid synostosis. Frequencies of the various sutures involved are (i) sagittal: 40% to 58% while the etiology is unknown; (ii) unicoronal: 20% to 29%, estimated one third caused by single-gene mutations; (iii) metopic: 4% to 10%, etiology unknown; and (iv) lambdoid: 2% to 4%, etiology unknown [109].

Most cases of syndromic craniosynostosis are caused by genetic mutations [110] contrary to non-syndromic craniosynostosis which has proven to be a difficult task due to the complex nature of the disease [111].

4.2.1. The craniosynostosis dataset (GSE27976)

In this experiment, we use the GSE27976[39] [111] gene expression data from 199 patients with isolated sagittal (n = 100), unilateral coronal (n = 50), and metopic (n = 49) synostosis, compared against a control population (n = 50).

Stamper et al [111] concluded that FGF7, SFRP4, and VCAM1 emerged as potential genetic biomarkers for single-suture craniosynostosis due to their significantly large changes in gene expression compared to the control population. Authors also reported differentially regulated gene networks which were extracted using two thousand genes with the highest gene information content scores (the percent variance explained by the first eigengene obtained from a decomposition

[39] http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE27976 (last day visited 11/08/2014)

of the probe-level data for each gene). These genes were uploaded into DAVID [112] web application in order to identify biological pathways associated with genes in the dataset that had consistent changes in expression at the probe level. Using this gene list, focal adhesion and ECM-receptor interaction were the two most significantly implicated pathways. In addition, the TGF-beta signaling pathway, regulation of actin cytoskeleton, cell adhesion molecules (CAMs), and gap junction were also identified as significantly enriched pathways ($p<0.01$). The following table reports the full list of the significantly enriched pathways.

Table 9: Significantly enriched pathways of GSE27976 using DAVID web application

hsa04510	Focal adhesion
hsa04512	ECM-receptor interaction
hsa05412	Arrhythmogenic right ventricular cardiomyopathy (ARVC)
hsa05200	Pathways in cancer
hsa05410	Hypertrophic cardiomyopathy (HCM)
hsa05414	Dilated cardiomyopathy
hsa04350	TGF-beta signaling pathway
hsa00480	Glutathione metabolism
hsa05222	Small cell lung cancer
hsa04810	Regulation of actin cytoskeleton
hsa04115	p53 signaling pathway
hsa04514	Cell adhesion molecules (CAMs)
hsa04360	Axon guidance
hsa00980	Metabolism of xenobiotics by cytochrome P450
hsa05218	Melanoma
hsa04540	Gap junction
hsa04610	Complement and coagulation cascades
hsa05220	Chronic myeloid leukemia
hsa00010	Glycolysis / Gluconeogenesis
hsa04010	MAPK signaling pathway
hsa04020	Calcium signaling pathway
hsa05210	Colorectal cancer
hsa04110	Cell cycle

DAVID uses over-representation analysis (ORA), which statistically evaluates the fraction of genes in a particular pathway found among the set of genes showing changes in expression. The main limitation of the ORA algorithms is that assumes that each gene is independent of the other genes neglecting that gene regulatory networks are complex networks of interactions between genes. Furthermore ORA uses only the most significant genes, discards the rest genes and assumes that each pathway is independent of other pathways, which is erroneous.

We ran MinePath for the GSE27976 dataset using as classes all the synostosis cases (199 samples) versus the control cases (50 samples). We selected all the human KEGG pathway (in total 221). The best (2471) sub-paths were evaluated based on the support vector machines algorithm from the Weka software which performed in both 10-fold and leave-one-out cross validation accuracies over 97.5%. Detailed results for the leave-one-out cross validation can be found in Table 10.

Table 10: LOOCV results of best sub-paths from the GSE27976 (Weka SVM)

Accuracy	98.39%
Precision	0.984
Recall	0.984

Confusion Matrix		
	Synostosis	Control
Synostosis	198	1
Control	3	47

The fibroblast growth factor-7 (FGF7) is member of the fibroblast growth factor FGF family. FGF members which are known for broad mitogenic and cell survival activities, and are involved in a variety of biological processes, including embryonic development, cell growth, morphogenesis, tissue repair, tumor growth and invasion. FGF7 protein is a potent epithelial cell-specific growth factor, whose mitogenic activity is predominantly exhibited in keratinocytes but not in fibroblasts and endothelial cells [113]. FGF7 identified as the most discriminant and potential genetic biomarker for single-suture craniosynostosis along with SFRP4, and VCAM1 proteins by Stamper et al [111]. MinePath identified Rap1 signaling pathway as one of the most discriminant pathways out of the 221 and the most informative for Synostosis (contains the most functional sub-paths for this class).

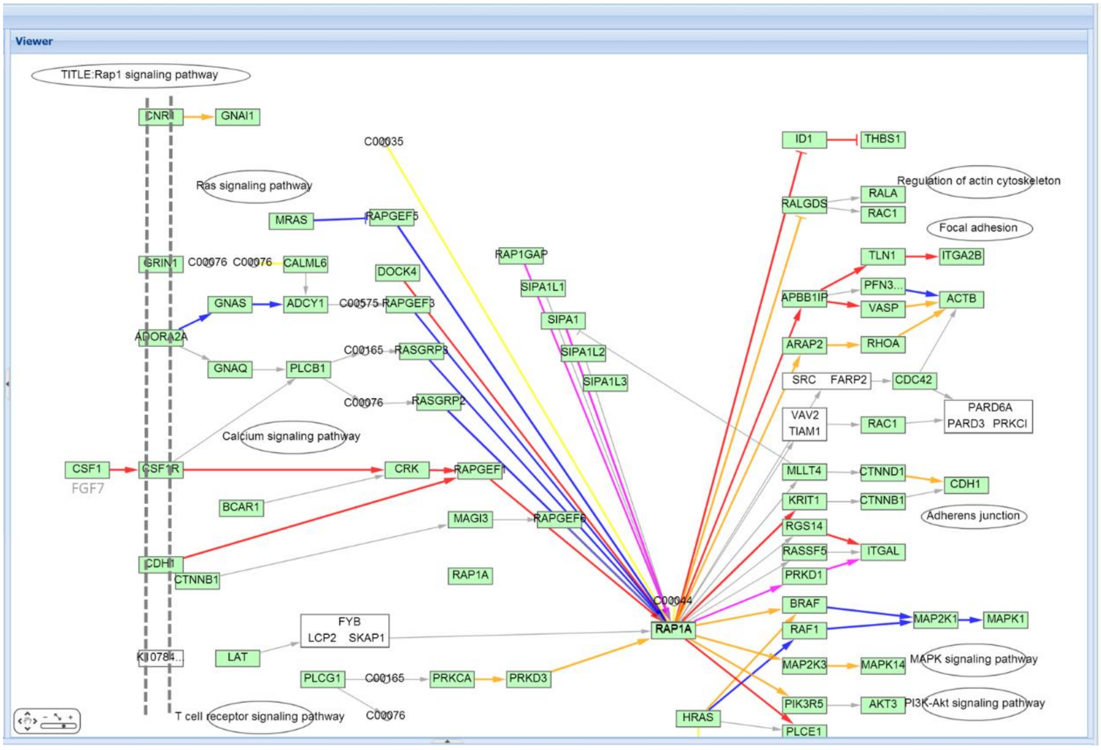


Figure 48: Rap1 signaling pathway for craniosynostosis. Red indicates relations active at synostosis, blue indicates relations active at control, magenta indicates overlapping relations and orange for sub-paths that are always active.

Figure 48 shows the discriminant sub-paths identified by MinePath for the Rap1 signaling pathway. Red indicates relations active at synostosis, blue indicates relations active at control, magenta indicates overlapping relations and orange for sub-paths that are always active. The vertical dashed lines distinguish the outer from the inner cell and we can see that only a group of genes belongs to the outer cell, the CSF1 which contains among others the FGF7 protein. MinePath identified a functional sub-path only for the Synostosis cases starting from this group of genes. Specifically the CSF1→CSF1R→CRK→RAPGEF1→RAP1A→APBB1P→TLN1→ITGA2B which leads to the focal adhesion pathway is considered to be discriminant for the two phenotypes and functional only in Synostosis. We can see that MinePath not only validated the results of Stamper et al [111] but also identified the path from FGF7 (the most discriminant gene according to Stamper et al) to the focal adhesion pathway (the most discriminant pathway according to Stamper et al).

Another finding of MinePath is the discriminant sub-path which is functional only for the Synostosis cases and starts again from the extracellular protein CSF1/FGF7 to the RAP1A hub gene which finally activates the PLCE1, leading to the PI3K-Akt signaling pathway. Dufour et al [114] identified that PI3K/Akt attenuation plays important role in the control of osteoblast survival by FGFR2 signaling (member of the fibroblast growth factor FGFR family).

Furthermore MinePath identified pathways functional only or mainly in one of the two phenotypes. P53 signaling pathway (in Figure 49 upper left) dominated by Synostosis meaning that the discriminant sub-paths are functional only for the Synostosis cases while discriminant sub-paths in Prolactin/Ras (in Figure 49 upper right), ErBb (in Figure 49 lower left) and Chemocine (in Figure 49 lower right) signaling pathways are mainly functional for the control samples. In Figure 49 we can see these four pathways where the red links indicate sub-paths functional in Synostosis, blue links indicate sub-paths functional in control samples and magenta links indicate sub-paths functional in both phenotypes.

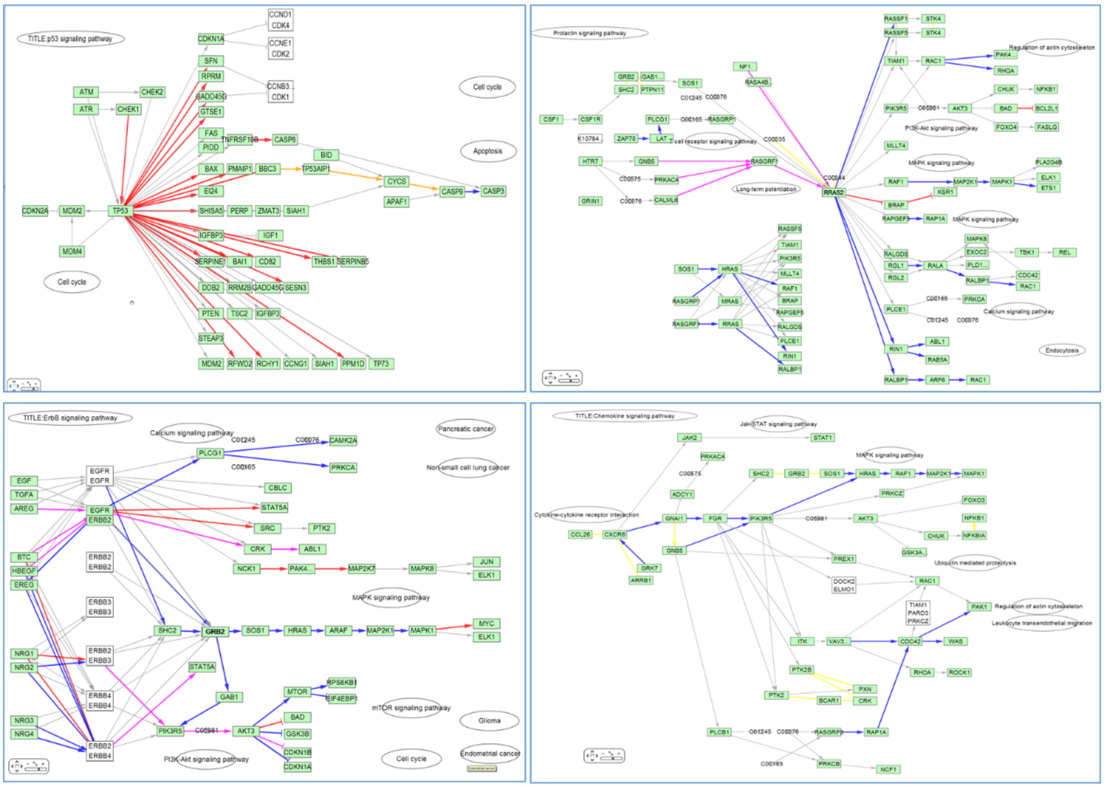


Figure 49: Pathways functional only or mainly in one of the two phenotypes (synostosis and control).

According to Moenning et al [115], PDGFRα signaling stimulates osteogenesis of neural crest cells derived osteoblasts by activating the PLC-γ pathway, using transgenic mice in vivo and in vitro experiments. Because the phenotype of transgenic mice resembles human craniosynostosis, the authors aimed to detect an involvement of PDGFRα in the etiology of human craniosynostosis. A sequencing analysis of the PDGFRα gene in 15 patients did not reveal PDGFRα mutations in the known hot-spot regions involved in autoactivation of the receptor. Nevertheless, the possibility of identifying mutations by screening an expanded group of craniosynostosis patients and sequencing the complete PDGFRα gene remains. The PDGFRα is part of the extra-celullar gene group (CSF1R) of Prolactin/Ras signaling pathway which was identified by MinePath as a descriptive pathway for control samples.

Feeding the best sub-paths, based on the MinePath ranking, in a C4.5 decision tree algorithm we can see that one pattern covers most of the Synostosis samples (154/199) while only one out of the fifty control samples follow this pattern. The pattern contains 7 sub-paths from which the two most discriminant (first and second selection of the decision tree algorithm) must be functional and the rest non-functional for a sample to be classified as Synostosis case. The pattern is shown in the following table.

Table 11: Pattern using sub-paths for the prediction of Synostosis.

```
hsa:4615-->hsa:3569 hsa:3569 = 1 (at Legionellosis)

|   hsa:5584 #hsa:5590_--> hsa:56288--> hsa:150084 hsa:50848 hsa:58494 hsa:83700 = 1
(at Tight junction)

|   |   hsa:22800 hsa:22808 hsa:3265 hsa:3845 hsa:4893 hsa:6237_--> hsa:10928 = 0 (at
RAS signaling)

|   |   |   hsa:84152--| hsa:5499 hsa:5500 hsa:5501--| hsa:775 hsa:775 hsa:776
@hsa04728 = 0 (at Dopaminergic synapse)

|   |   |   |   hsa:355--> hsa:8772--> hsa:843 = 0 (at Apoptosis)

|   |   |   |   |   hsa:2323--> hsa:2322--> hsa:2885 = 0 (at Pathways in Cancer)

|   |   |   |   |   |   hsa:1794--> hsa:5879 hsa:5879 hsa:5879 hsa:5880--> hsa:5058
hsa:5058--> hsa:3984 hsa:3985 = 0 (at Fc gamma R-mediated phagocytosis)
```

4.3. MinePath vs. Original Gene-Expression Data (MinePath as prognostic/Diagnostic predictor for GSE3494)

Most of breast cancer (BRCA) cases are estrogen responsive, implying the activation of a series of growth-promoting pathways, for example the estrogen receptor (ER) related ErbB signalling GRN. In an effort to reveal the underlying regulatory mechanisms that govern BRCA patients' treatment responses we applied our methodology on public gene-expression studies from the GEO repository.

This experiment is based on the public dataset from the GEO repository GSE3494[40] [116]. The biological tumour samples (i.e. breast tumour specimens) consisted of freshly frozen breast tumours from a population-based cohort of women representing 65% of all breast cancers resected in Uppsala County, Sweden, from January 1, 1987 to December 31, 1989. Estrogen receptor status was determined by biochemical assay as part of the routine clinical procedure. Tran-

[40] http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE3494 (last day visited 11/08/2014)

script profiles of 251 primary breast tumours were assessed by using Affymetrix U133 oligonucleotide microarrays.

In the original paper of the study, Miller et al [116] evaluated several linear learning methods including: diagonal linear discriminant analysis, k-nearest neighbours and support vector machines. In each case, the optimal gene classifier was obtained by leave-one-out cross validation, where the linear model-fit procedure was iteratively applied to all samples minus the left-out sample. The resulting prediction accuracies were highly similar, ranging from 84.9% to 85.7%.

We used the same dataset (gene expressions only from Affymetrix Human Genome U133A Array) targeting the ER phenotypic status of the respective patients, i.e., ER+ (ER positive) vs. ER- (ER negative). We targeted 14 pathways all of which are engaged within the 'Pathways in Cancer' integrated pathway of KEGG (hsa05200) as shown in Table 5.

We used the default values for the parameters of the web based MinePath, specifically:

- ***Min % for both up (80%):*** A sub-path to be considered as up regulated for both classes must cover at least 80% of the cases for each class.
- ***Add StD at Thr (0):*** No StD added at the threshold (remained the median)
- ***Boost true positives (True):*** Polarity filtering
- ***Use percentage for Ranking (True):*** Relative polarity ranking formula

MinePath identified 4632 sub-paths from the 14 pathways and the probes list of U133A platform. From the 4632 sub-paths 746 were selected using the ranking algorithm.

The best (746) sub-paths evaluated based on the support vector machines algorithm from the Weka software performed in leave-one-out cross validation.

Table 12: LOOCV results of best sub-paths from the GSE3494 (Weka SVM)

Accuracy	95.95%
Precision	0.959
Recall	0.96

Confusion Matrix		
	ERpos	ERneg
ERpos	209	4
ERneg	6	28

The proposed methodology performed better than the three different algorithms used in the gene signature of the Miller et al [116] (accuracies ranging from 84.9% to 85.7%).

4.4. MinePath as A discovery of New Biological/Clinical Knowledge (4 breast cancer datasets)

The '4ERdatasets' dataset (can be found in the datasets of MinePath) is a set of four independent discretized and then merged gene-expression studies targeting the ER phenotypic status respective patients, i.e., ER+ (ER positive) vs. ER- (ER negative), from the GSE2034[41] [117], GSE2990[42] [118], GSE3494[43] [116] and GSE7390[44] [119] studies.

For the discretization, the same methodology as in MinePath was used in the level of probes. Each dataset was discretized individually and then the four datasets were merged. The merging after the discretization was straight forward since these four clinical trials used the same microarray platform. The platform used was the GPL96 HG-U133A Affymetrix Human Genome U133A Array[45]. The U133 set (U133A & U133B) includes 2 arrays with a total of 44928 entries and was indexed 29-Jan-2002. The set includes over 1.000.000 unique oligonucleotide features covering more than 39.000 transcript variants, which in turn represent greater than 33.000 of the best characterized human genes. The HG-U133A Array includes representation of the RefSeq database sequences and probe sets (22282 probes) related to sequences previously represented on the Human Genome U95A Array. More details regarding the datasets, the numbers of samples per class, the number of probes and the annotated KEGG Id genes can be found in Table 13.

Table 13: Details for the four ER datasets; Each column is one dataset while rows provide information regarding the platform, the class, the number of samples per class, the number of probes for the platform and the identified number of genes (as KEGG/Entez IDs)

Dataset	GSE2034	GSE2990	GSE3494	GSE7390	4datasets
Platform	Affy-U133A	Affy-U133A	Affy-U133A	Affy-U133A	Affy-U133A
Class	ER	ER	ER	ER	ER
ER+ samples	209	149	213	134	705
ER- samples	77	34	34	64	209

[41] http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=gse2034 (last day visited 11/08/2014)
[42] http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE2990 (last day visited 11/08/2014)
[43] http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE3494 (last day visited 11/08/2014)
[44] http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=gse7390(last day visited 11/08/2014)
[45] http://www.affymetrix.com/estore/browse/products.jsp?productId=131536&categoryId=35760#1_1

Probes	22283	22283	22283	22283	22283
Annotated to KEGG	20967	20967	20967	20967	20967

For this experiment we did a validation on the independent datasets. Initially we compared the original gene expression data on independent datasets and we did the same using sub-paths. The results of the independent (train on one dataset and test on another) for the four ER datasets for genes and sub-paths can be found in Table 14. As we can see the sub-paths provide better accuracies in most of the cases and show a high consistency compared to genes. We must say that such an outcome is expected since sub-paths contain more information and provide more consistent and meaningful information.

Table 14: Validation on independent datasets for genes and sub-paths

Genes

	GSE2034	GSE2990	GSE3494	GSE7390
GSE2034		**81.42%**	**86.23%**	67.67%
GSE2990	26.92%		**86.23%**	39.89%
GSE3494	26.92%	**91.25%**		38.88%
GSE7390	73.07%	21.85%	13.76%	

Sub-paths (Best syb-paths vs All sub-paths)

	GSE2034	GSE2990	GSE3494	GSE7390
GSE2034		53.55%	85.02%	**70.20%**
GSE2990	**73.07%**		**86.23%**	**67.67%**
GSE3494	**77.27%**	54.64%		**79.29%**
GSE7390	**83.56%**	**73.77%**	**89.87%**	

On the next step, we ran Minepath for the five datasets (including the merged “4datasets”). Downloaded the results and used for each dataset the best sub-paths .arff file as train in weka. Then we evaluated the all sub-paths .arff file of the rest datasets as test in the weka trained model. Table 15 summarizes the results of this independent validation. As we can see the merged dataset performed the best accuracies overall and the average of accuracies is 99.595%. Even though the merged dataset actually contains the test subset each time, its trained model provided very high accuracies (over 99%) overall the datasets. This finding is in compliance with the conclusions of the authors from [120] who proved that “*due to the small sample sizes relative to the complexity of the entire expression profile, existing methods suffer certain limitations, namely the prevalence of study-specific signatures and difficulties in validating the prognostic tests constructed from these signatures on independent data. Integrating data from multiple studies to obtain more samples appears to be a promising way to overcome these limitations.*”

Table 15: Rows represent the train (best sub-paths) datasets versus the test (all sub-paths) in the columns across the four independent datasets and the merged dataset using the SMV implementation of Weka. Accuracy (Acc.), Precision, Recall and area under the curve (ROC area) for each train versus test experiment is reported.

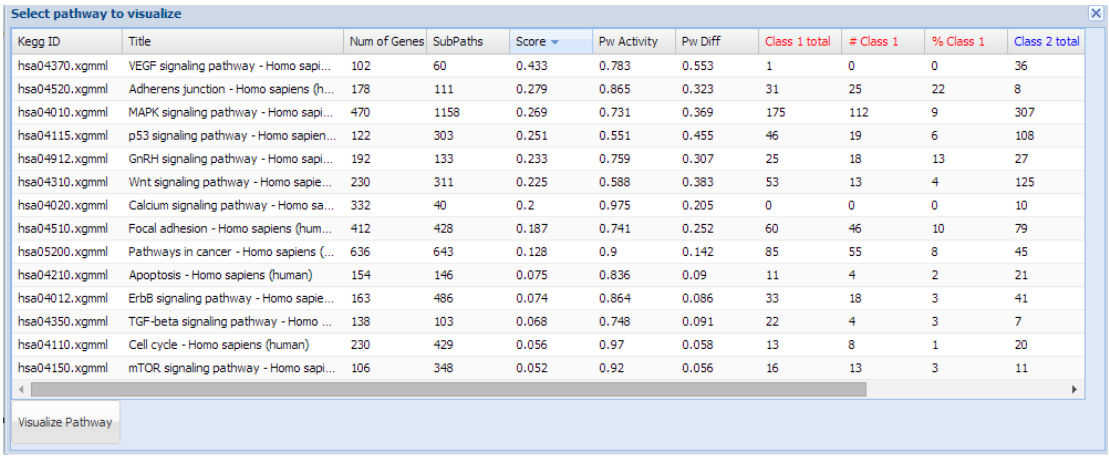
			Test (using all sub-paths)							
		Dataset	GSE2034				GSE2990			
	Sub path		Acc	Precision	Recall	ROC Area	Acc	Precision	Recall	ROC Area
Train (best sub-paths)	645	GSE2034	86.71% Acc. (10-fold)				53.550	0.604	0.536	0.329
	1264	GSE2990	73.07	0.534	0.731	0.500	87.43% Acc. (10-fold)			
	746	GSE3494	77.27	0.778	0.773	0.721	54.644	0.627	0.546	0.370
	794	GSE7390	83.56	0.829	0.836	0.748	73.770	0.891	0.738	0.839
	1013	4ER datasets	**99.30**	0.993	0.993	0.987	**100**	1.000	1.000	1.000

		Dataset	GSE3494				GSE7390			
	Sub path		Acc	Precision	Recall	ROC Area	Acc	Precision	Recall	ROC Area
Train (best sub-paths)	645	GSE2034	85.02	0.867	0.850	0.740	70.202	0.786	0.702	0.747
	1264	GSE2990	86.23	0.744	0.862	0.500	67.670	0.458	0.677	0.500
	746	GSE3494	95.54% Acc. (10-fold)				79.292	0.812	0.793	0.794
	794	GSE7390	89.87	0.888	0.899	0.694	87.87% Acc. (10-fold)			
	1013	4ER datasets	**99.59**	0.996	0.996	0.985	**99.49**	0.995	0.995	0.992

		Dataset	4ER datasets				**AVERAGE**			
	Sub path		Acc	Precision	Recall	ROC Area	Acc	Precision	Recall	ROC Area
Train (best sub-paths)	645	GSE2034	80.19	0.829	0.802	0.777	72.241	0.772	0.723	0.648
	1264	GSE2990	80.96	0.847	0.810	0.584	76.984	0.646	0.770	0.521
	746	GSE3494	79.64	0.812	0.796	0.747	72.714	0.757	0.727	0.658
	794	GSE7390	86.43	0.888	0.864	0.867	83.408	0.874	0.834	0.787
	1013	4ER datasets	87.41% Acc. (10-fold)				99.595	0.996	0.996	0.991

But MinePath is not only a pathway analysis methodology, is a complete web based platform that aims to aid the quest of functional or disrupted sub-paths in known pathways. Going one step further, after the validation of the algorithmic process, we try to identify biological insights using the specific dataset for the distinction of the ER positive and the ER negative patients.

The statistics of the selected (for our experiment) pathways are shown in the following figure.



Kegg ID	Title	Num of Genes	SubPaths	Score	Pw Activity	Pw Diff	Class 1 total	# Class 1	% Class 1	Class 2 total
hsa04370.xgmml	VEGF signaling pathway - Homo sapi...	102	60	0.433	0.783	0.553	1	0	0	36
hsa04520.xgmml	Adherens junction - Homo sapiens (h...	178	111	0.279	0.865	0.323	31	25	22	8
hsa04010.xgmml	MAPK signaling pathway - Homo sapi...	470	1158	0.269	0.731	0.369	175	112	9	307
hsa04115.xgmml	p53 signaling pathway - Homo sapien...	122	303	0.251	0.551	0.455	46	19	6	108
hsa04912.xgmml	GnRH signaling pathway - Homo sapi...	192	133	0.233	0.759	0.307	25	18	13	27
hsa04310.xgmml	Wnt signaling pathway - Homo sapie...	230	311	0.225	0.588	0.383	53	13	4	125
hsa04020.xgmml	Calcium signaling pathway - Homo sa...	332	40	0.2	0.975	0.205	0	0	0	10
hsa04510.xgmml	Focal adhesion - Homo sapiens (hum...	412	428	0.187	0.741	0.252	60	46	10	79
hsa05200.xgmml	Pathways in cancer - Homo sapiens (...	636	643	0.128	0.9	0.142	85	55	8	45
hsa04210.xgmml	Apoptosis - Homo sapiens (human)	154	146	0.075	0.836	0.09	11	4	2	21
hsa04012.xgmml	ErbB signaling pathway - Homo sapie...	163	486	0.074	0.864	0.086	33	18	3	41
hsa04350.xgmml	TGF-beta signaling pathway - Homo ...	138	103	0.068	0.748	0.091	22	4	3	7
hsa04110.xgmml	Cell cycle - Homo sapiens (human)	230	429	0.056	0.97	0.058	13	8	1	20
hsa04150.xgmml	mTOR signaling pathway - Homo sapi...	106	348	0.052	0.92	0.056	16	13	3	11

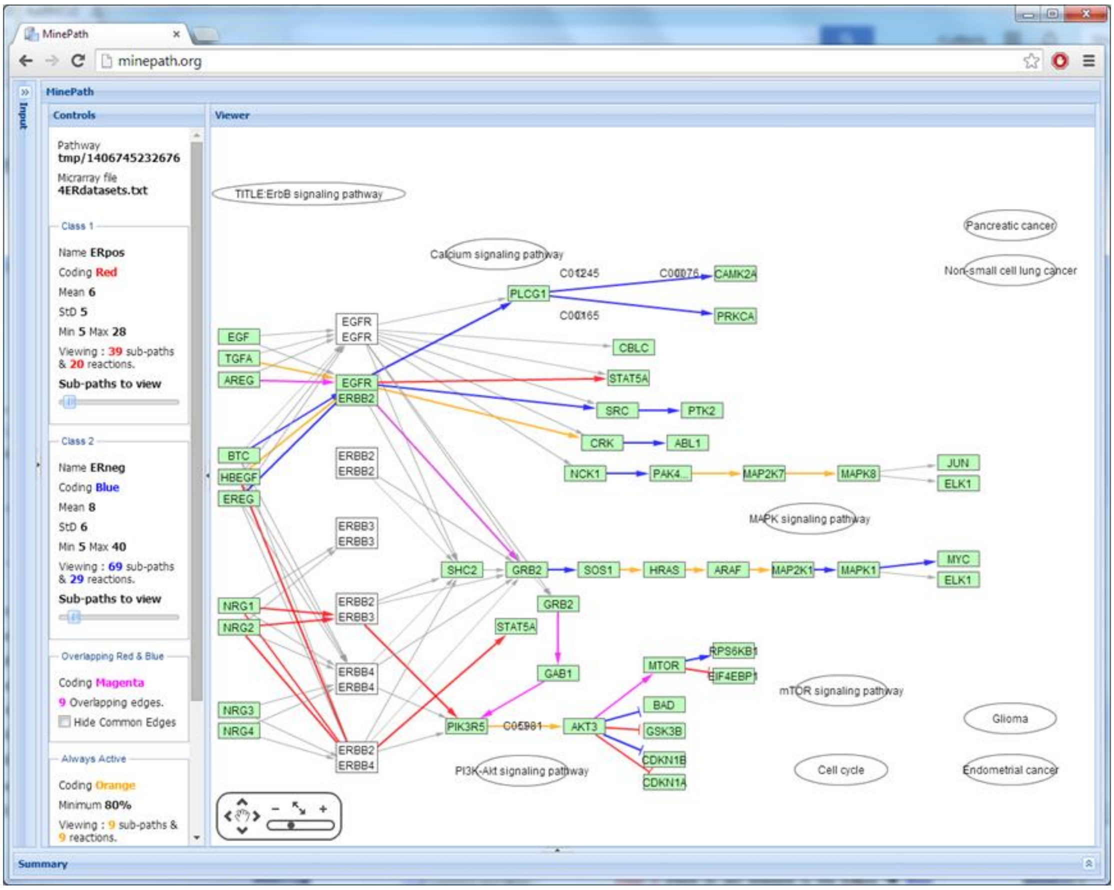
Figure 50: Statistics (from the MinePath web application) of the selected pathways for the 4ERdatasets dataset

It is known that ErbB-1 is overexpressed in many cancers [121]. Hence ErbB signalling pathway is one of the most important pathways to explore. The visualization of the MinePath results for the ErbB signalling (hsa04012) can be found in Figure 51.

As described in chapter 2, the MinePath web based graph GUI preserves the KEGG layout topology. It is enriched with the expressed regulatory mechanisms (relations) between genes that differentiate between the two phenotypes and the colour coding is as follows:

- Red indicates relations active at class 1, which in our example is the ERpos
- Blue indicates relations active at class 2 (ERneg)
- Magenta indicates overlapping relations in the two classes
- Orange for sub-paths that are always active.

The figure highlights only the “interesting” sub-paths, which in our case are the most discriminant sub-paths for the specific two phenotypes.



Pathway
tmp/1406745232676
Micrarray file
4ERdatasets.txt

Class 1

Name ERpos
Coding Red
Mean 6
StD 5
Min 5 Max 28
Viewing : 39 sub-paths & 20 reactions.
Sub-paths to view

Class 2

Name ERneg
Coding Blue
Mean 8
StD 6
Min 5 Max 40
Viewing : 69 sub-paths & 29 reactions.
Sub-paths to view

Overlapping Red & Blue

Coding Magenta
9 Overlapping edges.
☐ Hide Common Edges

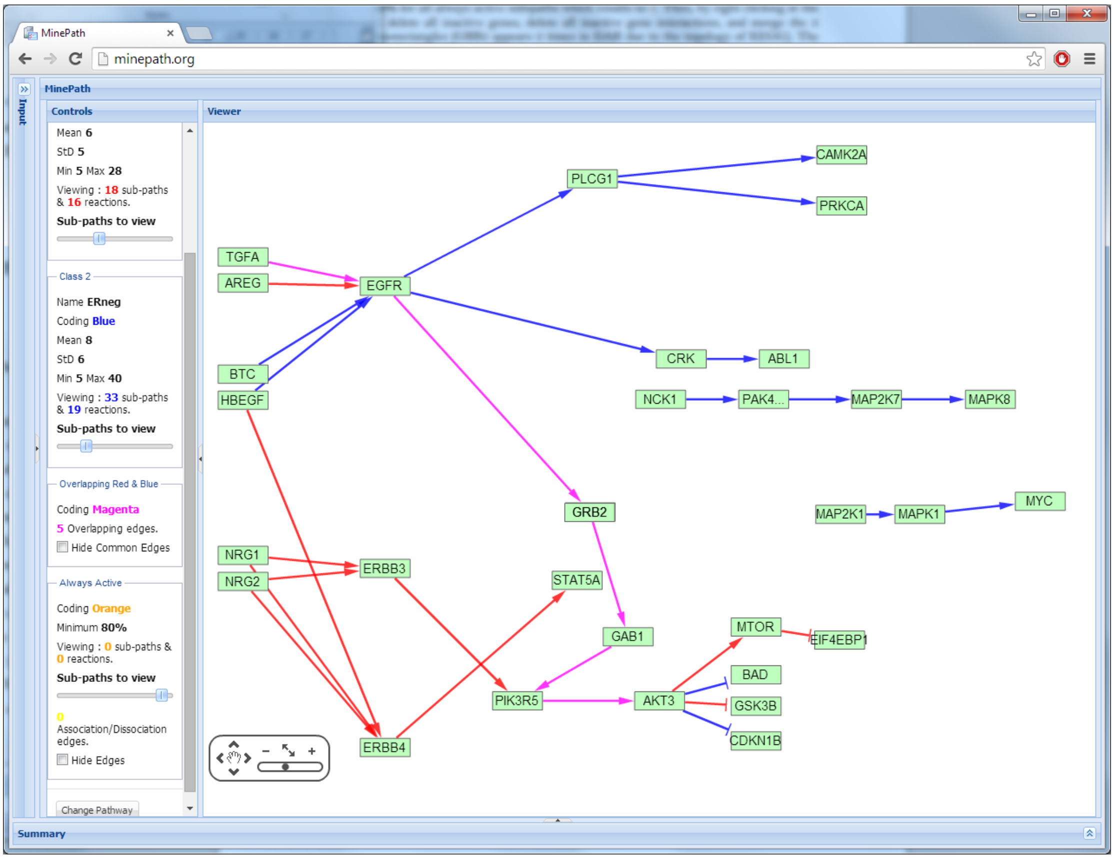
Always Active

Coding Orange
Minimum 80%
Viewing : 9 sub-paths & 9 reactions.

Figure 51: Visualizing ErBb for the 4ERdatasets using MinePath

Once we have the visual representation of the specific pathway, we can start exploring biological meaningful paths and sub-paths. Armed with the visualization functionalities of MinePath we can stress the thresholds to retain '*strong*' sub-paths per phenotype (class): Using **13** as threshold for class 1 (ERpos) results to **18** sub-paths and again **13** for class 2 (ERneg) results to **33** sub-paths; we also use **100%** for all always active sub-paths, which results to **0**. Then, we can "clean up" our pathway from the non-functional genes and reactions. By right click in the pathway viewer we select "Delete all inactive genes" and then we select the Delete all inactive gene interactions".

Then we merge the 2 GRB2 gene-rectangles (GRB2 appears 2 times in ErbB due to the topology of KEGG). The resulting (reduced) pathway will become as the one in Figure 52.


Mean 6
StD 5
Min 5 Max 28
Viewing : 18 sub-paths & 16 reactions.
Sub-paths to view

Class 2
Name ERneg
Coding Blue
Mean 8
StD 6
Min 5 Max 40
Viewing : 33 sub-paths & 19 reactions.
Sub-paths to view

Overlapping Red & Blue
Coding Magenta
5 Overlapping edges.
□ Hide Common Edges

Always Active
Coding Orange
Minimum 80%
Viewing : 0 sub-paths & 0 reactions.
Sub-paths to view
0 Association/Dissociation edges.
□ Hide Edges

Figure 52: Exploring ErBb for the 4ERdatasets using MinePath. Thresholds 13 for class 1 (ERpos), 13 for class 2 (ERneg), 100% for always active sub-paths and deleted all inactive genes and gene interactions

Both phenotypes (ER positive and ER negative) have extra-cellular origins:

- MinePath identified that the ER positive path originates from AREG (amphiregulin) that activates EGFR and consequently we have an activation of a common path (ER positive and ER negative) from EGFR→GRB2→GAB1→PI3K→PKB/Akt. It continues with two different sub-paths. The first one guide to the activation of mTOR, which leads to the inhibition of the EiF4EBP1 gene and blocks “protein synthesis” and the second one act as inhibitor of GSK-3 and blocking of “Metabolism”. Another clear path that leads to the same biological mechanisms for ER positive start from the extra-cellular NRG1, NRG2 (neuregulin1,2) growth factors that activate ErbB-3 and ErbB-4 viral oncogenes followed by the PI3K → PKB/Akt activation reaction.
- The ER negative path originates from the extra-cellular BTC (betacellulin) and HB-EGF (Heparin-binding EGF-like growth factor), shares the same sub-path with ER positive (EGFR→GRB2→GAB1→PI3K→PKB/Akt) but now this sub-path leads to the inhibition of BAD that is linked to “cell sur-

vival" and the inhibition of the CDKN1B protein, which blocks the "cell cycle progression".

According to recent literature, the aforementioned results are quite relevant to the estrogen-receptor status. Based on a search of the related biomedical literature we focus our exploration on the mechanisms underlying the resistance to pure estrogen antagonists (e.g., fulvestrant - a drug treatment of hormone receptor-positive metastatic breast cancer in postmenopausal). Recent studies show the significant role of both ErbB3 and ErbB4 as alternative targets for the treatment of BRCA patients. As Sutherland notes in [122]: "the initial growth inhibitory effects of fulvestrant appear compromised by cellular plasticity that allows rapid compensatory growth stimulation via ErbB-3/4. Further evaluation of pan-ErbB receptor inhibitors in endocrine-resistant disease appears warranted".

In addition, Hutcheson et al. in [123] investigated whether induction of ErbB3 and/or ErbB4 may provide an alternative resistance mechanism to antihormonal action. Their conclusion is that fulvestrant treatment is sensitive to the actions of the ErbB3/4 ligand HRGb1 (NRG1) with enhanced ErbB3/4-driven signalling activity and significant increases in cell proliferation.

4.5. MinePath using miRNAs (a clinical predictive model)

MicroRNAs (miRNAs) are endogenous molecules containing about 22 nucleotides that can play an important regulatory role in animals and plants by targeting mRNAs for cleavage or translational repression [124]. miRNA research has revealed multiple roles in negative regulation [125] (transcript degradation and sequestering, translational suppression) and possible involvement in positive regulation (transcriptional and translational activation). A miRNA controls gene expression post-transcriptionally either via the degradation of target mRNAs or the inhibition of protein translation. Using high-throughput profiling, dysregulation of miRNAs has been widely observed in different stages of cancer [126], [127]. The up-regulation (overexpression) of specific miRNAs could lead to the repression of tumour suppressor gene expression and conversely the down-regulation of specific miRNAs could result in an increase of oncogene expression; both these situations induce subsequent malignant effects on cell proliferation, differentiation and apoptosis that lead to tumour growth and progress [128], [129].

As Chen et al stated [128], miRNAs play key roles in human cancer, identifying the underlying pathways will provide a more complete understanding of their functions and regulations during cancer progression and may have clinical applications in the future. It is known that miRNAs affect (target or down-regulate) genes and that interactions between genes exist (pathways or parts of it). Therefore activations of miRNAs can result in the posttranscriptional down-regulation or up-regulation of the expression of certain genes [130].

In this experiment, we merge miRNAs and MinePath in order to identify disrupted sub-paths from miRNA expressions in known pathways. The methodology and the extension of MinePath to support miRNAs have been described in the Methodology chapter (section miRNAs to disrupted sub-paths).

The reference cohort for the experiment is based on the hsa (human) KEGG pathways (223 in total) and the GSE38419 public miRNA dataset [131]. The dataset contains 23 samples taken from Wilm's tumour patients prior to chemotherapy and 19 samples with the consent of healthy controls. The mean age of the treated patients was 3.3 years +/- 2.2 and the mean age of healthy controls was 37.8 years +/- 14.2. The microfluidic biochip (Geniom Biochip Homo sapiens v12, febit biomed GmbH, Heidelberg, Germany) contained 7 replicates of 848 miRNAs as annotated in the Sanger miRBase [132] version 12.0.

The clinical variable of GSE38419 dataset is the characterization to a Wilm's tumour patient or to a healthy person and the model has been trained to predict one of these two classes.

The discretization process of MinePath applied to each miRNA separately and the final dataset is a matrix of discretized values. Initially the expression levels of each miRNA over the total number of samples are sorted in descending order. Then the midpoints between each two consecutive values are calculated and for each midpoint, the samples are clustered into two sub-groups, high and low. For each midpoint, the information gain formula is applied, which computes the entropy of the system with respect to its division into subgroups. Finally, the midpoint that results in the highest information gain is selected as the one that best discriminates against the two subgroups and all the samples in the high group are considered to be overexpressed getting a value of 1, whereas the ones in the low group are the non-expressed/under-expressed, getting a value of 0.

Many miRNA-related database systems have been developed in recent years to provide further insight into miRNAs and their target genes. For the identification of the targeted genes we used the miRTarBase[46], a comprehensive collection of miRNA–target interactions (MTI), which are validated experimentally. The biological features of miRNA - target duplex are observed based on the largest collection of human MTIs currently available.

[46] http://mirtarbase.mbc.nctu.edu.tw (last day visited 11/08/2014)

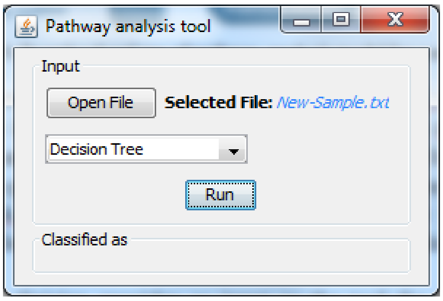
miRTarBase has accumulated more than fifty thousand miRNA-target interactions, which are collected by manually surveying pertinent literature after data mining of the text systematically to filter research articles related to functional studies of miRNAs. The collected MTIs are validated experimentally by reporter assay, western blot, microarray and next-generation sequencing experiments. While containing the largest amount of validated MTIs, the miRTarBase provides the most updated collection by comparing with other similar, previously developed databases. We used the current release (release 4.5), which contains:

- Number of articles: 2,636
- Number of species: 18
- Number of target genes: 17,520
- Number of miRNAs: 1,232
- Number of miRNA-target interactions: 51,460

For the specific dataset (GSE38419), that contains 848 miRNAs, we identified 7067 validated microRNA-target interactions from the miRTarBase.

After the decomposition of each of these pathways into its functional and disrupted sub-paths, the ranking formula of MinePath identified the most discriminant sub-paths (980). Then using the WEKA [91] machine learning library we created and trained two different predictive models able to predict new sample's category (healthy or Wilms tumour patient).

The second part of the scenario comes from the treatment domain (e.g. the hospital) and aims to predict if a new patient is characterized (according to his/her miRNA expression data) to Wilms tumour patient or to healthy person. When a new patient, who is candidate for Wilms tumour, arrives in the hospital the clinician requests for a miRNA exam and searches in the p-medicine workbench for tools able to predict the disease based on disrupted pathways from miRNA expression data. The pathway analysis tool is identified as a candidate tool and the clinician downloads the tool. Figure 53 shows the standalone prediction tool for Wilms tumour or healthy individuals based on miRNA expression data.



Selected File: New-Sample.txt

Decision Tree

Figure 53: The pathway analysis standalone prediction tool

The clinician provides as input to the tool the miRNA expression data of the patient and the tool normalizes/discretizes the genomic data according to the reference cohort (from step 1). Then the MirTarBase database is used to identify targeted genes from the miRNAs. In the last step, MinePath extracts the disrupted pathways for the specific patient and feeds the prediction model (created at step 1) to identify if the sample belongs to Wilms tumour patient or to a healthy person according to his/her miRNA expression data.

4.5.1. Support Vector Machines model

The support vector machines linear kernel classifier created a model using 780 sub-paths out of the 980 most discriminant and the remaining 200 sub-paths characterized as zero biased from the linear kernel. Randomized V-fold cross validation was performed using 10- fold and leave-one-out. 10 fold means that we divide the data into 10 subsets of (approximately) equal size. We train the classifier 10 times, each time leaving out one of the subsets from training for measuring “out-of-sample” performances. Then we measure the accuracy, which is the proportion of true results both true positives and true negatives in the population. The overall accuracy is the measured as the mean of the accuracies achieved in the 10 runs. Leave-one-out implies that all cases but one are used to train the model and then the model is tested using the left-out case. The process is repeated as many times as the number of records and the final results aggregate successes and misses.

The performance of our support vector machines linear kernel model was measured using the 10-fold cross validation and the leave-one-out cross validation methods, which both achieved 100% accuracy.

4.5.2. Decision Tree learning model

The decision tree learning (C4.5 [92] software Weka J48) was applied using data of the disrupted sub-paths as variables and Wilms tumour or healthy as different classes. The C4.5 algorithm builds a decision tree from the top; first the most discriminative variable (sub-path PLCβ→PKC→MEKK from GnRH signalling pathway) for classifying between Wilms tumour or healthy is selected. Then, the algorithm searches for the next best informative variable (sub-path PDK1→AKT→CREB from the PI3K-AKT signalling pathway) of the tree to improve the model. The third and final node of the decision tree is the P50→COX2 sub-path from the NF-KAPPA B signalling pathway. Figure 54 provides a graphical representation of the decision tree. Feature selection is a part of the decision tree algorithm. Interactions between features are taken into account. To measure the performance of the models, we calculated the accuracy for train versus train, which was 100%, for 10-fold cross validation 80% and for leave-one-out cross validation 78%.

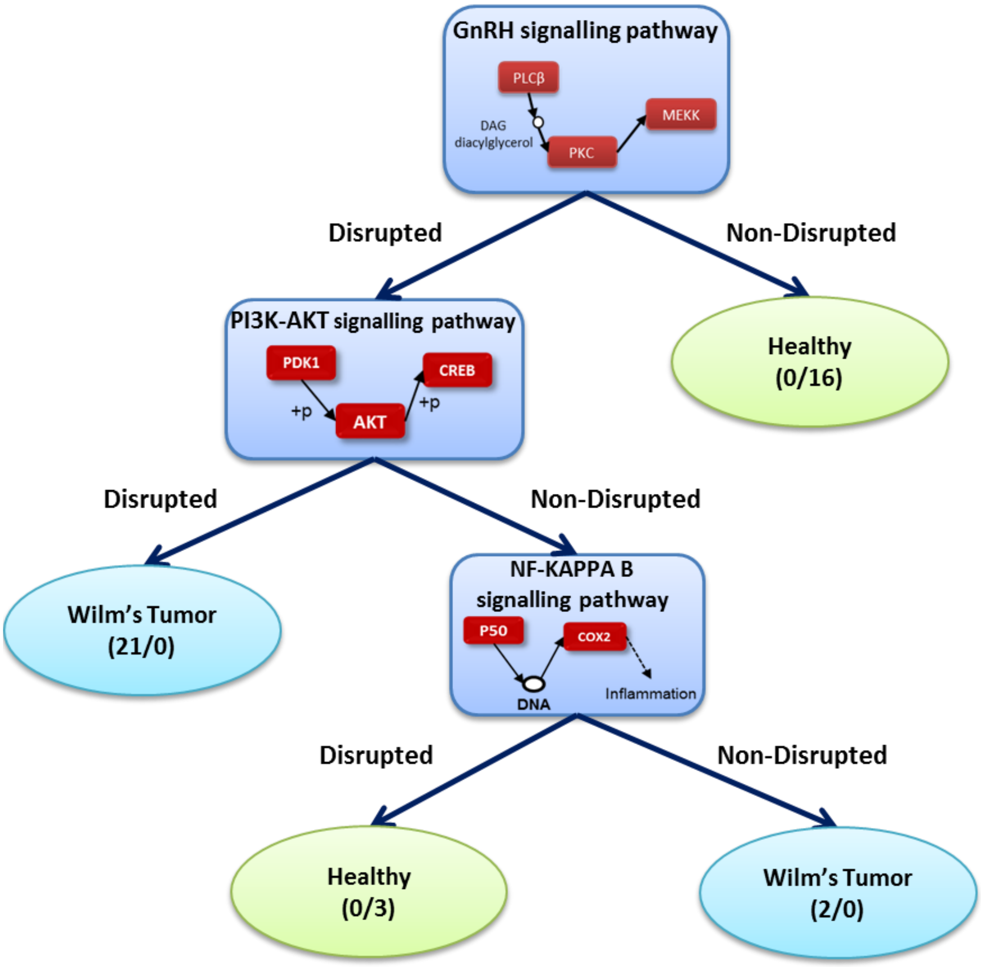


Figure 54: Decision tree for Wilms tumour prediction model. Starting from the top the most discriminative sub-path PLCβ→PKC→MEKK from GnRH signalling pathway is selected then the PDK1→AKT→CREB sub-path from the PI3K-AKT signalling pathway and the final node of the decision tree is the P50→COX2 sub-path from the NF-KAPPA B signalling pathway

Even though the decision tree model did not achieve the accuracy of the support vector machines model in leave-one-out cross validation (78% and 100% respectively), it is interesting that the decision tree uses only three sub-paths to predict new samples. Investigating the three selected sub-paths for the decision tree model we can see (Figure 55, Figure 56 and Figure 57 in red) that these sub-paths have a central role, in terms of topology and number of connections, in their respective pathways.

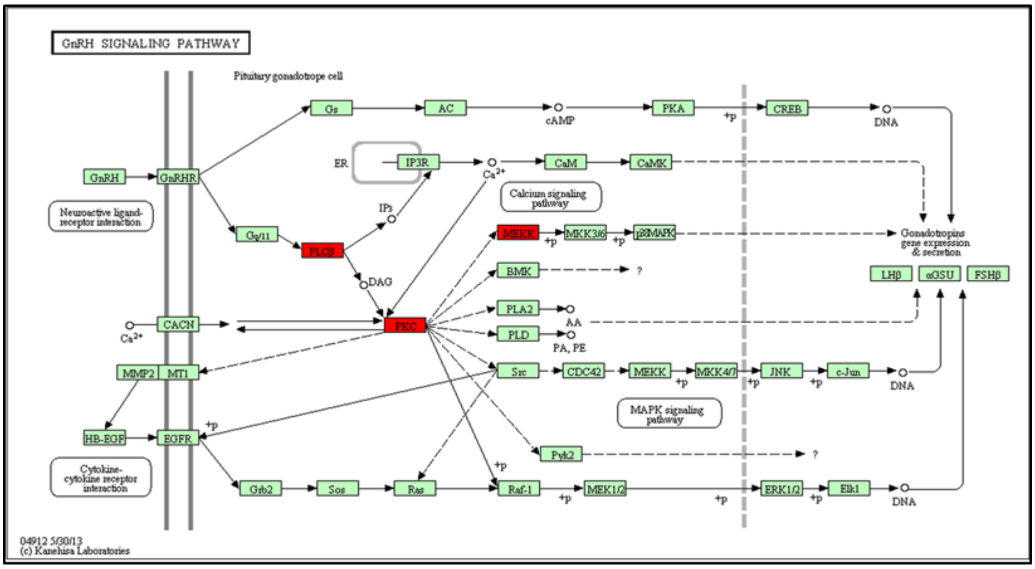


Figure 55: The PLCβ→PKC→MEKK disrupted sub-path (red) in the GnRH signalling pathway

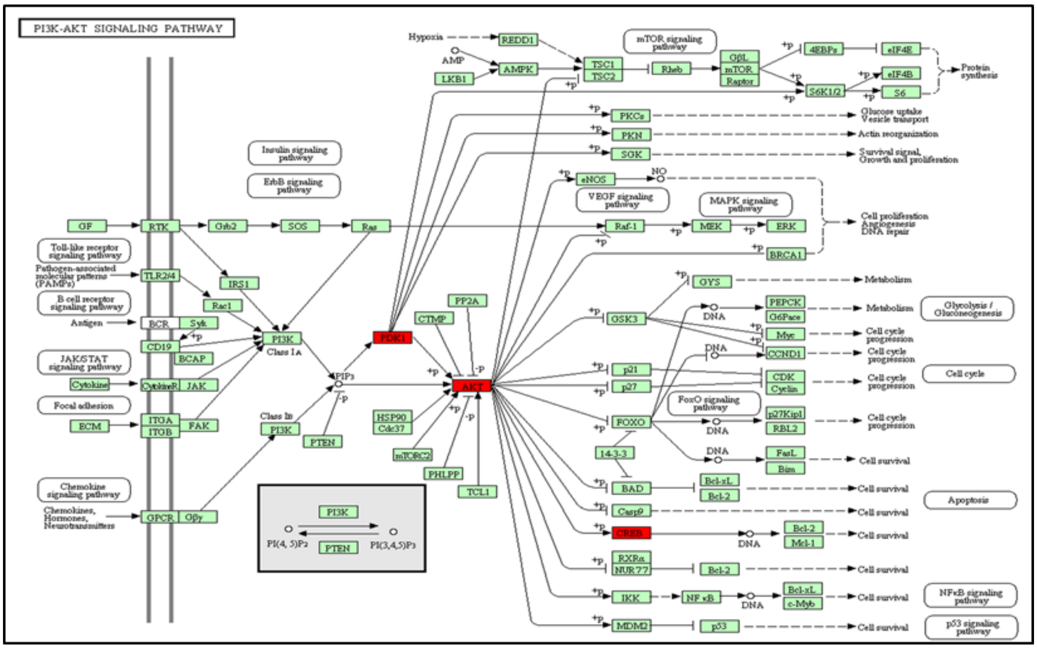


Figure 56: The PDK1→AKT→CREB disrupted sub-path (red) in the PI3K-AKT signalling pathway

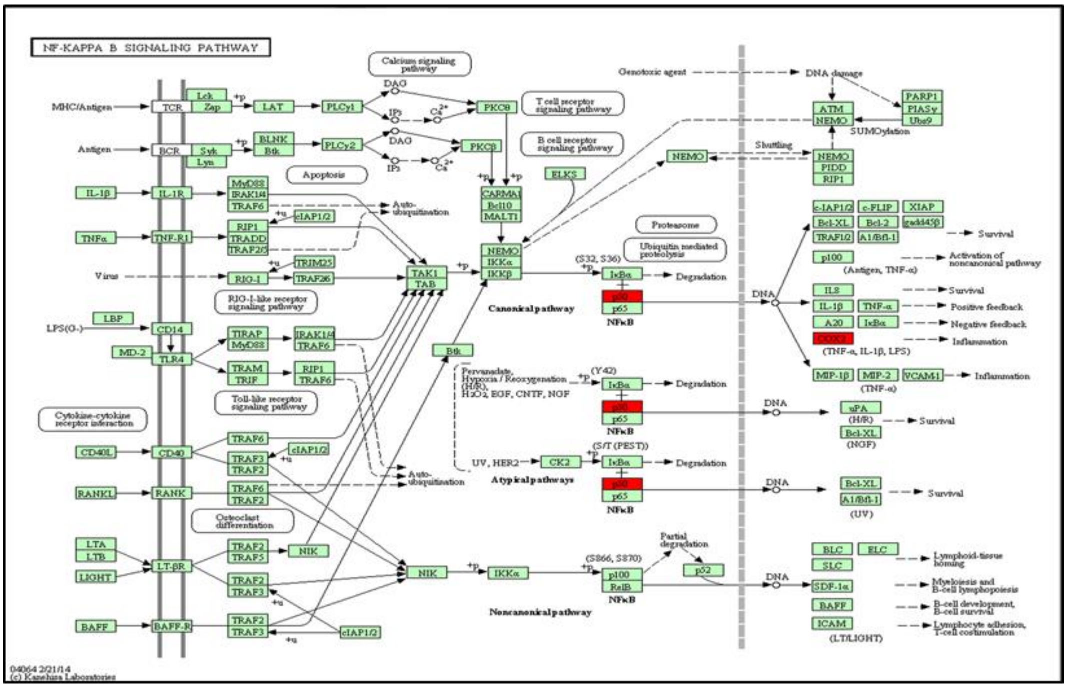


Figure 57: The P50→COX2 disrupted sub-path (red) in the NF-KAPPA B signalling pathway

The results are in agreement and justifies an already known finding about the regulatory role of miRNAs: miRNAs preferentially regulate hub nodes, i.e., top 5% of the highly connected nodes in the network, and the network cut points which are the bottle-necks of metabolic flows, however, avoid regulating intermediate nodes which are the nodes between the hub nodes, cut points, upstream nodes and the output nodes [133].

Furthermore, the protein kinase C (PKC), which has been identified by the decision tree as the most discriminant (the first) disrupted sub-path, is implicated in the regulation of neuroblastoma (pediatric kidney tumor) cell growth and proliferation [134]. Zeidman et al [135] proved that PKCε through its regulatory domain can induce immature neurite-like processes via a mechanism that appears to be of importance for neurite outgrowth during neuronal differentiation in neuroblastoma cells.

The second sub-path of our model comes from the PI3K-AKT signaling pathway. In many types of tumor PI3K-AKT pathway inhibition can lead to a wide spectrum of direct effects including cell-cycle arrest, induction of autophagy, inhibition of metastasis as well as cell differentiation and death [136]. Recently, Santo et al [137] identified the forkhead transcription factor FOXO3a as a key target of the PI3K/AKT pathway in neuroblastoma and concluded that the inactivation of FOXO3a by AKT was essential for neuroblastoma cell survival.

Similarly, Brown et al [138] using morphoproteomic analysis revealed the activation of the NF-kappaB pathway in high risk neuroblastoma cases. Preclinical studies such as the Brignole et al [139] and Michealis et al [140] using the proteosome inhibitor bortezomib, proved that NF-kappaB pathway regulates the proliferation of human neuroblastoma cells in vitro.

In conclusion, we identified, by supervised machine learning algorithms, a complex of potential causative factors for Wilms tumour: the simultaneous suppression of specific signalling sub-paths as discriminators between healthy and non-health. On the basis of these variables, patterns may be recognized to identify individuals at risk for Wilms tumour.

Around 10% of Wilms tumour patients are diagnosed having a concurrent syndrome that enhances the risk of Wilms tumour. But not all of these patients will develop such a tumour [141]. A screening method for early detection of Wilms tumour in these patients would be beneficial as the size or stage of a tumour is related with outcome [142]. In addition the detection of tumour specific disrupted pathways might help to find targeted therapies for individual patients. In one child with relapse of a bilateral nephroblastomatosis and disrupted retinoic acid pathway the treatment with retinoic acid did cure the child without tumour surgery [143]. If it can be shown that this pathway analysis tool is beneficial for Wilms tumour it can serve as a proof of principle for usage in other cancer. From a technological point of view a translation in other domains is easy as it is only necessary to link the tool with the corresponding database of patient specific miRNAs in other clinical domains.

5. Conclusions

Microarray experiments have advanced life scientists' ability not only to detect but also to quantify gene expression for target phenotypes. Initially the belief was that microarrays would reveal genotype categories (gene signature) for specific phenotypes. Unfortunately, microarray data mining has a number of limitations with most prominent (i) the noisy content (ii) the low reproducibility of the experiments and (iii) the fact that different gene-selection methodologies and techniques, even for gene-expression data acquired from the same experiment, produce gene lists that are strikingly different [9].

On the other hand, gene regulatory relations are restricted to what might be observed in an experiment. A change in the expression of a regulator gene modulates the expression of a target gene mainly via protein-DNA interactions. In other words, there are genes that causally regulate other genes. A change in the expression of these genes might change dramatically the behaviour of a part or the network as a whole. The identification and prediction of such changes is a challenging task, with the extraction and utilization of knowledge from GRNs to be of paramount importance.

Recently, bioinformatics community focused on more enhanced gene-selection methods, mainly by utilizing knowledge from other sources such as GRNs. Initial efforts used GRN information as groups (plain list) of associated genes in order to identify the most discriminant and phenotype-differentiating genes. Molecular pathways effectively reduced the resulting sets of genes, extracted from a gene set analysis approach and in some cases improved prediction performance but GRNs encompass much more knowledge form just a plain list of genes.

More and more methods take advantage of the GRNs topology and the underlying gene interaction patterns. In addition, most of the developed tools to take advantage of advanced network visualization toolkits in order to map and display the differentiating genes on target gene regulatory networks e.g., Cytoscape[47] and KEGG Mapper[48].

Pathway selection methodologies show similarities with gene signatures in terms of level of information used over the years. Although GRNs hold important information about the structure and correlation among genes that should not be neglected, most of the currently available methods in pathway selection do not fully exploit it. Analysing the literature we identified three categories of methodologies that focus on the identification and selection of discriminant pathways

[47] http://www.cytoscape.org/ (last day visited 11/08/2014)

[48] http://www.genome.jp/kegg/mapper.html (last day visited 11/08/2014)

and sub-pathways, based on the different levels of knowledge extraction from target GRNs. Initially the focus was on the identification of differentially expressed pathways (as a whole) using microarray data. Then the efforts concentrated on the knowledge of the GRN topology using decomposition mechanisms to reveal discriminant sub-pathways based on the graph theory concepts and network visualization toolkits. Recently more advanced methodologies are developed, which takes in consideration not only the topology of the GRNs but also, the regulation type (activation/inhibition) of the interaction link that connects two or more genes.

We classified the methodologies into three categories according to the level of the utilised GRN information. The categories are: pathway selection using GRNs as list of genes, sub-pathway selection using the topology of GRNs and sub-pathway selection methodologies using the underlying GRN gene regulatory interactions.

I. The first category naming "pathway selection" focus on the identification of differentially expressed pathways using microarray data. Nine (9) methodologies fall into this category. The proposed methodologies extract knowledge from gene regulatory networks trying, with the use of gene-expression data, to identify those pathways that contain the most discriminant genes.

II. The second category "sub-pathway selection using topology" includes eleven (11) methodologies. The respective methods go one step further and focus on the extraction of the discriminant pathways or, parts of pathways. Chuang et al [55] proved that the identified sub-networks are significantly more reproducible between different breast cancer cohorts than individual marker genes selected without network information. The authors also stated that network-based classification achieves higher accuracy than individual marker genes in prediction of independent validation data sets.

III. The third and most informative category is the "sub-pathway selection using regulatory mechanisms". While the previous approaches are useful, the valuable information from GRNs - such as the inherent gene regulatory relations found in biological pathways, is not taken in consideration. This category takes advantage not only of the topology of the GRNs but of the underlying gene relation types as well (i.e., activation or inhibition). This approach solves the major problem of the set enrichment strategies that refers to the conflicting constrains between GRNs and gene-expression data. A typical example of the conflicting constrains is reflected in the situation when two significantly up-regulated genes increase the enrichment of the set in microarray expression data, even if the first gene inhibits the other in a GRN.

The last category – being in its infancy, exhibits the fewer methodologies so far, but it takes the most out of GRNs and gene-expression data compared to the other two and is a promising alternative for the identification of the regulatory mechanisms that underlie and putatively govern various phenotypes.

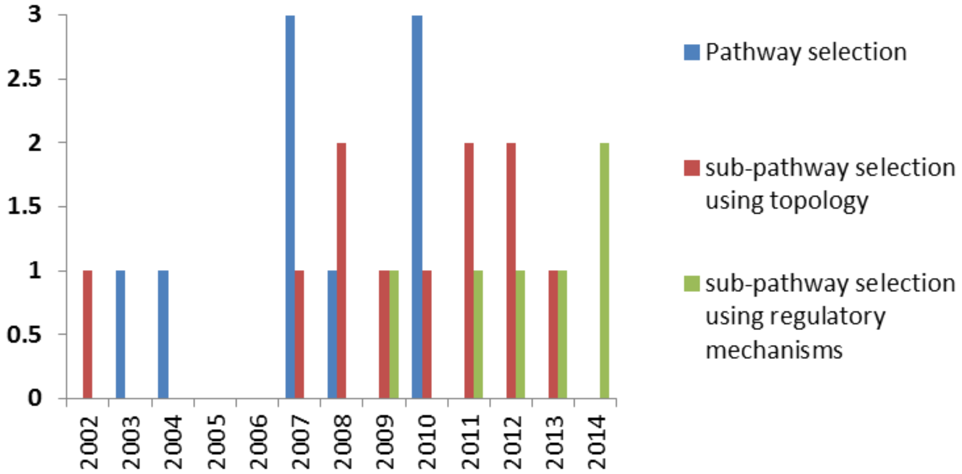


Figure 58: Number of methodologies for each category over the years

An overview of the number of the developed methodologies over the last years in the three reviewed categories is illustrated in Figure 58. It can be observed that for the pathway selection category the methodologies range from 2003 to 2010. The second category (sub-pathway selection using topology) has its first publication on 2007 and exhibits a relatively stable pattern until today. The most advanced and newer category is the third one (sub-pathway selection using regulatory mechanisms), which seems that it is at its first steps and could possibly gain a momentum. Our assumption for that momentum amplifies with the similarities we can find between the discriminant gene regulatory (sub)-networks and microarray gene selection methodologies.

Apart from the proposed procedure, only four (4) other tools take advantage of the underlying GRN gene regulation mechanisms, naming GGEA [16], SPIA [60], TEAK [15] and PATHOME [13]. The main difference of the proposed methodology from these four systems is the handling of the gene regulatory mechanisms. To our knowledge all the other methodologies count with a +1 the activations and -1 the inhibitions. Each sub-path gets a final score, which is also used as a ranking mechanism. Contrary, our approach strictly checks and takes into account only sub-paths that are functional (according to the gene relations and the expression values). Our approach is binary and leads to distinction between functional and non-functional sub-paths per sample instead of a representation of the sub-path per class (the sum).

MinePath relies on a novel approach for GRN processing that takes into account all possible functional interactions of the network. The phenotype information is extracted from microarrays and all the selected GRNs are evaluated for the identification of the most informative GRNs at the specific phenotype. The efficient ranking of sub-paths provides the most differentiating and prominent GRN functional sub-paths for the respective target phenotypes. The formulas possess a polarity characteristic according the class phenotype, i.e., positive for class S1 and negative for class S2. These sub-paths present evidential molecular mechanisms that govern the disease itself, its type, its state or other targeted disease phenotypes (e.g., positive or negative response to specific drug treatment). The methodology was applied on gene-expression studies including the target of identifying putative mechanisms that underlie and govern the treatment response of breast cancer patients according to their ER-status profiles. Results were quite indicative and strongly supported by the relevant biomedical literature.

Another advantage of MinePath over the similar tools is the productive environment with efficient, interactive and user-friendly visualization that offers rich exploratory capabilities towards the insight of key phenotype regulatory mechanisms, a fact that all the other solutions does not facilitate and inspection of results limits the exploratory potential of the users. Some gene set enrichment analysis methodologies and tools utilize pathway visualization approaches to overcome this problem. However, since they are based on a gene-oriented approach, they are still unable to handle differentially expressed pathways or even differentially expressed sub-paths. Solutions such as the KEGG Atlas/Mapper [95], WebGestalt [96], NetworkTrial [97] or even Graphite Web [98] visualize just the pathway genes using some colour scale or colour-coding schema. This problem is apparent even for small pathways. For example, the inhibition relation A —| B when up-regulation of A inhibits B and when down-regulation of A turns B up-regulated. For such different cases, different colours should be assigned to the genes. The situation becomes even more complicated when one has to visualize the phenotype inclination of an interaction. MinePath overcomes the aforementioned problems offering an effective identification and visualization of differentially active GRN sub-paths in real time on a solely Web-based platform.

Furthermore, MinePath takes also into account and visualizes sub-paths fully functional in both phenotypes. These sub-paths have no discriminant power but in the area of gene regulatory networks, the sub-paths that are always activated can link the gap (functional interaction) between two sub-paths and reveal a complete functional root, which is biologically valuable (e.g. link the gap between extracellular gene interactions and final biological reaction such as apoptosis).

The MinePath platform and its Web-based implementation aim to effectively address these issues. Its core algorithm determines differentially expressed pathway sub-paths and relations instead of just differential genes. These sub-paths present evidential molecular mechanisms that govern the disease itself, its subtype, state or other targeted disease phenotypes. In this form, MinePath introduces a new and efficient representation of the differentially expressed sub-paths over a Web-based human-computer interface. Furthermore, MinePath supports live interaction, immediate visualization of regulatory relations and it is equipped with special topological and network-adjustment functionalities.

Armed with the aforementioned features, MinePath serves the users' exploratory needs to reveal the regulatory mechanisms that underlie and putatively govern the expression of target phenotypes.

The current version of MinePath has been thoroughly tested for its stability. Exploratory results are quite satisfactory and the modular implementation of the core MinePath algorithm gives us the ability to "build on demand" new tools such as the miRNA scenario.

Additional functionality is foreseen in planned future releases of the methodology, the algorithm and the platform. The modular implementation gives us the ability to "build on demand" new tools based on end user scenarios. Such an example is the miRNA scenario/extension and we plan to create a validation tool of candidate sub-paths (GRN reconstruction validation).

For the methodology we plan to:

- Introduce new ranking algorithms
- Introduce other pre-processing methodologies (apart discretization)
- Support multi-class datasets
- Support other quantified gene-expression data (e.g., RNA-seq)

For the platform we plan to:

- Create automated uploading system of microarray data from public sources (e.g., GEO)
- Add merging of gene-expression datasets (to serve meta-analysis needs)
- Visualize two or more pathways in order to enrich exploratory quests.

It is known that integrating heterogeneous data sources is more effective than working within the boundaries of a single data domain, an observation that is particularly valid for the biomedical domain [144]. Bioinformatics and systems biology have demonstrated that knowledge across domains can better aid relevant scientific communities in their research endeavours or even reveal and create new research domains, such as translational bioinformatics [145]. Methodological approaches for pathway analysis have moved from employing algorithms using simple gene lists to the utilization of the topology and the regulatory mechanisms of biological networks.

Extracting out the most of the knowledge will always give us more natural and meaningful, as well as more accurate results.

References

[1] P.O. & Botstein, D. Brown, "Exploring the new world of the genome with DNA microarrays," *Nature genetics*, vol. 21, pp. 33–37, 1999.

[2] Dolled-Filhart M, King BL, Rimm DL. Camp RL, "Quantitative analysis of breast cancer tissue microarrays shows that both high and normal levels of HER2 expression are associated with poor outcome.," *Cancer Research*, vol. 63, no. 7, pp. 1445–1448, 2003.

[3] Smeds J, George J, Vega VB, Vergara L, Ploner A, Pawitan Y, Hall P, Klaar S, Liu ET, Bergh J. Miller LD, "An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 38, pp. 13550–13555, 2005.

[4] J Quackenbush, "Computational approaches to analysis of DNA microarray data," *Yearbook of medical informatics*, pp. 91–103, 2006.

[5] Sørlie T., "Molecular portraits of breast cancer: tumour subtypes as distinct disease entities.," *European Journal of Cancer*, vol. 40, no. 18, pp. 2667–2675, 2004.

[6] McShane LM, Simon R. Radmacher MD, "A Paradigm for Class Prediction Using Gene Expression Profiles. ," *Journal of Computational Biology*, vol. 9, no. 3, pp. 505–511, 2002.

[7] Robert, and Jennifer Shoemaker. Nadon, "Statistical issues with microarrays: processing and analysis.," *TRENDS in Genetics*, vol. 18, no. 5, pp. 265-271, 2002.

[8] Ray L., B. Dolenko, and Richard Baumgartner Somorjai, "Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions.," *Bioinformatics*, vol. 19, no. 12, pp. 1484-1491, 2003.

[9] Kela I, Getz G, Givol D, Domany E Ein-Dor L, "Outcome signature genes in breast cancer: is there a unique set?," *Bioinformatics* , vol. 21, no. 2, pp. 171–178, 2005.

[10] Takayuki Iwamoto and Lajos Pusztai, "Predicting prognosis of breast cancer with gene signatures: are we lost in a sea of data?," vol. 2, no. 11, p. 81, 2010.

[11] Tun-Hsiang Yang, Zhenjun Hu, Zhiping Weng and Charles DeLisi Jui-Hung Hung, "Gene set enrichment analysis: performance evaluation and usage guidelines," *Briefings in Bioinformatics*, vol. 13, no. 3, pp. 281-291, 2012.

[12] Albert-László, Natali Gulbahce, and Joseph Loscalzo Barabási, "Network medicine: a network-based approach to human disease.," *Nature Reviews Genetics*, vol. 1, no. 12, pp. 56-68, 2011.

[13] S., Chang, H. R., Kim, K. T., Kook, M. C., Hong, D., Kwon, C. H.,. & Kim, Y Nam, "PATHOME: an algorithm for accurately detecting differentially expressed subpathways.," *Oncogene*, 2014.

[14] Brad T. Sherman, and Richard A. Lempicki Da Wei Huang, "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.","

Nature protocols 4, vol. 1, pp. 44-57, 2008.

[15] Cole Johnson, Anuj Kumar and Dongxiao Zhu Thair Judeh, "TEAK: Topology Enrichment Analysis frameworK for detecting activated biological subpathways," *Nucleic Acids Research*, vol. 41, no. 1, pp. 1425-37, 2013.

[16] Gergely Csaba, Robert Küffner, Nicola Mulde and Ralf Zimmer Ludwig Geistlinger, "From sets to graphs towards a realistic enrichment analysis of transcriptomic systems," *Bioinformatics*, vol. 27, no. 13, pp. i366-373, 2011.

[17] Sandro Lambecka, Susanne Toepferb, Eugene van Somerenc, Reinhard Guthke Michael Heckera, "Gene regulatory network inference: Data integration in dynamic models—A review," *Biosystems*, vol. 96, no. 1, pp. 86–103, 2009.

[18] Chen H, Martinez JD. Leroy G, "A Shallow Parser Based on Closed-class Words to Capture Relations in Biomedical Text," *Journal of Biomedical Informatics* , vol. 36, pp. 145-158, 2003.

[19] Hyun Sook Kim , Jung Jae Kim Jong C. Park, "Bidirectional Incremental Parsing For Automatic Pathway Identification with Combinatory Categorical Grammar," *Pacific Symposium on Biocomputing.* , vol. 6, pp. 396-407, 2001.

[20] Yuryev A, Egorov S, Novichkova S, Nikitin A, Mazo I. Daraselia N, "Extracting Human Protein Interactions from MEDLINE Using a Full-sentence Parser," *Journal of Bioinformatics*, vol. 20, no. 5, pp. 604-611, 2004.

[21] Kra P, Yu H, Krauthammer M, Rzhetsky A. Friedman C, "GENIES: a Natural-language Processing System For theExtraction of Molecular Pathways From Journal Articles," *Journal of Bioinformatics* , vol. 17, no. 1, pp. 74-82, 2001.

[22] G. Demetriou, P. J. Artymiuk and P. Willett R. Gaizauskas, "Protein Structures And Information Extraction From Biological Texts: thePASTA System," *Journal of Bioinformatics*, vol. 19, no. 1, pp. 135-143, 2003.

[23] Zhou M and Cui Y Li H, "Ranking Gene Regulatory Network Models with Microarray Data and Bayesian Network," *Lecture Notes in Computer Science* , vol. 3327, pp. 109-118, 2005.

[24] J Li, H Su, G Watts, H Chen Z Huang, "Large-scale regulatory network analysis from microarray data: modified Bayesian network learning and association rule mining," *Decision Support Systems*, vol. 43, no. 4, 2007.

[25] Yoneda K and Wu R. Wachi S, "Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues," *BioInformatics* , vol. 21, no. 23, pp. 4205-4208, 2005.

[26] Anders Wallqvist and Jaques Reifman Bhaskar Dutta, "PathNet: a tool for pathway analysis using topological information," *Source Code for Biology and Medicine*, vol. 7, no. 10, 2012.

[27] Hua Dong, Li Jin and Momiao Xiong. Hoicheong Siu, "New Statistics for Testing Differential Expression of Pathways from Microarray Data," *complex science*, vol. 4, pp. 277-285, 2009.

[28] Bing Zhang, Russell D. Wolfinger, Xi Chen. Lily Wang, "An Integrated Approach for

the Analysis of Biological Pathways using Mixed Models. ," *PLoS Genetics*, vol. 4, no. 7, 2008.

[29] Leslie Cope and Giovanni Parmigiani Rosemary Braun, "Identifying differential correlation in gene/pathway combinations," *BMC Bioinformatics* , vol. 9, p. 488, 2008.

[30] Feng Tai and Wei Pan., "Incorporating prior knowledge of gene functional groups into regularized discriminant analysis of microarray data. ," *Bioinformatics* , vol. 23, pp. 3170–3177, 2007.

[31] Zervakis M, Tsiknakis M, and Kafetzopoulos D. Sfakianakis S, "Integration of Biological Knowledge in the Mixture-of-Gaussians Analysis of Genomic Clustering. ," in *10th International Conference on Information Technology and Applications in Biomedicine*, 2010.

[32] Lisa Rizzetto, Raffaele Paola, Philippe Rocca-Serra, Luca Gambineri, Cristina Battaglia, Duccio Cavalieri Luca Beltrame, "Using Pathway Signatures as Means of Identifying Similarities among Microarray Experiments," *PLoS ONE* , vol. 4, no. 1, 2009.

[33] James R. Heath, Michael E. Phelps, Biaoyang Lin Leroy Hood, "Systems Biology and New Technologies Enable Predictive and Preventative Medicine," *Science Magazine*, vol. 306, no. 5696 , pp. 640-643 , 2004.

[34] Malabat C, Weber C, Moszer I, Aittokallio T, Letondal C, Rousseau S. Clément-Ziza M, "Genoscape: a Cytoscape plug-in to automate the retrieval and integration of gene expression data and molecular networks.," *Bioinformatics*, vol. 25, no. 19, pp. 2617-2618, 2009.

[35] Ono K, Ideker T, Maere S. Smoot M, "PiNGO: a Cytoscape plugin to find candidate genes in biological networks.," *Bioinformatics*, vol. 27, no. 7, pp. 1030-1, 2011.

[36] Smoot M, Cerami E, Kuchinsky A, Landys N, Workman C, Christmas R, Avila-Campilo I, Creech M, Gross B, Hanspers K, Isserlin R, Kelley R, Killcoyne S, Lotia S, Maere S, Morris J, Ono K, Pavlovic V, Pico AR, Vailaya A, Wang PL, Adler A, Conklin BR Cline MS, "Integration of biological networks and gene expression data using Cytoscape.," *Nature Protocols*, vol. 2, no. 10, pp. 2366-82, 2007.

[37] Ye Tian1, Lu Jin1, Huai Li2, Ie-Ming Shih3, Subha Madhavan4, Robert Clarke4, Eric P. Hoffman5, Jianhua Xuan1, Leena Hilakivi-Clarke4 and Yue Wang Bai Zhang1, "DDN: A caBIG analytical tool for differential network analysis," *Bioinformatics* , vol. 27, no. 7, pp. 1036-8, 2011.

[38] Sabah Jassim, Michael A Cawthorne, Kenneth Langlands Maysson Al-Haj Ibrahim, "A Pathway-based Gene Selection Method Provides Accurate Disease Classification," *International Journal of Digital Society*, vol. 2, no. 4, pp. 566-573, 2011.

[39] A. Baudot, N. Krasnogor, A. Valencia E. Glaab, "TopoGSA: network topological gene set analysis ," *Bioinformatics*, vol. 26, no. 9, p. 1271, 2010.

[40] P. Khatri, R. P. Martins,G. C. Ostermeier, S. A. Krawetz. S. Draghici, "Global functional profiling of gene expression. ," *Genomics*, vol. 81, no. 2, pp. 98-104, 2003.

[41] Shanker Kalyana-Sundaram, Vasudeva Mahavisno, Radhika Varambally, Jianjun Yu, Benjamin B. Briggs, Terrence R. Barrette, Matthew J. Anstet, Colleen Kincead-Beal, Prakash Kulkarni, Sooryanaryana Varambally, Debashis Ghosh, Arul M. Chinnaiy Daniel R. Rhodes, "Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles," *Neoplasia* , vol. 9, no. 2, pp. 166-180, 2007.

[42] Castagnini C, Toti S, Maciag K, Kelder T, Gambineri L, Angioli S, Dolara P. Cavalieri D, "Eu.Gene Analyzer a tool for integrating gene expression data with pathway databases," *Bioinformatics*, vol. 23, no. 19, pp. 2631-2632, 2007.

[43] Dinu I, Potter JD, Liu Q, Yasui Y. Adewale AJ, "Pathway analysis of microarray data via regression. ," *Journal of Computational Biology*, vol. 15, no. 3, pp. 269-277, 2008.

[44] Michael R Kosorok. Shuangge Ma, "Detection of gene pathways with predictive power for breast cancer prognosis.," *BMC Bioinformatics*, vol. 11, no. 1, 2010.

[45] Bingbing Yuan, Fran Lewitter, Roded Sharan, Brent R. Stockwell and Trey Ideker Brian P. Kelley, "PathBLAST: a tool for alignment of protein interaction networks," *Nucleic Acids Research*, vol. 32, no. Web Server issue , pp. 83–88, 2004.

[46] Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, Franz M, Grouios C, Kazi F, Lopes CT, Maitland A, Mostafavi S, Montojo J, Shao Q, Wright G, Bader GD, Morris Q. Warde-Farley D, "The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function.," *Nucleic Acids Research*, vol. 38, no. Web server issue, pp. 214-220, 2010.

[47] Zuberi K, Rodriguez H, Kazi F, Wright G, Donaldson SL, Morris Q, Bader GD. Montojo J, "GeneMANIA Cytoscape plugin: fast gene function predictions on the desktop.," *Bioinformatics*, vol. 26, no. 22, pp. 2927-2928, 2010.

[48] Critchley-Thorne R, Lee P, Holmes S. Nacu S, "Gene expression network analysis and applications to immunology," *Bioinformatics*, vol. 23, no. 7, pp. 850-858, 2007.

[49] Xu J, Huang B, Li J, Wu X, Ma L, Jia X, Bian X, Tan F, Liu L, Chen S, Li X. Chen X, "A sub-pathway-based approach for identifying drug response principal network.," *Bioinformatics*, vol. 27, no. 5, pp. 649-54, 2011.

[50] Fidel Ramírez, Sven-Eric Schelhorn, Thomas Lengauer and Mario Albrecht Yassen Assenov, "Computing topological parameters of biological networks," *Bioinformatics*, vol. 24, no. 2, pp. 282-284, 2008.

[51] Krishnamurthy A, Karp RM, Shamir R litsky I, "DEGAS: De Novo Discovery of Dysregulated Pathways in Human Diseases.," *PLoS ONE*, vol. 5, no. 10, 2010.

[52] Kucuk H, Weile J, Wipat A, Baumbach J Alcaraz NM, "KeyPathwayMiner - Detecting case-specific biological pathways by using expression data.," *Internet Mathematics*, vol. 7, no. 4, pp. 299-313, 2011.

[53] Friedrich T, Kötzing T, Krohmer A, Müller J, Pauling J, Baumbach J. Alcaraz N, "Efficient key pathway mining: combining networks and OMICS data.," *Integrative Biology*, vol. 4, no. 7, pp. 756-764, 2012.

[54] Owen Ozier, Benno Schwikowski and Andrew F. Siegel Trey Ideker, "Discovering

regulatory and signalling circuits in molecular interaction networks," *Bioinformatics*, vol. 18, no. 1, pp. 233-240, 2002.

[55] Eunjung Lee, Yu-Tsueng Liu, Doheon Lee, Trey Ideker Han-Yu Chuang, "Network-based classification of breast cancer metastasis," *Molecular Systems Biology,* vol. 3, no. 1, 2007.

[56] Guanming Wu and Lincoln Stein, "A network module-based method for identifying cancer prognostic signatures," *Genome Biology*, vol. 13, no. 12, 2012.

[57] Sales G, Massa MS, Chiogna M, Romualdi C. Martini P, "Along signal paths: an empirical gene set approach exploiting pathway topology.," *Nucleic Acids Research*, vol. 41, no. 1, 2013.

[58] S.A, Yoo-Ah Kim, Pei, B., Ravi N, Rowe D.W. Hsin-Wei Wang, Wong A, Dong-Guk Shin. Kazmi, "Meta Analysis of Microarray Data Using Gene Regulation Pathways," *Bioinformatics and Biomedicine*, pp. 37-42, 2008.

[59] Li X,Miao Y,Wang Q,Jiang W,Xu C,Li J,Han J,Zhang F,Gong B,Xu L, Li C, "SubpathwayMiner: a software package for flexible identification of pathways," *Nucleic Acids Research*, vol. 37, no. 19, p. 131, 2009.

[60] Draghici S, Khatri P, Hassan SS, Mittal P, Kim JS, Kim CJ, Kusanovic JP, Romero R. Tarca AL, "A novel signaling pathway impact analysis.," *Bioinformatics*, vol. 25, no. 1, pp. 75-82, 2009.

[61] Calura E, Martini P, Romualdi C. Sales G, "Graphite Web: web tool for gene set analysis exploiting pathway topology.," *Nucleic Acids Research*, no. Web Server issue, pp. 89-97, 2013.

[62] Winston A., Roger Higdon, Larissa Stanberry, Dwayne Collins, and Eugene Kolker Haynes, "Differential expression analysis for pathways," *PLoS computational biology 9, no. 3*, 2013.

[63] Minoru, and Susumu Goto Kanehisa, "KEGG: kyoto encyclopedia of genes and genomes.," *Nucleic acids research*, vol. 28, no. 1, pp. 27-30, 2000.

[64] I., Larrañaga, P., Blanco, R., & Cerrolaza, A. J. Inza, "Filter versus wrapper gene selection approaches in DNA microarray domains.," *Artificial intelligence in medicine*, vol. 31, no. 2, pp. 91-103, 2004.

[65] Huiqing, Jinyan Li, and Limsoon Wong Liu, "A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns.," *Genome Informatics Series*, pp. 51-60, 2002.

[66] Ruggero G., Claire Leschi, Jérémy Besson, and Jean-François Boulicaut Pensa, "Assessment of discretization techniques for relevant pattern discovery from gene expression data.," in *BIOKDD*, Seattle, WA, USA, 2004, pp. 24-30.

[67] Alexander J., David K. Gifford, Tommi S. Jaakkola, and Richard A. Young Hartemink, "Maximum-likelihood estimation of optimal scaling factors for expression array normalization.," in *International Symposium on Biomedical Optics*, 2001, pp. 132-140.

[68] Claude E. Shannon, "A Mathematical Theory of Communication.," *Bell System*

Technical Journal, vol. 27, no. 3, pp. 379–423, 1948.

[69] G., Koumakis, L. & Moustakis, V. Potamias, "Gene Selection via Discretized Gene-Expression Profiles and Greedy Feature-Elimination.," *Methods and Applications of Artificial Intelligence. Springer Berlin Heidelberg.*, pp. 256-266, 2004.

[70] L., Weinberg, C. R., Darden, T. A., & Pedersen, L. G. Li, "Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method," *Bioinformatics*, vol. 17, no. 12, pp. 1131-1142, 2001.

[71] Andrea, Damian Szklarczyk, Sune Frankild, Michael Kuhn, Milan Simonovic, Alexander Roth, Jianyi Lin et al Franceschini, ""STRING v9. 1: protein-protein interaction networks, with increased coverage and integration.," *Nucleic acids research*, vol. 41, no. 1, 2013.

[72] D. Nishimura, "BioCarta," *Biotech Software & Internet Report: The Computer Software Journal for Scient*, vol. 2, no. 3, pp. 117-120, 2001.

[73] G., Marc Gillespie, Imre Vastrik, Peter D'Eustachio, Esther Schmidt, Bernard de Bono, Bijay Jassal et al. Joshi-Tope, "Reactome: a knowledgebase of biological pathways.," *Nucleic acids research*, vol. 33, pp. 428-432, 2005.

[74] Emek, Michael P. Cary, Suzanne Paley, Ken Fukuda, Christian Lemer, Imre Vastrik, Guanming Wu et al. Demir, "The BioPAX community standard for pathway data sharing.," *Nature biotechnology*, vol. 28, no. 9, pp. 935-942, 2010.

[75] Ethan G., Benjamin E. Gross, Emek Demir, Igor Rodchenkov, Özgün Babur, Nadia Anwar, Nikolaus Schultz, Gary D. Bader, and Chris Sander. Cerami, "Pathway Commons, a web resource for biological pathway data.," *Nucleic acids research*, vol. 39, pp. 685-690, 2011.

[76] M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M.,. & Yamanishi, Y. Kanehisa, "KEGG for linking genomes to life and the environment. ," *Nucleic acids research*, vol. 36, pp. D480-D484., 2008.

[77] Martin A., and Gert Vriend. Ott, "Correcting ligands, metabolites, and pathways.," *BMC bioinformatics*, vol. 7, no. 1, p. 517, 2006.

[78] Smoot ME, Ono K, Ruscheinski J, Wang PL, Lotia S, Pico AR, Bader GD, Ideker T. Saito R, "A travel guide to Cytoscape plugins," *Nature Methods*, vol. 9, no. 11, pp. 1069-76, Nov 2012.

[79] S. A. Kauffman, "Metabolic stability and epigenesis in randomly constructed genetic nets.," *Journal of theoretical biology*, vol. 22, no. 3, pp. 437-467, 1969.

[80] Stuart A. Kauffman, *The Origins of Order: Self-Organization and Selection in Evolution.*. New York: Oxford Univ. Press.

[81] P. & Draghici, S. Khatri, "Ontological analysis of gene expression data: current tools, limitations, and open problems," *Bioinformatics*, vol. 21, pp. 3587-95, 2005.

[82] Da Wei, Brad T. Sherman, and Richard A. Lempicki. Huang, "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.," *Nature protocols*, vol. 4, no. 1, pp. 44-57, 2008.

[83] Timothy J., Michaela A. Dinan, Mark Dewhirst, Mariano A. Garcia-Blanco, and James L. Pearson Robinson, "SplicerAV: a tool for mining microarray expression data for changes in RNA processing.," *BMC bioinformatics* , vol. 11, no. 1, p. 108, 2010.

[84] Martin Gardner, *Logic machines, diagrams and Boolean algebra*. New York: Dover Publications, 1968.

[85] C. E., Tang, L. J., Brown, K. A., & Pietenpol, J. A. Barbieri, "Loss of p63 leads to increased cell migration and up-regulation of genes involved in invasion and metastasis. ," *Cancer Research*, vol. 66, no. 15, pp. 7589-7597, 2006.

[86] H. J., & Clayton, D. G. Cordell, "Genetic association studies. ," *The Lancet*, vol. 366, no. 9491, pp. 1121-1131, 2005.

[87] C., Taminau, J., Meganck, S., Steenhoff, D., Coletta, A., Molter, C. & Nowe, A. Lazar, "A survey on filter techniques for feature selection in gene expression microarray analysis. ," *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. 9, no. 4, pp. 1106-1119, 2012.

[88] M. A. Hall, Correlation-based feature selection for machine learning, 1999, Doctoral dissertation, The University of Waikato.

[89] R., & John, G. H. Kohavi, "Wrappers for feature subset selection.," *Artificial intelligence*, vol. 97, no. 1, pp. 273-324, 1997.

[90] P. W. Baim, "A Method for Attribute Selection in Inductive Learning Systems," *Pattern Analysis and Machine Intelligence, IEEE Transactions* , vol. 10, no. 6, pp. 888-896, 1988.

[91] Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Mark Hall Eibe Frank, "The WEKA Data Mining Software: An Update;," *SIGKDD Explorations*, vol. 11, no. 1, 2009.

[92] J. R. Quinlan, *C4. 5: programs for machine learning*, 1st ed.: Morgan kaufmann, 1993.

[93] Pat Langley George H. John, "Estimating Continuous Distributions in Bayesian Classifiers.," in *Eleventh Conference on Uncertainty in Artificial Intelligence*, San Mateo, 1995, pp. 338-345.

[94] Platt J, *Fast Training of Support Vector Machines using Sequential Minimal Optimization.*, Advances in Kernel Methods - Support Vector Learning, ed., B. Schoelkopf and C. Burges and A. Smola, Ed., 1998.

[95] Shujiro, Takuji Yamada, Masami Hamajima, Masumi Itoh, Toshiaki Katayama, Peer Bork, Susumu Goto, and Minoru Kanehisa Okuda, "KEGG Atlas mapping for global analysis of metabolic pathways.," no. web server, 2008.

[96] J., Duncan, D., Shi, Z., Zhang, B Wang, "WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013," no. Web Server, 2013.

[97] O. Müller, T. Kehl, A. Gerasch, C. Backes, A. Rurainski, A. Keller, M. Kaufmann, and H. Lenhof D. Stöckel, "NetworkTrail—a web service for identifying and visualizing deregulated subnetworks.," *Bioinformatics*, vol. 29, no. 13, pp. 1702-1703, 2013.

[98] Gabriele, Enrica Calura, Paolo Martini, and Chiara Romualdi. Sales, "Graphite Web: web tool for gene set analysis exploiting pathway topology.," no. web server, 2013.

[99] Andreas, Christina Backes, Andreas Gerasch, Michael Kaufmann, Oliver Kohlbacher, Eckart Meese, and Hans-Peter Lenhof Keller, "A novel algorithm for detecting differentially regulated paths based on gene set enrichment analysis.," *Bioinformatics* , vol. 25, no. 21, pp. 2787-2794, 2009.

[100] Heidi S., Samir Kharbanda, Ruihuan Chen, William F. Forrest, Robert H. Soriano, Thomas D. Wu, Anjan Misra et al Phillips, "olecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis.," *Cancer cell*, vol. 9, no. 3, pp. 157-173, 2006.

[101] Andrew I., Tim Wiltshire, Serge Batalov, Hilmar Lapp, Keith A. Ching, David Block, Jie Zhang et al Su, "A gene atlas of the mouse and human protein-encoding transcriptomes.," *Proceedings of the National Academy of Sciences of the United States of America* , vol. 101, no. 16, pp. 6062-6067, 2004.

[102] Weinberg RA Hanahan D, "Hallmarks of cancer: the next generation.," *Cell*, vol. 144, pp. 646-674, 2011.

[103] Fang WG. Shi YH, "Hypoxia-inducible factor-1 in tumour angiogenesis. ," *World J Gastroenterol*, vol. 10, no. 8, pp. 1082-7, Apr 2004.

[104] Olga, Jiri Zavadil, Mine Esencay, Yevgeniy Lukyanov, Daniel Santovasi, Shu-Chi Wang, Elizabeth W. Newcomb, and David Zagzag. Méndez, "Knock down of HIF-1α in glioma cells reduces migration in vitro and invasion in vivo and impairs their ability to form tumor spheres," *Molecular Cancer*, vol. 9, no. 133 , 2010.

[105] David H., Susan Saporito-Irwin, Jeffrey E. DeClue, Ralf Wienecke, and Abhijit Guha Gutmann, "Alterations in the rap1 signaling pathway are common in human gliomas.," *Oncogene*, vol. 15, no. 13, pp. 1611-1616, 1997.

[106] J. Y., Lim, J. Y., Cheong, J. H., Park, Y. Y., Yoon, S. L., Kim, S. M.,. & Lee, J. S. Cho, "Gene expression signature–based prognostic risk score in gastric cancer.," *Clinical Cancer Research*, vol. 17, no. 7, pp. 1850-1857, 2011.

[107] B., & Kinzler, K. W. Vogelstein, "Cancer genes and the pathways they control.," *Nature medicine*, vol. 10, no. 8, pp. 789-799, 2004.

[108] Robert J., Meyer Michael Cohen, and Raoul CM Hennekam Gorlin, *Syndromes of the Head and Neck.*, 819th ed. New York: Oxford University Press, 1990.

[109] Virginia, June-Anne Gold, Trevor L. Hoffman, Jayesh Panchal, and Simeon A. Boyadjiev Kimonis, "Genetics of craniosynostosis.," *In Seminars in pediatric neurology*, vol. 14, no. 3, pp. 150-161, 2007.

[110] Serti Eacute AE, Jehee FS, Fanganiello R, Yeh E Passos-Bueno MR, "Genetics of craniosynostosis: genes, syndromes, mutations and genotype-phenotype correlations.," *Frontiers of Oral Biology*, vol. 12, pp. 107–143, 2008.

[111] Brendan David, Brig Mecham, Sarah S. Park, H. Wilkerson, Federico M. Farin, Richard P. Beyer, Theo K. Bammler, Lara M. Mangravite, and Michael L. Cunningham Stamper, "Transcriptome correlation analysis identifies two unique

craniosynostosis subtypes associated with IRS1 activation.," *Physiological genomics*, vol. 44, no. 23, pp. 1154-1163, 2012.

[112] Sherman BT, Lempicki RA. Huang DW, "Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. ," *Nature Protocols*, vol. 4, no. 1, pp. 44-57, 2009.

[113] [Online]. http://www.ncbi.nlm.nih.gov/gene?Db=gene&Cmd=ShowDetailView&TermToSearch=2252

[114] Cécilie, Hind Guenou, Karim Kaabeche, Daniel Bouvard, Archana Sanjay, and Pierre J. Marie Dufour, "FGFR2-Cbl interaction in lipid rafts triggers attenuation of PI3K/Akt signaling and osteoblast survival.," *Bone*, vol. 42, no. 6, pp. 1032-1039, 2008.

[115] Anne, Richard Jäger, Angela Egert, Wolfram Kress, Eva Wardelmann, and Hubert Schorle. Moenning, "Sustained platelet-derived growth factor receptor alpha signaling in osteoblasts results in craniosynostosis by overactivating the phospholipase C-gamma pathway.," *Molecular and cellular biology*, vol. 29, no. 3, pp. 881-891, 2009.

[116] Lance D., Johanna Smeds, Joshy George, Vinsensius B. Vega, Liza Vergara, Alexander Ploner, Yudi Pawitan et al. Miller, "An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 38, pp. 13550-13555, 2005.

[117] Yixin, Jan GM Klijn, Yi Zhang, Anieta M. Sieuwerts, Maxime P. Look, Fei Yang, Dmitri Talantov et al. Wang, "Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer.," *The Lancet 365*, vol. 9460 , pp. 671-679, 2005.

[118] Christos, Pratyaksha Wirapati, Sherene Loi, Adrian Harris, Steve Fox, Johanna Smeds, Hans Nordgren et al. Sotiriou, "Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis.," *Journal of the National Cancer Institute*, vol. 98, no. 4, pp. 262-272, 2006.

[119] Piette F, Loi S, Wang Y et al. Desmedt C, "Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. ," *Clin Cancer Res*, vol. 13, no. 11, pp. 3207-14, Jun 2007.

[120] Lei, Aik C. Tan, Raimond L. Winslow, and Donald Geman. Xu, "Merging microarray data from separate breast cancer studies provides a robust prognostic test.," *Bmc Bioinformatics*, vol. 9, no. 1, p. 125, 2008.

[121] Chengsen, Jeffrey Wyckoff, Fubo Liang, Mazen Sidani, Stefania Violini, Kun-Lin Tsai, Zhong-Yin Zhang, Erik Sahai, John Condeelis, and Jeffrey E. Segall. Xue, "Epidermal growth factor receptor overexpression results in increased tumor cell motility in vivo coordinately with enhanced intravasation and metastasis.," *Cancer research*, vol. 66, no. 1, pp. 192-197, 2006.

[122] R.L: Sutherland, "Endocrine resistance in breast cancer: new roles for ErbB3 and ErbB4. ," *Breast Cancer Research*, vol. 13, no. 3, p. 106, 2011.

[123] I.R., et al.: Hutcheson, "Heregulin beta1 drives gefitinib-resistant growth and invasion in tamoxifen-resistant MCF-7 breast cancer cells.," *Breast Cancer Research*, vol. 9, no. 4, p. 50, 2007.

[124] D. P. Bartel, "MicroRNAs: genomics, biogenesis, mechanism, and function," *Cell*, vol. 116, pp. 281–297, 2004.

[125] Lee P., Nelson C. Lau, Philip Garrett-Engele, Andrew Grimson, Janell M. Schelter, John Castle, David P. Bartel, Peter S. Linsley, and Jason M. Johnson Lim, "Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs.," *Nature*, vol. 433, no. 7027, pp. 769-773, 2005.

[126] Marilena V., and Carlo M. Croce Iorio, "MicroRNA dysregulation in cancer: diagnostics, monitoring and therapeutics. A comprehensive review.," *EMBO molecular medicine*, vol. 4, no. 3, pp. 143-159, 2012.

[127] Huiping Liu, "MicroRNAs in breast cancer initiation and progression.," *Cellular and Molecular Life Sciences*, vol. 69, no. 21, pp. 3587-3599, 2012.

[128] Pai-Sheng, Jen-Liang Su, and Mien-Chie Hung Chen, "Dysregulation of microRNAs in cancer.," *J Biomed Sci*, vol. 19, no. 1, p. 90, 2012.

[129] Ramiro, George A. Calin, and Carlo M. Croce. Garzon, "MicroRNAs in cancer," *Annual review of medicine*, vol. 60, pp. 167-179, 2009.

[130] Qinghua, Zhenbao Yu, Enrico O. Purisima, and Edwin Wang Cui, "Principles of microRNA regulation of a human cellular signaling network.," *Molecular systems biology*, vol. 2, no. 1, 2006.

[131] Backes C, Nourkami-Tutdibi N, Leidinger P et al Schmitt J, "Treatment-independent miRNA signature in blood of Wilms tumor patients.," *BMC Genomics*, vol. 7, no. 13, p. 379, Aug 2012.

[132] Ana, and Sam Griffiths-Jones Kozomara, "miRBase: integrating microRNA annotation and deep-sequencing data.," *Nucleic acids research*, p. 1027, 2010.

[133] Chabane, and Edwin Wang. Tibiche, "MicroRNA regulatory patterns on the human metabolic network.," *The Open Systems Biology Journal*, vol. 1, pp. 1-8, 2008.

[134] Karin, Ruth Zeidman, Ulrika Trollér, Anna Schultz, and Christer Larsson Svensson, "Protein kinase C beta1 is implicated in the regulation of neuroblastoma cell growth and proliferation.," *Cell growth & differentiation: the molecular biology journal of the American Association for Cancer Research*, vol. 11, no. 12, pp. 641-648, 2000.

[135] Ruth, Bjarne Löfgren, Sven Påhlman, and Christer Larsson. Zeidman, "PKCε, via its regulatory domain and independently of its catalytic domain, induces neurite-like processes in neuroblastoma cells.," *The Journal of cell biology*, vol. 145, no. 4, pp. 713-726, 1999.

[136] Yuqing, Boyi Gan, Debra Liu, and J. H. Paik. Zhang, "FoxO family members in cancer.," *Cancer Biol Ther*, vol. 12, no. 4, pp. 253-259, 2011.

[137] Evan E., Peter Stroeken, Peter V. Sluis, Jan Koster, Rogier Versteeg, and Ellen M. Westerhout Santo, "FOXO3a is a major target of inactivation by PI3K/AKT

signaling in aggressive neuroblastoma.," *Cancer research*, vol. 73, no. 7, pp. 2189-2198, 2013.

[138] Robert E., Dongfeng Tan, Jeffrey S. Taylor, Michal Miller, Jeffrey W. Prichard, and Marylee M. Kott. Brown, "Morphoproteomic confirmation of constitutively activated mTOR, ERK, and NF-kappaB pathways in high risk neuro-blastoma, with cell cycle and protein analyte correlates.," *Annals of Clinical & Laboratory Science*, vol. 37, no. 2, pp. 141-147, 2007.

[139] Chiara, Danilo Marimpietri, Fabio Pastorino, Beatrice Nico, Daniela Di Paolo, Michela Cioni, Federica Piccardi et al Brignole, "Effect of bortezomib on human neuroblastoma cell growth, apoptosis, and angiogenesis.," *Journal of the National Cancer Institut*, vol. 98, no. 16 , pp. 1142-1157, 2006.

[140] Fichtner I, Behrens D, Haider W, Rothweiler F, Mack A, Cinatl J, Doerr HW, Cinatl J Jr Michealis M, "Anti-cancer effects of bortezomib against chemoresistant neuroblastoma cell lines in vitro and in vivo.," *Int J Oncol*, vol. 28, pp. 439–446, 2006.

[141] N.Graf, "Biomarker and Wilms Tumor.," *Highlight Pediatr Blood Cancer*, vol. 61, no. 2, pp. 185–186, Nov 2-13.

[142] Norbert, Harm van Tinteren, Christophe Bergeron, François Pein, Marry M. van den Heuvel-Eibrink, Bengt Sandstedt, Jens-Peter Schenk et al Graf, "Characteristics and outcome of stage II and III non-anaplastic Wilms' tumour treated according to the SIOP trial and study 93-01.," *European Journal of Cancer*, vol. 48, no. 17, pp. 3240-3248, 2012.

[143] Olaf, Susanne Hämmerling, Clemens Stockklausner, Jens-Peter Schenk, Patrick Günther, Wolfgang Behnisch, Bajes Hamad, Naima Ali Al Mulla, and Andreas Kulozik Witt, "13-cis retinoic acid treatment of a patient with chemotherapy refractory nephroblastomatos," *Journal of pediatric hematology/oncology*, vol. 31, no. 4, pp. 296-299, 2009.

[144] Philip RO Payne, "Biomedical Knowledge Integration," vol. 8, no. 12, 2012.

[145] Russ B. Altman, "Introduction to Translational Bioinformatics Collection.," vol. 12, no. 8, 2012.

Appendix I (KEGG pathways)

KEGG is part of the GenomeNet[49] project of the Kyoto University. KEGG initiated in 1995 for sequence information from a number of organisms into metabolic or regulatory pathways. KEGG consists of 4 main databases: PATHWAY, GENES, LIGAND, and BRITE.

The KEGG PATHWAY database is a collection of manually drawn graphical diagrams, called KEGG pathway maps, representing molecular pathways for metabolism, genetic information processing, environmental information processing, other cellular processes, human diseases, and drug development. Pathway maps are based on extensive survey of published literature. If available, different organisms are compared. The pathway map is drawn and updated with the notation shown below.

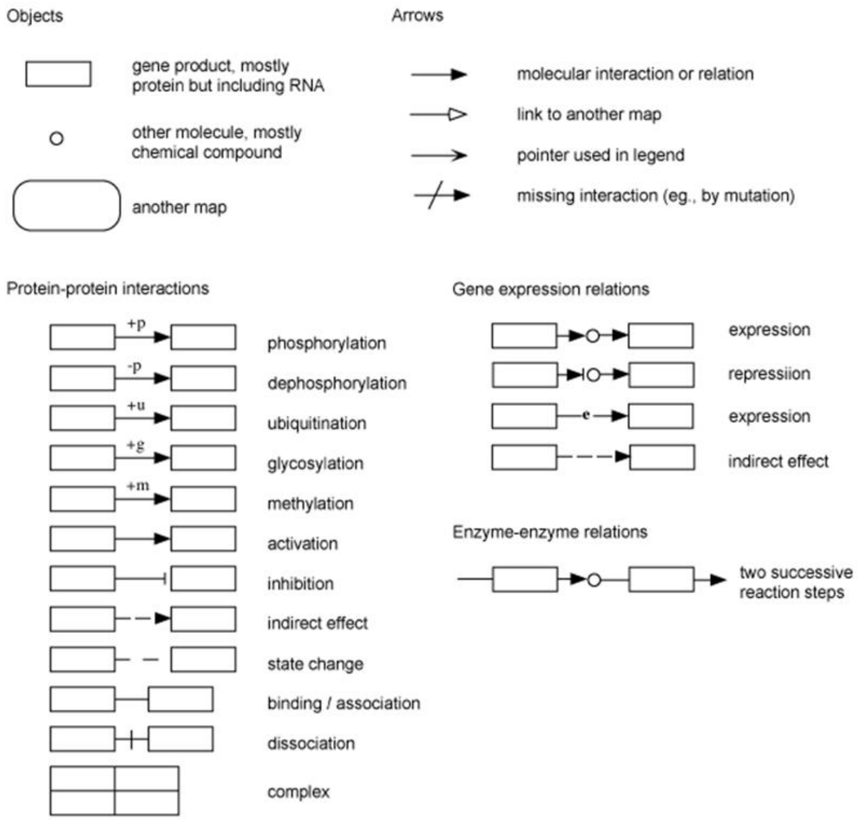


Figure 59: KEGG pathway notations (source KEGG help documentation)

There are two types of KEGG pathways, (i) reference pathways which are manually drawn and (ii) organism-specific pathways which are computationally generated based on reference pathways.

[49] http://www.genome.jp/

In the organism-specific pathways, green boxes are hyperlinked to GENES entries by converting K numbers (KO identifiers) to gene identifiers in the reference pathway, indicating the presence of genes in the genome and also the completeness of the pathway.

Maps are available both as GIF-files and as XML version. These KEGG Markup Language (KGML) files contain computerized information about graphical objects and their relations in the KEGG pathways as well as information about orthologous gene assignments in the KEGG GENES database. Each pathway is identified by a five-digit number preceded by one of: *map, ko, ec, rn*, and three- or four-letter organism code.

In KGML the pathway element specifies one graph object with the entry elements as its nodes and the relation and reaction elements as its edges. The relation and reaction elements indicate the connection patterns of rectangles (gene products) and the connection patterns of circles (chemical compounds), respectively, in the KEGG pathways. The two types of graph objects, those consisting of entry and relation elements and those consisting of entry and reaction elements, are called the protein network and the chemical network, respectively. Since the metabolic pathway can be viewed both as a network of proteins (enzymes) and as a network of chemical compounds, another distinction of KEGG pathways is:

- metabolic pathways viewed as both protein networks and chemical networks
- regulatory pathways viewed as protein networks only

The following figure shows an overview of KGML.

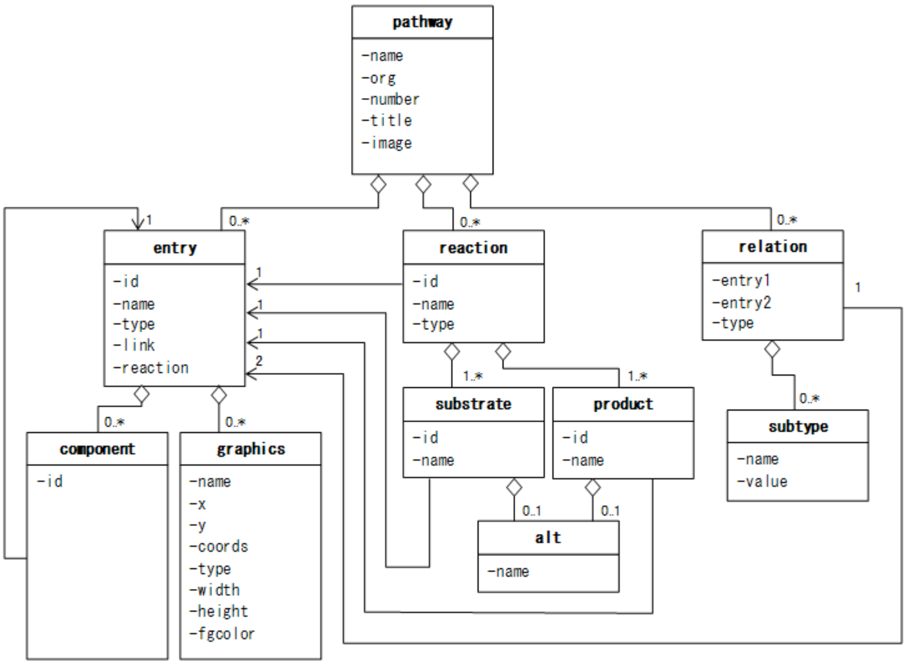


Figure 60: Overview of KGML schema (source the KEGG Markup Language)

The pathway element is a root element, and one pathway element is specified for one pathway map in KGML. The entry, relation, and reaction elements specify the graph information, and additional elements are used to specify more detailed information about nodes and edges of the graph.

More details can be found in the KEGG Markup Language[50].

The pathway maps are classified into the following sections:

- Metabolism
- Genetic information processing (transcription, translation, replication and repair, etc.)
- Environmental information processing (membrane transport, signal transduction, etc.)
- Cellular processes (cell growth, cell death, cell membrane functions, etc.)
- Organismal systems (immune system, endocrine system, nervous system, etc.)
- Human diseases
- Drug development

The metabolism section contains aesthetically drawn global maps showing an overall picture of metabolism, in addition to regular metabolic pathway maps. The low-resolution global maps can be used, for example, to compare metabolic capacities of different organisms in genomics studies and different environmental samples in metagenomics studies. In contrast, KEGG modules in the KEGG MODULE database are higher-resolution, localized wiring diagrams, representing tighter functional units within a pathway map, such as sub-pathways conserved among specific organism groups and molecular complexes. KEGG modules are defined as characteristic gene sets that can be linked to specific metabolic capacities and other phenotypic features, so that they can be used for automatic interpretation of genome and metagenome data.

[50] http://www.kegg.jp/kegg/xml/docs/

Appendix II (Datasets)

MinePath uses microarray experiments and respective gene-expression data to identify discriminant sub-paths in known GRNs. Currently provides 15 public gene expression datasets from the Gene Expression Omnibus database for 6 different disease categories naming (i) breast cancer, (ii) leukemia, (iii) craniosynostosis, (iv) lung cancer, (v) colon cancer and (vi) mental disorder. The user can select one of the annotated datasets or upload his/her own dataset. Details for the preparation of a private dataset and upload to the MinePath server can be found in section 3.5.2.1.1 (Select or upload gene expression dataset).

In the following sections we describe in short the (currently) available datasets.

Breast cancer

Most of the datasets currently available in the web based MinePath application fall into the breast cancer category. This series represents 180 lymph-node negative relapse free patients and 106 lymph-node negate patients that developed a distant metastasis.

GSE2034

GSE2034 dataset [117] comes from a breast cancer relapse free survival study. The Erasmus Medical Center (Rotterdam, Netherlands) tumour bank used for the frozen tumour samples from patients with lymph-node-negative breast cancer who were treated during 1980–1995, but who did not receive systemic neoadjuvant or adjuvant therapy. Tumour samples were submitted to the laboratory from 25 regional hospitals for measurements of steroid-hormone receptors. Analysis conducted with Affymetrix Human U133a GeneChips, the expression of 22 000 transcripts from total RNA of the frozen tumor samples.

GSE2990

The patients coming from Uppsala Hospital have been also used in other studies as in GSE3494. The dataset contains 64 microarray experiments from primary breast tumours used in the original publication [118] as training set to identify genes differentially expressed in grade 1 and 3 and 129 microarray experiments from primary breast tumours of untreated patients used as validation set to validate the list of genes and its correlation with survival. No replicate, no reference sample in the dataset. Analysis conducted with Affymetrix Human U133a GeneChips.

GSE3494

The biological tumour samples (breast tumour specimens) consisted of freshly frozen breast tumours from a population-based cohort of 315 women representing 65% of all breast cancers resected in Uppsala County, Sweden, from January 1, 1987 to December 31, 1989 [116]. Oestrogen receptor status was determined by biochemical assay as part of the routine clinical procedure. All tumour specimens were assessed on Affymetrix Human U133 A and B arrays.

GSE7390

Gene expression profiling of frozen samples from 198 lymph node-negative systemically untreated breast cancer patients was done at the Bordet Institute, blinded to clinical data and independent of Veridex. The Veridex organization is dedicated to providing physicians with high-value in vitro diagnostic oncology products, including CELLSEARCH[51] Circulating Tumour Cell testing for more than a decade. Genomic risk was defined by Veridex, blinded to clinical data. The original paper [119] tried to predict distant metastases and the study conducted by TRANSBIG project.

E-GEOD-13671

The E-GEOD-13671 dataset included duplicates from four normal controls and from two BRCA1 mutation carriers and single arrays from another two BRCA1 mutation carriers using a three-dimensional culture technique to grow mammary epithelial cells ex vivo. Ten colonies were collected and RNA was isolated using the Absolutely RNA Nanoprep kit (Stratagene). Samples were hybridized to the Human Genome U133 Plus 2.0 (Affymetrix) at the Partners Genomics Centre.

E-GEOD-20685

The primary goal of this study is to identify molecular subtypes of breast cancer through gene expression profiles of 327 breast cancer samples and determine molecular and clinical characteristics of different breast cancer subtypes. Expression signatures of different cellular functions (e.g., cell proliferation/cell cycle, wound response, tumor stromal response, vascular endothelial normalization, drug esponse genes, etc.) in different breast cancer molecular subtypes investigated and assessed how microarray-based breast cancer molecular subtypes may be used to guide treatment. Gene expression profiles of 327 breast cancer samples were determined using total RNA and Affymetrix U133 plus 2.0 arrays.

[51] https://www.cellsearchctc.com

GSE22035

43 ER-positive breast tumours including 14 tumours with PIK3CA mutations and 29 tumours without PIK3CA mutations were used as screening set for microarray. PI3K/AKT pathway plays one of pivotal roles in breast cancer development and maintenance. The ERα-positive breast tumours PIK3CA mutations have been observed in 30% to 40%. However, genes expressed in connection to the pathway activation in breast tumorigenesis remain largely unknown. Samples were hybridized to the Affymetrix U133 plus 2.0 arrays

4ERdatasets

The '4ERdatasets' dataset is a set of four independent discretized and then merged gene-expression studies targeting the ER phenotypic status respective patients, i.e., ER+ (ER positive) vs. ER- (ER negative), from the GSE2034, GSE2990 GSE3494 and GSE7390studies.

The four datasets used the same hybridization platform, the GPL96 HG-U133A Affymetrix Human Genome U133A Array, making the procedure of merging relatively easy. For the discretization, the same methodology as in MinePath was used in the level of probes. Each dataset was discretized individually and then the four datasets were merged.

Leukaemia

Leukaemia in MinePath is currently represented by one dataset, the GSE18239, an expression data from JAK1 wild-type and JAK1 mutation-positive T cell acute lymphoblastic leukaemia blasts. The Janus kinase 1 (JAK1) gene encodes a cytoplasmic tyrosine kinase that noncovalently associates with a variety of cytokine receptors and plays a nonredundant role in lymphoid cell precursor proliferation, survival, and differentiation. Somatic mutations in JAK1 occur in individuals with acute lymphoblastic leukemia. The study used microarray to compare the gene expression profile of JAK1 mutation positive or negative acute lymphoblastic leukaemia blasts. The hybridization platform Human Genome U133 Plus 2.0 (Affymetrix) was used.

Glioma

A glioma is a type of tumor that starts in the brain or spine. Three types of normal glial cells can produce tumors—astrocytes, oligodendrocytes, and ependymal cells. These tumors are usually highly malignant (cancerous) because the cells reproduce quickly and they are supported by a large network of blood vessels. In the adult population, glioblastoma multiforme (GBM), is a common and one of the most malignant primary brain tumors, representing up to 50% of all primary brain gliomas[52].

In MinePath you can find a dataset which is a merging of two different studies using as classes the glioma cases from the GSE4271 (100 samples) versus the control cases from the GSE1133 (158 samples).

Craniosynostosis

Craniosynostosis is a disease defined by premature fusion of one or more cranial sutures. In MinePath currently we can find one annotated dataset for craniosynostosis, the GSE27976. In this study, gene expression data from 199 patients with isolated sagittal (n= 100), unilateral coronal (n = 50), and metopic (n = 49) synostosis are compared (all together) against a control population (n = 50). For the study, the HuGene-1_0-st Affymetrix Human Gene 1.0 ST Array [transcript (gene) version] was used.

Lung cancer

Lung cancer in MinePath is currently represented by one dataset containing sixty pairs of tumour and adjacent normal lung tissue specimens from non-smoking female lung cancer patients who were admitted to National Taiwan University Hospital or Taichung Veterans General Hospital were analysed by using GeneChip Human Genome U133 Plus 2.0 expression arrays (Affymetrix) by Partek (Partek, Inc.) for mRNA expression levels. The mean ± SD age of patients used for microarray experiments was 61 ± 10 years. Most of the tumours were adenocarcinomas (93%), and 78% of the samples were in stage I or II. Because the cancer and normal tissues were from the same individual, paired t tests and Bonferroni post hoc P value adjustment were used.

Colon cancer

In MinePath currently we can find one annotated dataset for colon cancer. The specific dataset (GSE4107) extracted RNA from colonic mucosa of healthy controls (10samples) and patients (12samples) were analysed using Affymetrix Human Genome U133 Plus 2.0 Array. Patients and controls were age- (50 or less), ethnicity- (Chinese) and tissue-matched.

Tumour specimens and adjacent grossly normal-appearing tissue at least 8 cm away were routinely collected and archived from patients undergoing colorectal resection at the Singapore General Hospital. Young (≤50 years old) Chinese pa-

[52] CBTRUS. Statistical Report: Primary Brain Tumors in the United States. 1998–2002:2005. Central Brain Tumor Registry of the United States.

tients whose tumours were classified as microsatellite-stable were included in this retrospective study. Five to seven pinch biopsies from several locations throughout the colon were obtained from Chinese individuals (≤50 years old) undergoing colonoscopic examination and were found to have no polyps and no known family history or previous CRC incidence: these were designated as healthy controls.

Mental disorder

The GSE12649 mental disorder study has been annotated and uploaded in the MinePath web platform. Since the dataset contains three phenotypical categories and MinePath operates over datasets with two phenotypes, we split the study data into three independent datasets, the bipolar disorder versus control, the schizophrenia versus control and the bipolar disorder versus schizophrenia.

The study is screened a total of 102 postmortem brains obtained from the Stanley Medical Research Institute were used for DNA microarray analysis. Fresh frozen samples were used for RNA extraction. RNA samples extracted from the prefrontal cortices Broadmann's Area 46 (part of the frontal cortex in the human brain). They contain total RNA samples from 35 individuals in each of the three diagnostic groups, bipolar disorder, schizophrenia and controls. Diagnoses had been made according to the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV; American Psychiatric Association).