Journal of
Data Mining in Genomics & Proteomics

Yepes and Torres, J Data Mining Genomics Proteomics 2016, 7:1
DOI: 10.4172/2153-0602.1000185

Review Article Open Access

Mining Datasets for Molecular Subtyping in Cancer

Sally Yepes* and Maria Mercedes Torres

Facultad de Ciencias, Departamento de Ciencias Biológicas, Universidad de los Andes, Bogotá D.C., Colombia.

Abstract

Given the heterogeneity in the clinical behavior of cancer patients with identical histopathological diagnosis, the search for unrecognized molecular subtypes, subtype-specific markers and the evaluation of their clinical-biological relevance are a necessity. This task is benefiting today from the high-throughput genomic technologies and free access to the datasets generated by the international genomic projects and the repositories of information. Machine learning strategies have proven to be useful in the identification of hidden trends in large datasets, contributing to the understanding of the molecular mechanisms and subtyping of cancer. However, the translation of new molecular subclasses and biomarkers into clinical settings requires their analytic validation and clinical trials to determine their clinical utility. Here, we provide an overview of the workflow to identify and confirm cancer subtypes, summarize a variety of methodological principles, and highlight representative studies. The generation of public big data on the most common malignancies is turning the molecular pathology into a database-driven discipline.

Introduction

The diagnosis of cancer is made primarily through histopathological classification systems that take into account the morphological characteristics of the tumor, allowing their identification and clinical stage assignment. The histopathological classification systems, despite their contribution to reducing cancer-related mortality, still present uncertainties in providing prognostic information or guidance for determining the most appropriate therapeutic direction. It is also clear that such systems fail to provide information about the underlying molecular mechanisms, which may be the origin of the clinically observed differences. Furthermore, the existing histopathological subtypes are heterogeneous; this is evident at the levels of molecular pathogenesis, clinical course, and treatment responsiveness [1,2]. These limitations necessitate the discovery of new molecular subtypes and the evaluation of their clinical and biological significance. Low-dimensional approaches that consider a limited number of genes and patients are insufficient to address the problem of cancer subtyping. It is necessary to identify patterns in large datasets and at a genome-wide scale using machine learning strategies. This task benefits from the high-throughput genomic technologies, the enormous amount of genomic datasets generated by the international genomic projects, and the availability of data analysis algorithms, allowing a comprehensive and unprecedented characterization of the disease.

The machine learning approaches can be used to dissect the complexity of cancer. These are the computational tools that recognize and classify patterns based on models derived from the data. The motivation for this mini review is provide an overview of the workflow for molecular subtyping in cancer. Although there are various methods available for classification, a common analytical framework emerges across several research studies. This common workflow with its outstanding techniques is covered here with interest in the methodological principles and the biological interpretation.

Despite great efforts in cancer biomarkers several factors have impeded translation of research findings into clinical practice [3]. The precise role in the management of patients, of new molecular subclasses and predictors identifies by machine learning approaches, need to be refined and strongly validated. However, it is clear that they have the potential to provide insights about the underling molecular mechanisms and help to dissect the molecular heterogeneity of the disease. Machine learning for cancer subtyping has been performed mainly with expression data. However, this technique can also be applied to other levels of biological information, such as promoter methylation, miRNAs, and single nucleotide polymorphisms, analyzed with hybridization array technology or next generation sequencing, allowing the study of the data structure in many different levels and providing an integrated view of the biological processes involved.

Unsupervised and Supervised Learning for Cancer Study

There two main types of statistical problems associated with tumor classification: the identification of unknown tumor classes and the classification of malignancies into known classes. These two issues can be addressed using in a complementary manner unsupervised and supervised machine learning. These methodologies have supported the discovery of subgroups with biological significance and clinical implications in multiple types of cancer (Table 1). Representative examples include: 1) the distinction between acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML) without the previous knowledge of the classes they belong to [4], a distinction that is critical for successful treatment. The class discovery was made through the use of self-organizing maps (SOM); the clusters identified by this method were compared to the known classes through linear discriminant analysis [4]. 2) The subclassification of diffuse large B-cell lymphoma (DLBCL) into groups related to the differences in the stages of B-cell differentiation (germinal center, activated B -cell) and its association with differences in patient prognosis [5,6]. Average linkage hierarchical clustering was used in this case as the unsupervised strategy. The use of hierarchical clustering analysis and supervised analysis also allowed the classification of breast cancer into at least 4 subtypes (basal, luminal A, luminal B, HER2+) [7,8]. The refinement and comprehensive study of these subtypes has allowed the identification of differences with regard to clinical features,

***Corresponding author:** Sally Yepes, Facultad de Ciencias, Departamento de Ciencias Biologicas, Universidad de los Andes, Bogota DC, Colombia, Tel: +57-1-3394949; Fax: +57-1-2841890 E-mail: sl.yepes233@uniandes.edu.co

Received December 21, 2015; **Accepted** January 12, 2016; **Published** January 20, 2016

Citation: Yepes S, Torres MM (2016) Mining Datasets for Molecular Subtyping in Cancer. J Data Mining Genomics Proteomics 7: 185. doi:10.4172/2153-0602.1000185

Table 1: Some machine learning methods used in cancer subtyping.

Reference	Clustering method	Classification method	Cancer	Clinico-biological significance
Golub TR et al., 1999 [4]	Self-organizing maps	Linear discriminant analysis	ALL vs. AML	A distinction that is critical for successful treatment
Alizadeh AA et al., 2000 [5]	Hierarchical clustering	N/A	DLBCL	Subtypes with differences in the stages of B-cell differentiation (germinal center, activated B-cell) and their relation with prognosis
Perou CM et al., 2000 [8]; Sorlie T et al., 2003 [7]; van 't Veer LJ et al., 2002 [9]; van de Vijver MJ et al., 2009 [10]	Hierarchical clustering	Various: prediction analysis of microarrays, correlation with categories	Breast cancer	Subtypes (basal, luminal A, luminal B, HER2+) with differences with regard to clinical features, response to treatment, and prognosis.
Salazar R et al., 2011 [12]	N/A	A nearest mean classifier	CRC	Classifier that can predict patients with early-stage CRC
Collisson EA et al., 2011 [28]	Non-negative matrix factorization	N/A	PDA	Subtypes of PDA and their responses to therapy
Sadanandam A et al., 2013 [29]	Non-negative matrix factorization	Prediction analysis of microarrays	CRC	Subtypes of CRC and their relation with cellular phenotype and response to therapy
Armstrong SA et al., 2003 [40]	N/A	K-nearest-neighbors	MLL	Rearrangement of the MLL gene distinguishes a unique leukemia
Tan IB et al., 2011 [58]	Non-negative matrix factorization	Support vector machines	GC	Subtypes of GC and their relation with survival and response to chemotherapy
Budinska et al., 2013 [59]	Hierarchical clustering	Multiclass linear discriminant analysis	CRC	Subtypes with molecular heterogeneity

ALL: acute lymphoblastic leukemia, AML: acute myeloid leukemia, DLBCL: diffuse large B-cell lymphoma, CRC: colorectal cancer, PDA: pancreatic ductal adenocarcinoma, MLL: mixed lineage leukemia, GC: gastric cancer

response to treatment, and prognosis. 4) The supervised classification approaches have had a major impact on the ability to influence clinical management as they help predict the outcome of the disease; most efforts have focused on identifying prognostic (clinical outcome) and predictive (response to treatment) markers. Supervised predictors have motivated the development of large-scale validation efforts, especially in breast and colon cancer. Predictors in the areas of breast and colon cancer (*MammaPrint*, *Oncotype DX*, *Coloprint*) [9-12] have been noted for their progress in clinical trials.

A common methodological framework to identify subtypes can be found in many studies [13], while it does not use the same analytical techniques, follows a similar workflow: discovery cohorts are chosen and pre-processed, unsupervised clustering techniques are applied, supervised classifiers are developed, and clustering and classification are validated in independent cohorts (Figure 1).

Cohorts and data pre-processing

The starting point should be the selection of characterized cohorts based on histopathological evaluation and clinical monitoring. The clinical relevance of a classification system lies mainly in the stratification of patients based on clinical outcomes such as survival or therapeutic response (if it is the case of cohorts with therapeutic intervention). The subdivisions are evaluated by methods such as the Kaplan-meier estimator or the cox proportional hazard model, for this purpose, it is necessary to have information about the current status of the patient as well as during long clinical follow-up periods, which will allow the estimation of survival endpoints (e.g. overall survival, recurrence-free survival). In the case of microarray data, raw data are pre-processed in a process that involves three steps: background correction to adjust the intensity readings for nonspecific signals; adjustment of the intensity readings for technical variability to ensure that the measurements of all samples are comparable (normalization); and computation of a summary value for the different probes representing each gene (summarization). The most commonly used types of normalization are the RMA [14], the Quantile [15], the Loess [16], and the VNS [17,18]. A filtration step is recommended for removing non- informative genes that may represent noise for clustering; genes that show little variation between patients and those with low signal intensity should be eliminated [19].

Given the heterogeneity of the disease, the cohort must represent a broad spectrum of molecular events to realize the identification of rare subgroups. Therefore, cohorts with a large number of patients are preferred. Because of the difficulty of finding large cohorts, it is a common practice to combine normalized datasets, which will have the challenge of making corrections for non-biological variation such as differences in origin and technical processing. Several correction or adjustment processes have been proposed. These include SVA [20], Combat [21], and DWD [22]. The use of different platforms (e.g. Affymetrix, Agilent) is also a limiting factor when combining cohorts, requiring homogeneous annotation of the probes corresponding to the same gene symbol.

Clustering of patients

The goal of clustering is the identification of natural or inherent structures in a dataset. The clustering divides patients into groups that may represent subtypes of the disease, using measures of distance or similarity (e.g. geometrical distances or correlation). The cluster analysis employs two basic strategies, namely hierarchical clustering and partitioning, in addition to hybrid methods.

The hierarchical clustering is one of the most widely used methods. This generates tree-like structures between elements and can be divided into agglomerative (bottom-up) and divisive (top-down). In bottom-up methods, each data set is considered a cluster. The clusters are iteratively grouped based on their similarity measures. The top-down method starts with a cluster containing all data, and splits are performed recursively until each cluster contains a single data. The type of clustering algorithm, the distance metric, the type of linkage or inter-cluster distances (when appropriate), and the number of clusters must be selected when employing these methods; guidance can be found in the work of Drăghici S [23,24]. The use of a hierarchical grouping system for the subdivision of the data may involve the notion of development or transition between the elements, whereas a non-hierarchical system may involve greater independence or independent

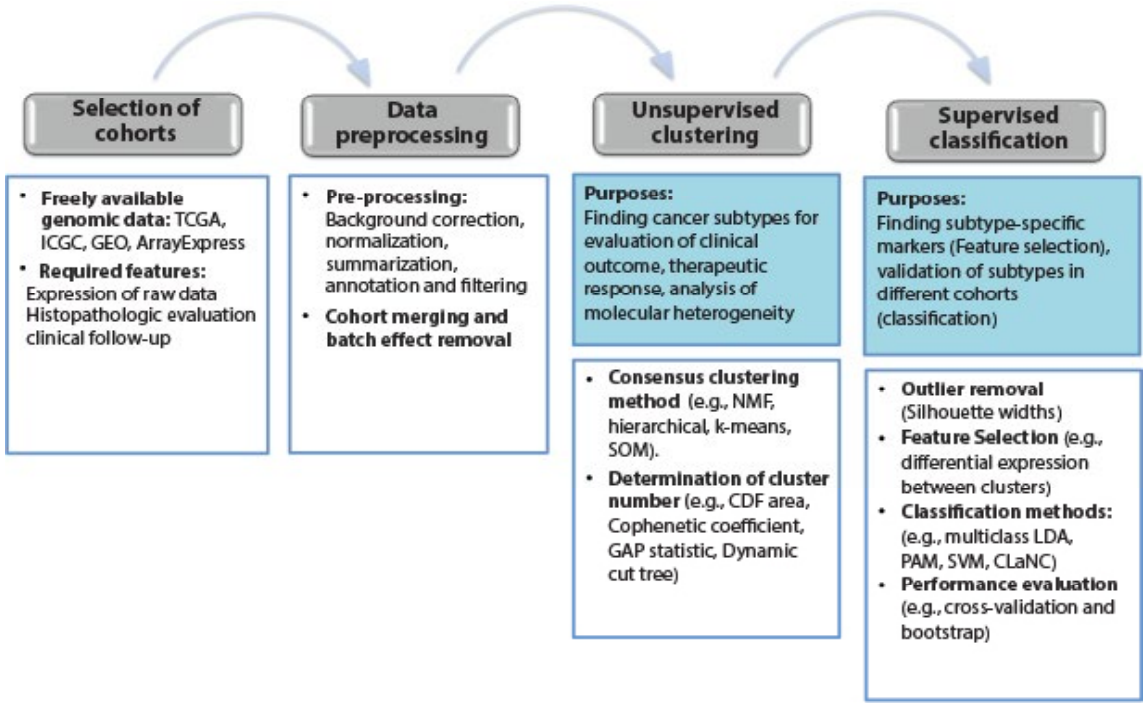

Figure 1: A common methodology workflow to identify and confirm cancer subtypes. The unsupervised or class discovery methods can be used to identify the unrecognized tumor subtypes, and the supervised or classification methods can be used to allocate the samples into previously defined classes. The entire workflow also can help to identify potential biomarkers, through feature selection methods applied before classification.

emergence. On the other hand, partitioning methods are a group of methods based on a variety of mathematical models. Among the most widely used methods of partitioning are the self-organizing maps (SOM) and the K-means [24].

SOM is a type of artificial neural network (ANN), which computes a set of reference vectors (prototypes) representing the local means of the data. SOM then partitions the data set, with each prototype defining a cluster consisting of the data points nearest to it. The user specifies the number of clusters to be identified [25]. The k-means starts with a fixed a priori number of cluster centers (k). Each data point is assigned to the nearest center, based on its distance from each center, to form a set of temporary clusters. An iterative process recalculates the position of the cluster centers based on the current membership of each cluster and reassigns the points to the k clusters. This process continues until stabilization is achieved [26]. The most common methods for identifying robust subgroups (tolerant to outliers) include the use of a clustering algorithm together with a consensus clustering process originally proposed by Monti et al. [27]. The consensus clustering performs sub-sampling and, for each subsample, runs a particular type of clustering, estimating consensus values for different numbers of clusters, allowing the assessment of the stability of the clusters discovered. The number of clusters can be determined from the empirical cumulative distribution function (CDF) area [27].

Recently, the non-negative matrix factorization (NMF) consensus clustering has been extensively used [28-33]. The NMF is a dimensionality reduction method that can summarize outstanding functional properties in a small number of metagenes (positive linear combination of genes). This is accomplished via a decomposition of the gene expression matrix into two matrices with nonnegative entries. Each sample is assigned to a subtype or cluster by finding the metagene that is most closely related to the sample's expression profile. The robustness of clustering is evaluated by repeating the factorization process using different random initial conditions for the factorization algorithm. This creates a consensus matrix to assess the stability of the resulting clusters [34]. The NMF appears to have some advantages over other methods: it is not based on distances, does not assume a hierarchical structure and provides a quantitative measure to identify the number of clusters. The latter is performed by means of the cophenetic coefficient.

Given the large number of genes and patients, virtually everything can be clustered. On the other hand, given the different nature of clustering algorithms, it is possible to modify the parameters to generate different results using the same data (e.g. the clustering produced by a given algorithm is dependent on the distance metric used). Therefore, the clustering is expected to be useful in the discovery of groups of patients with functional, survival, or phenotypic differences. The value of clustering is demonstrated by the biological information it provides, the utility of the markers found and the extrapolation of the results.

Techniques such as multidimensional scaling (MDS) can be useful for the visualization and initial recognition of high-dimensional data and to identify patterns and evaluate metrics used for the separation of elements. The method converts a similarity matrix to a simple geometrical picture [35]. The principal component analysis (PCA) can be used with exploratory intent and for the purpose of dimensionality reduction. In this method, new features, principal components with the larges variance, are identified and used instead of the original ones [36].

Classification and validation of results

Once the clustering is accomplished, the next phase in many studies consists of exploring the potential clinical utility and validation using

different cohorts; for these purposes, the supervised analysis techniques are used. The goal is to design a classifier that is able to accurately predict the class membership of new samples (test data) using data with known class membership (training data) [37]; samples used for training and testing should be large and independent for obtaining reliable results. Once a classification model is executed, it is important to estimate the classifier performance with respect to the sensitivity (true positives), specificity (true negatives), and accuracy (total number of correct predictions). Among the methods used for evaluating the performance of a classifier by splitting the initially labeled data into subsets are the cross-validation and bootstrap methods [26,38].

Two of the most common classification algorithms used for mining genomic data on cancer is the support vector machines (SVMs) and the classifiers based on nearest centroids, such as prediction analysis of microarrays (PAM). The SVMs map the input vector into a feature space of higher dimensionality and identify the hyperplane that separates the data points into two classes. The marginal distance between the decision hyperplane and the instances that are closest to the boundary are maximized. The resulting classifier achieves considerable generalizability for the classification of new samples [38]. The PAM calculates a standardized centroid for each class; the method takes the gene expression profile of a new sample and compares it to each of the class centroids. The class whose centroid it is closest to is the predicted class for that new sample. The method uses a shrinkage technique to assess the contributions of genes to classification as an automated gene selection step [39].

An important consideration for the classification process is the choice of the most discriminative genes for the analysis since this specific group of genes, which can be considered subtype- specific markers, is what makes the distinction between classes possible. The choice of a group of genes is also important to avoid over fitting. If increased error rates of the classification are observed despite the decrease in the error rates during the training process, over fitting may be the cause [38]; this is associated with the presence of a disproportionate number of genes with respect to the number of samples and can be prevented with the use of feature selection methods. Selected features can lead to better classification performance, provide insights into the disease and offer biomarkers with clinical value. These markers could be tested experimentally for evaluate their functional value in the disease and validated in different cohorts for corroborate their role as biomarker. To mention just one example: Armstrong et al. [40] found that lymphoblastic leukemias with MLL translocations can be separated from conventional acute lymphoblastic and acute myelogenous leukemias, they identified a target gene FLT3 that was shown experimentally to be a drug target [40,41].

Approaches to feature selection can be divided into two categories: filter methods and wrapper methods. In the former methods, a statistical measure of the marginal relevance of the features is used (e.g. t-test, SAM); those methods perform explicit feature selection before the classifier construction. Wrapper methods use the accuracy of a resulting classifier to evaluate the features, for example, classification techniques such as the decision trees and random forests [42,43] intrinsically contain a feature selection step that evaluate the "variable importance".

Methods such as significance analysis of microarrays (SAM) can be used as filter to find highly discriminative genes between subgroups. This method identifies differential gene expression relative to the spread of expression across all genes. Sample permutation is used to estimate false discovery rates (FDR) [44]. By adjusting the threshold, it is possible to control the estimated FDR associated with the gene sets. Decision tree is nonparametric and have the advantage of be easy to interpret, it follows a tree-structured where the nodes represent the input variables and the leaves correspond to decisions outcomes. A decision tree classifies by posing a series of questions or decision rules, each question is contained in a node, and every internal node points to one child node for each possible answer to its question [45]. Random forests uses an ensemble of unpruned decision tres (grown fully), each of which is built on a bootstrap sample of the training data using a randomly subset of variables [42] (Figure 2).

The identification of subtype-specific markers or cluster markers is sensitive to outliers. Therefore it is advisable to use only those samples that belong statistically to the core of each of the clusters before the extraction of biomarkers. This can be accomplished using the silhouette width [46]. Classifiers with fewer genes tend to perform poorly. Therefore, a compromise between the size and the performance of the classifier should be considered, for example, considering a sufficiently large number of genes that do not generate error rates higher than 5%. Once a good classifier is generated, it is necessary to reduce the number of genes for potential clinical application and validation.

Once the list of genes that differentiates groups has been obtained, these can be examined in search of functional information, such as signaling pathways involved or biological processes affected, through enrichment analysis. DAVID is a tool used for pathway enrichment and gene ontology analyses [47]. The gene sets enrichment analysis (GSEA) can be used to evaluate whether the differentially expressed genes are enriched in specific gene sets [48].

Despite research efforts in recent years, only a few molecular markers have been established in clinical practice. For example, although several studies have been reported in breast cancer, markers recommended by the American Society of Clinical Oncology (ASCO) are reduced to the status of a few molecules, ER and PR indicated for endocrine therapy, HER2 for anti-HER2 therapy, and the 21-gene recurrence score to determine prognosis. [49-51]. Other representative examples include KRAS mutations in colorectal cancer to select patients for treatment with antibodies against epidermal growth factor receptor (EGFR) [52], and EGFR and ALK alterations for therapeutic direction in lung cancer [53].

For a tumor marker to be used in clinical settings, issues related to analytic and clinical validity, and clinical utility must be addressed. Analytic validity relate to analytic accuracy, reliability, and reproducibility. Clinical validity is the demonstration that the marker has a strong association with a clinical outcome. Clinical utility entail that use of the marker has shown to result in a favorable balance between benefit and harm, leading to improved outcomes compared with nonuse of the marker [3]. Recently, Yuan et al. [54] addressed a key issue related to the lack of cancer biomarkers with clinical utility: statistical significance vs. magnitude difference. Predictive models in cancer have relied on statistical significance (P value) to evaluate clinical utility but the size of the difference in the patient outcome should also be considered [55]. For their applications in the clinical management of patients, the new molecular subclasses and predictors identified by machine learning approaches need to be refined and strongly validated. However, it is clear that they have the potential to complement traditional histopathological systems and to provide insight into the underlying molecular mechanisms.

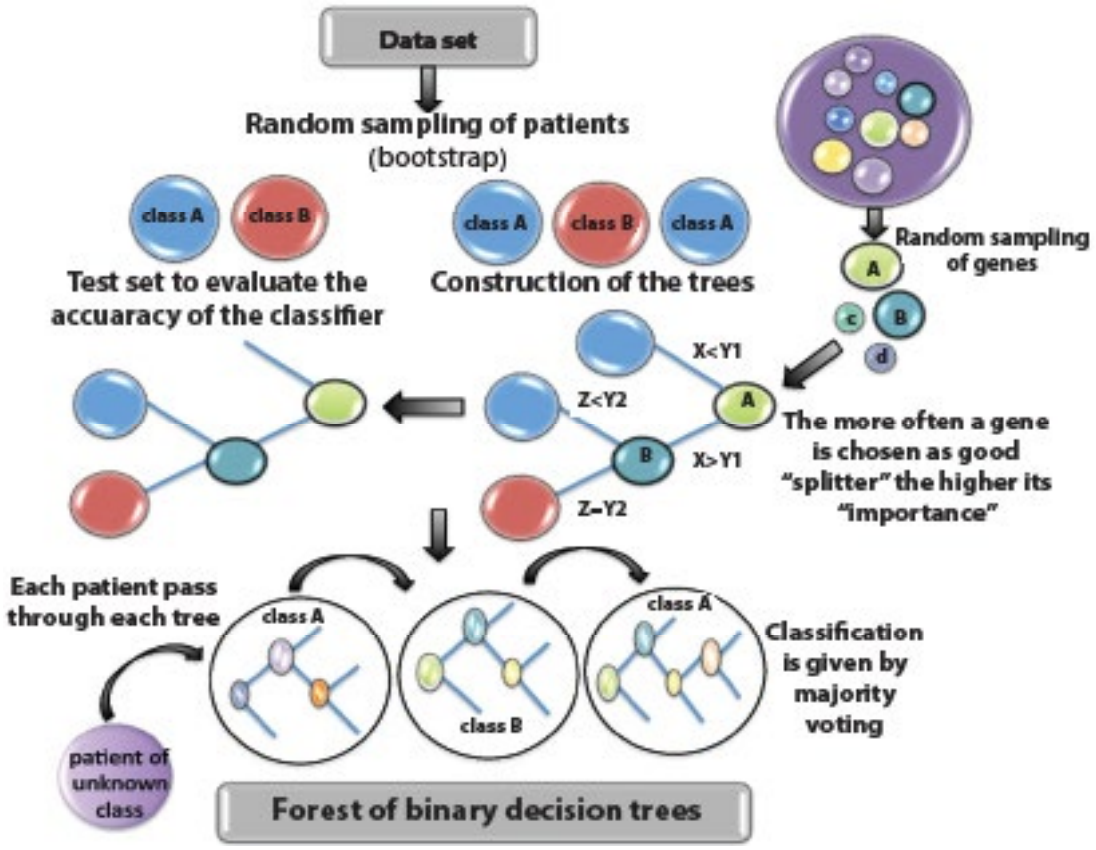

Figure 2: An illustration of Random Forest. A Random Forest is a collection of decision trees. Each tree cast a vote in classifying and it is formed by first selecting at random, at each node, a small group of input features (genes) to split on and, secondly, by calculating the best split based on these features in the training set. The tree is grown to maximum size (without pruning). This randomization scheme is blended with bootstrap to resample, the training data set each time a new individual tree is grown. Each patient not used in the construction of the tree is used as test set (out of bag), the tree is then tested against the "out of bag" patients to estimate the accuracy of the classifier. The entire process is repeated with new random data set divisions and new random gene sets for selection of splitter variables to produce ultimately a forest. The forest can then also be applied to independent patients of unknown class.

Datasets and Analysis Tools

Data generated by genomic cancer projects can be accessed freely or in a controlled manner. Allowing free access to genomic data has become a practice in international projects that generate high-throughput data. This democratization of scientific information has allowed such data to become the starting point of analysis, development of workflows and analytical tools and the generation of new questions from the scientific community [56,57].

Among the projects with genomic data on cancer are The Cancer Genome Atlas (TCGA) (http://cancergenome.nih.gov/dataportal) and The International Cancer Genome Consortium (ICGC) (http://dcc.icgc.org). TCGA is a collaborative effort put together to characterize the genomic changes in major cancers and is jointly conducted by The National Cancer Institute (NCI), The National Human Genome Research Institute (NHGRI) and the different centers and institutes of the National Institute of Health (NIH). This initiative has analyzed over 30 human types of cancer using high-throughput technologies: exome and genome sequencing, expression analysis, copy number variations, DNA methylation, and miRNAs, evaluated using microarray platforms and/or sequencing. The goal of the ICGC is a comprehensive description of genomic, transcriptomic, and epigenomic changes in 50 different tumor types and/or subtypes which are of clinical and societal importance across the globe. It is possible to find several levels of data processing in these projects. Based on the level of intervention and integration (raw, processed or normalized, interpreted and summarized), these are referred to as levels 1-4 and are freely available or controlled for different analytical platforms. Among the public repositories containing large numbers of expression data in compliance with basic rules for publication to the community, are the Gene Expression Omnibus (GEO) of the National Center for Biotechnology Information (NCBI) and the repository Array Express of the European Bioinformatics Institute (EBI).

For the analysis of high-throughput data, a wide range of methods are freely available through the Bioconductor project (http://www.bioconductor.org/). This resource has nearly 1104 software packages and an active user community. The Bioconductor uses the R programming language, which is open source and open development, allowing highly interactive protocols and providing an opportunity for programming one's own analysis. Another free software resource for use with the R environment and with a variety of solutions in statistical genomics is the Comprehensive R Archive Network (CRAN) with approximately 7590 packages. The Gene Pattern (http://www.broadinstitute.org/genepattern) stands out among the tools with a graphical interface, with hundreds of analysis tools and workflows for different types of genomic data.

Conclusion

Cancer subtyping schemes obtained by machine learning strategies and the use of clinically characterized cohorts are contributing to a better understanding of the molecular heterogeneity of cancer. This task is considerably more feasible today, as cancer datasets are freely available.

References

1. Kleppe M, Levine RL (2014) Tumor heterogeneity confounds and illuminates: assessing the implications. Nat Med 20: 342-344.
2. Almendro V, Marusyk A, Polyak K (2013) Cellular heterogeneity and molecular evolution in cancer Annu Rev Pathol 8: 277-302.
3. McShane LM, Hayes DF (2012) Publication of Tumor Marker Research Results: The Necessity for Complete and Transparent Reporting. J Clin Oncol 30: 4223-4232.
4. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, et al. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286: 531-537.
5. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, et al. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 403: 503-511.
6. Fu K, Weisenburger DD, Choi WW, Perry KD, Smith LM, et al (2008) Addition of rituximab to standard chemotherapy improves the survival of both the germinal center B-cell-like and non-germinal center B-cell-like subtypes of diffuse large B-cell lymphoma. J Clin Oncol 26: 4587-4594.
7. Sorlie T, Tibshirani R, Parker J, Hastie T, Marron JS, et al. (2003) Repeated observation of breast tumor subtypes in independent gene expression data sets. Proc Natl Acad Sci U S A 100: 8418-8423.
8. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, et al. (2000) Molecular portraits of human breast tumours. Nature 406: 747-752.
9. van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, et al (2002) Gene expression profiling predicts clinical outcome of breast cancer. Nature 415: 530-536.
10. van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, et al (2009) A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med 347: 1999-2009.
11. Tan IB, Tan P (2011) Genetics: an 18-gene signature (ColoPrint®) for colon cancer prognosis. Nat Rev Clin Oncol 8: 131-133.
12. Salazar R, Roepman P, Capella G, Moreno V, Simon I, et al (2011) Gene expression signature to improve prognosis prediction of stage II and III colorectal cancer. J Clin Oncol 29: 17-24.
13. Wang X, Markowetz F, De Sousa E Melo F, Medema JP, Vermeulen L (2013) Dissecting cancer heterogeneity--an unsupervised classification approach. Int J Biochem Cell Biol 45: 2574-2579.
14. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, et al (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4: 249-264.
15. Bolstad BM, Irizarry RA, Astrand M, Speed TP (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19: 185-193.
16. Smyth GK, Speed T (2003) Normalization of cDNA microarray data. Methods 31: 265-273.
17. Huber W, von Heydebreck A, Sueltmann H, Poustka A, Vingron M (2003) Parameter estimation for the calibration and variance stabilization of microarray data. Stat Appl Genet Mol Biol 2: 3.
18. Huber W, von Heydebreck A, Sültmann H, Poustka A, Vingron M (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics 1: S96-104.
19. Hackstadt AJ, Hess AM (2009) Filtering for increased power for microarray data analysis. BMC Bioinformatics 10: 11.
20. Leek JT, Storey JD (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet 3: 1724-1735.
21. Johnson WE, Li C, Rabinovic A (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 8: 118-127.
22. Benito M, Parker J, Du Q, Wu J, Xiang D, et al (2004) Adjustment of systematic microarray data biases. Bioinformatics 20: 105-114.
23. Drăghici S (2003) Data analysis tools for DNA microarrays. London: Chapman and Hall/CRC Press.
24. Drăghici S (2012) Statistics and data analysis for microarrays using R and Bioconductor. 2nd edition. Chapman and Hall/CRC Press
25. Kohonen T (1995) Self-organizing maps. Berlin: Springer.
26. Tarca AL, Careyaghici SVJ (2007) Chen Machine XW, Romelearningo R, Drand its applications to biology. PLoS Comput Biol 3: e116.
27. Monti S, Tamayo P, Mesirov J, Golub T (2003) Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. Machine Learning 52: 91–118.
28. Collisson EA, Sadanandam A, Olson P, Gibb WJ, Truitt M, et al (2011) Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy. Nat Med 17: 500-503.
29. Sadanandam A, Lyssiotis CA, Homicsko K, Collisson EA, Gibb WJ, et al (2013) A colorectal cancer classification system that associates cellular phenotype and responses to therapy. Nat Med 19: 619-625.
30. Cancer Genome Atlas Network (2011) Comprehensive molecular characterization of human colon and rectal cancer. Nature 487: 330-337.
31. Cancer Genome Atlas Research Network (2011) Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature 499: 43-49.
32. Cancer Genome Atlas Research Network (2011) Integrated genomic analyses of ovarian carcinoma. Nature 474: 609-615.
33. Yepes S, Torres MM, Andrade RE (2015) Clustering of Expression Data in Chronic Lymphocytic Leukemia Reveals New Molecular Subdivisions. PLoS One 10: e0137132.
34. Brunet JP, Tamayo P, Golub TR, Mesirov JP (2004) Metagenes and molecular pattern discovery using matrix factorization. Proc Natl Acad Sci USA 101: 4164-4169.
35. Chen Y, Meltzer PS (2005) Gene expression analysis via multidimensional scaling. Curr Protoc Bioinformatics Chapter 7: Unit 7.11.
36. Raychaudhuri S, Stuart JM, Altman RB (2000) Principal components analysis to summarize microarray experiments: application to sporulation time series. Pac Symp Biocomput 455-466.
37. Gentleman R, Carey V, Huber W, Irizarry R, Dudoit S, Editors (2005) Bioinformatics and Computational Biology Solutions Using R and Bioconductor. New York: Springer.
38. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI (2014) Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J 13: 8-17.
39. Tibshirani R, Hastie T, Narasimhan B, Chu G (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc Natl Acad Sci USA 99: 6567-6572.
40. Armstrong SA, Kung AL, Mabon ME, Silverman LB, Stam RW, et al (2003) Validation of a therapeutic target identified by gene expression based classification. Cancer Cell 3: 173-183.
41. Armstrong SA, Staunton JE, Silverman LB, Pieters R, den Boer ML et al (2002) MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. Nat Genet 30: 41-47.
42. Breiman L (2001) Random forests. Mach Learn 45: 5–32.
43. Breiman L, Friedman JH, Olsen RA, Stone CJ (1984) Classification and regression trees. New York: Wadsworth and Brooks.
44. Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci USA 98: 5116-5121.
45. Kingsford C, Salzberg SL (2008) What are decision trees? Nat Biotechnol 26: 1011-1013.
46. Rousseeuw PJ (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 20: 53-65.

47. Huang DW, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4: 44-57.

48. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA 102:15545-15550.

49. Harris L, Fritsche H, Mennel R, Norton L, Ravdin P, et al (2007) American Society of Clinical Oncology 2007 update of recommendations for the use of tumor markers in breast cancer. J Clin Oncol 25: 5287-5312.

50. Cardoso F, Saghatchian M, Thompson A, Rutgers E; TRANSBIG Consortium Steering Committee (2008) Inconsistent criteria used in American Society of Clinical Oncology 2007 update of recommendations for the use of tumor markers in breast cancer. J Clin Oncol 26: 2058-2059.

51. Duffy MJ, Walsh S, McDermott EW, Crown J (2015) Biomarkers in Breast Cancer: Where Are We and Where Are We Going? Adv Clin Chem 71: 1-23.

52. Allegra CJ, Jessup JM, Somerfield MR, Hamilton SR, Hammond EH, et al (2009) American Society of Clinical Oncology provisional clinical opinion: testing for KRAS gene mutations in patients with metastatic colorectal carcinomato predict response to anti - epidermal growth factor receptor monoclonal antibody therapy. J Clin Oncol 27: 2091-2096.

53. Leighl NB, Rekhtman N, Biermann WA, Huang J, Mino-Kenudson M, et al (2014) Molecular testing for selection of patients with lung cancer for epidermal growth factor receptor and anaplastic lymphoma kinase tyrosine kinase inhibitors: American Society of Clinical Oncology endorsement of the College of American Pathologists/International Association for the study of lung cancer/ association for molecular pathology guideline. J Clin Oncol 32: 3673-3679.

54. Yuan Y, Van Allen EM, Omberg L, Wagle N, Amin-Mansour A (2014) Assessing the clinical utility of cancer genomic and proteomic data across tumor types. Nat Biotechnol 32: 644-652

55. Henry NL, Hayes DF (2006) Uses and Abuses of Tumor Markers in the Diagnosis, Monitoring, and Treatment of Primary and Metastatic Breast Cancer. Oncologist 11: 541-552.

56. Joly Y, Dove ES, Knoppers BM, Bobrow M, Chalmers D (2012) Data sharing in the post-genomic world: the experience of the International Cancer Genome Consortium (ICGC) Data Access Compliance Office“ (DACO). PLoS Comput Biol 8: e1002549.

57. Walport M, Brest P (2011) Sharing research data to improve public health. Lancet 377: 537-539

58. Tan IB, Ivanova T, Lim KH, Ong CW, Deng N, et al. (2011) Intrinsic subtypes of gastric cancer, based on gene expression pattern, predict survival and respond differently to chemotherapy. Gastroenterology 141: 476-85

59. Budinska E, Popovici V, Tejpar S, D'Ario G, Lapique N, et al. (2013) Gene expression patterns unveil a new level of molecular heterogeneity in colorectal cancer. J Pathol 231: 63-76