Identification of Tumor Evolution Patterns by Means of Inductive Logic Programming

Vitoantonio Bevilacqua[1]*, Patrizia Chiarappa[2], Giuseppe Mastronardi[1], Filippo Menolascina[1,2], Angelo Paradiso[2], and Stefania Tommasi[2]

[1] *Polytechnic of Bari, 70125 Bari, Italy;* [2] *National Cancer Institute, 70126 Bari, Italy.*

In considering key events of genomic disorders in the development and progression of cancer, the correlation between genomic instability and carcinogenesis is currently under investigation. In this work, we propose an inductive logic programming approach to the problem of modeling evolution patterns for breast cancer. Using this approach, it is possible to extract fingerprints of stages of the disease that can be used in order to develop and deliver the most adequate therapies to patients. Furthermore, such a model can help physicians and biologists in the elucidation of molecular dynamics underlying the aberrations-waterfall model behind carcinogenesis. By showing results obtained on a real-world dataset, we try to give some hints about further approach to the knowledge-driven validations of such hypotheses.

Key words: array comparative genomic hybridization, breast cancer, cancer evolution model, gene selection, inductive logic programming

Introduction

The understanding of key genomic events that are considered as causal to cancer development and progression represents the holy grail of current research in oncology. Cancer remains a quite obscure disease at the molecular level even if great efforts have been spent in the last decades in order to defeat it. Several studies in this field have focused on the analysis of gene expression levels in several different contexts (*1*); however, a comprehensive outlook of the genetic mechanisms underlying the various types of cancer still lacks and a coherent model of evolution of cancer is out of our knowledge. This is a key point both from the research and clinical standpoints because such a dynamical model of cancer not only would be a precious aid for our understanding of cancer biology, but also may provide direct hints about the optimization of drug delivery in cancer therapies, which can result in a significant rising in disease-free survival of patients. A comprehensive introduction to tumor evolution investigation can be found in previous studies (*2–4*). The knowledge of molecular events behind cancer development and its key stages allows a more accurate classification of patients that does not rely on approximate evaluation of features like histological grading, but considers inner dynamics that are the real causal events of cancer. On the other hand, such a precious knowledge of this domain can help researchers in developing strategies both at the clinical (treatment planning systems) and pharmacological levels (compounds) in order to stop or reverse the aberrant process behind cancer.

In the last decades, researchers have proposed some models of cancer evolution, among which the Knudson's two-hit model (*5*) is probably the most famous. Knudson describes the deactivation of both alleles of a tumor suppressor gene as the initiating step of oncogenesis; however, no further information can be obtained from this model about following steps of cancer development. Current opinions in oncology tend to consider the subsequent progression toward aggressive malignancy as a multi-step process characterized by a lesser and lesser dependence of cells on growth signals as well as suppression of apoptotic pathways. This multi-step process is believed to be linear in some types of cancer, such as colorectal cancer, while to be strongly non-linear in other types like neuroblastoma or breast cancer (BC). Because of this strongly non-linear evolution, the investigation of BC is currently considered a very demanding task.

Due to the multiplicity of actors that cancer development requires, it could be argued that the ac-

***Corresponding author.**
E-mail: bevilacqua@poliba.it

tivity of several genes needs to be changed in order to develop any kind of cancer. This mutation process is inherently random and undirected; however, it is very unlikely that the necessary alterations could happen simultaneously by chance alone, particularly when more than a few genes need to be mutated. This has led Nowell to develop a model called "clonal evolution" (*6*); in Nowell's view, cancer is guided through evolution by a random mutation process that selects alterations providing a growth advantage to cancer cells. In this paper, we try to extract clonal evolution hallmarks in the context of BC using a data-mining approach based on association rule mining. In particular, we show how a particular approach to association rule mining can be used in the development of a coherent model of cancer evolution through the use of Tertius, an inductive logic programming (ILP)-based system proposed by Flach and Lachiche (*7*). ILP emphasizes the declarative aspect of knowledge representation, focusing on concepts more than on procedures (*8*, *9*). This aspect can result to be useful when complex environments (just like cancer-related ones) should be analyzed in order to extract knowledge on basic processes while a procedural approach is too hard when not unfeasible. For this reason, we have developed an evolution model for tumor progression in BC using Tertius. In this model, after collecting samples, DNA extraction is carried out, followed by hybridization and array scanning phases. A pre-processing stage is used in order to reduce the dimensionality of the dataset and to reduce computational time for analyses, that is, data normalization and filtering tasks are put at the base of this step. The Tertius algorithm is fed using pre-processed data and results are displayed in the form of a network of interactions.

Model

Efforts spent in the last years for the determination of tumor evolution pathways have been mainly focused on two approaches: graph building (*10*) and tree construction (*11*). Both paradigms belong to the class of graph theory based algorithms, but are characterized by significant differences. Graphs are defined as sets of nodes and edges connecting nodes. Trees are structures with a root node and branches that bring to nodes till the leaves are reached. Trees can be generalized as directed acyclic graphs, that is, graphs with directed edges and no cycles.

Graph building task can be faced in several different ways. In the last years, many different algorithms for tree and graph construction have been proposed. For example, SOTA algorithm has been used in order to build a tumor progression model for hepatocellular carcinoma (*12*). In another study, a branching tree and a distance-based tree were constructed for nasopharyngeal carcinoma (*11*).

All of these approaches are based on some kinds of metrics that establish relationships among cases; however, none of these approaches can consider adequately higher order relationships between variables (genes, BACs, and so on). The scientific community currently agrees in considering the linear relationships between genomic players' activity as a quite restrictive approximation, especially for human organism. For these reasons, in order to answer the need for more powerful tools, we propose an ILP approach to the problem of tumor evolution model building. ILP systems develop predicate descriptions from examples and background knowledge. The examples, background knowledge, and final descriptions are all described as logic programs. A unifying theory of ILP is being built up around lattice-based concepts such as refinement, least general generalization, inverse resolution, and most specific corrections.

ILP approach

In ILP, several entities can be defined; here we focus on objects (facts) and concepts (hypotheses). Having selected description languages for objects and concepts, a procedure is needed to establish if a given object belongs to a certain concept, that is, if the description of the object satisfies the description of the concept. If it is the case, then the concept description *covers* the object description. An example e for learning a concept C is a labeled fact, with a label O if the object is an instance of the concept C, and a label ϕ otherwise. The inductive logic learning task is then configured as:

Given a set ℓ of positive and negative examples of a concept C, find an hypothesis H, expressed in a given concept description language L such that:

(1) every positive example ε in $\ell+$ is covered by H;

(2) no negative example ε in $\ell-$ is covered by H.

Such procedure returns first-order logic rules that are used to maximize the coverage of the example set. Hypotheses drawn by these systems can be used to construct interaction models, frequently put in the

form of graphs, among variables. Understanding dynamics behind these hypotheses can greatly help in rising the definition of our knowledge about a precise process like a biological pathway in disease evolution.

Tertius algorithm

Tertius deals with learning first-order logic rules from the data that lack an explicit classification predicate. Learned rules are not restricted to predicate definitions as in supervised inductive logic programming. Tertius first performs an optimal search that tries to find the k most confirmed hypotheses belonging to the set H. The main contribution of Tertius in the field of ILP algorithms lays in its heuristic measure of confirmation (trade-off between *novelty*, defined as the relative decrease in counter-instances from expected to observed, and *satisfaction*, defined as the fraction of expected but non-observed counter-instances) for hypotheses and in the non-redundant refinement operator that avoids duplicates in the search. During its elaboration, Tertius tries to build first-order logic rules like the following:

$$\alpha =' (x_1 - y_1)'$$
$$\Rightarrow \beta =' (x_2 - y_2)' \text{ or } \gamma =' (x_3 - y_3)' \text{ or } \delta =' (x_4 - y_4)'$$

which puts in evidence the relationships existing between α, β, γ, and δ variables in the example set that the algorithm has been fed with. In this case, the rule, for example, states the situation when variable α in $(x_1 - y_1)$ has an impact on β, γ, and δ, specified by the ranges $(x_i - y_i)$. It is evident that this paradigm can be successfully translated in the bioinformatics field considering variables as biological actors and the ranges can represent gain/loss of copies or over/under expression of genes.

Results and Discussion

The Tertius algorithm has been run on the pre-processed data as described previously. In order to reduce the computational time needed for rule extraction, we fed the algorithm with a pre-filtered list of 40 genes. These genes were selected in order to split the case set in two subgroups based on tumor progression; this choice was driven by the fact that tumor progression is, evidently, a good indicator of hallmark events in tumor evolution model.

The Tertius analysis returned 21 associative rules (see Supporting Online Material) that were translated in Figure 1. This graph has been analyzed using the graph theory in order to individuate interesting actors in this pseudo-pattern. VEGFC (vascular endothelial growth factor C) and ATE1 (arginyltransferase 1) genes resulted to be *hubs* in this graph, that is, points

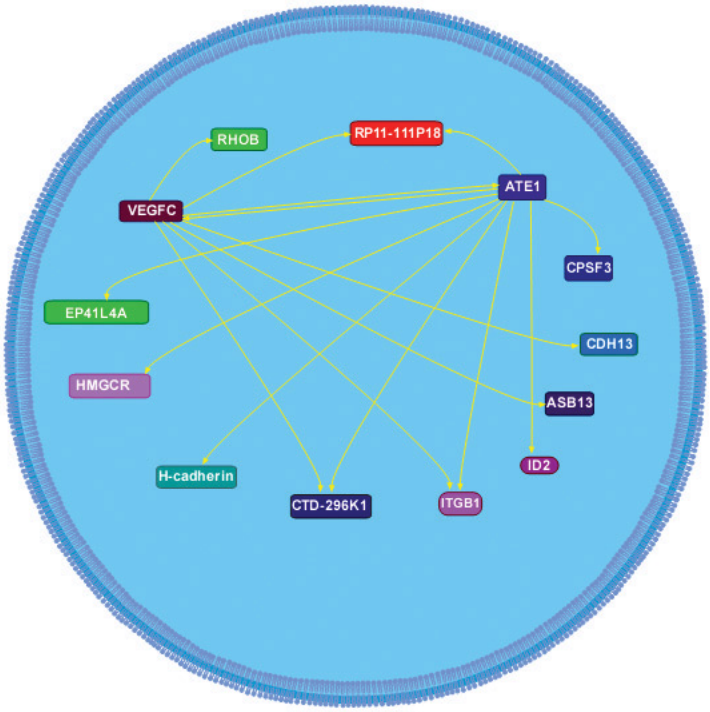

Fig. 1 Graph representation of the rules extracted by Tertius. Arrows represent control activity of one gene on another (causal relationships are represented by arrow orientation).

in which connections are concentrated (in formalisms we refer to in-degree and out-degree of nodes in order to define hubs). This is a quite interesting characteristic under the biological point of view, since both VEGFC and ATE1 are involved in vascularization bioprocesses.

VEGFC regulates the process of angiogenesis together with its two known receptor tyrosine kinases FLT1 and KDR/FLK1. Its role has been investigated by Hung *et al* (*13*); they tried to observe whether differential expression of VEGFC might explain the different propensity to lymph node metastasis in thyroid cancers. VEGFC's role in cancer is sustained by the fact that paired comparison of VEGFC expression between thyroid cancers and normal thyroid tissues from the same patients showed a significant increase of VEGFC expression in papillary thyroid cancer and a significant decrease of VEGFC expression in medullary thyroid cancer (*13*).

On the other hand, ATE1 functions as an oxygen sensor; in fact, through biochemical analyses, Kwon *et al* (*14*) demonstrated that the N-terminal cysteine, in contrast to N-terminal aspartate and glutamate, is oxidized before its arginylation by R-transferase, suggesting the sensor role of the arginylation branch of the N-end rule pathway. Recent works demonstrated that post-translational arginylation is critical for the survival of an organism, and the knockout of ATE1 results in embryonic lethality in mice. Moreover, it has been found that the lack of protein arginylation results in perturbation of angiogenesis, a process that is critical for tumor growth and metastasis development (*15*). Genomic analysis has identified a limited number of potential arginylation targets involved in the regulation of oncogenic transformation that have been hypothesized to play a role in prostate cancer (*16–18*).

As a matter of fact, these two genes result to be correlated in terms of function even if they haven't been found associated in the same context highlighted by this analysis. However, the biological snapshot returned by Tertius is clear: disruption in the activity of specific genes involved in angiogenesis is the turning point in BC progression. This aspect can be considered a relevant characteristic of BC for at least two reasons: Firstly, vascularization is a well-known mechanism through which tumors acquire energy, and this fact can explain the fast growing aggressiveness of some BC cases; Secondly, being abnormalities well localized from both a functional and a structural point of view, it could be argued that further studies should be carried out on approaches (such as drug therapies) aimed to prevent tumors from losing equilibrium in the mentioned hotspots.

In this paper, we presented an ILP-based approach to BC evolution modeling. Although the model has been demonstrated to be strongly non-linear, we tried to show how key steps in this process can be retrieved using first-order logic rules. In particular, the graph representation we used puts in evidence the complexity of events that underlie tumor progression, accounting for a small but interpretable non-linearity of the actual biological model. It can be argued that this aspect can result to be a relevant strength point of similar approaches: while not discarding the model, this algorithm provides tools to the researcher to draw a precise idea of the main actors of a certain pathway and how they interact in order to complete the process under investigation. Tertius algorithm, therefore, results to be a good trade-off between the expressive power needed for models in oncology and the necessary approximation that rises from the high intrinsic complexity of the processes underlying cancer. In particular, we argued that the roles of copy number levels of two genes, VEGFC and ATE1, can result to be critical in BC evolution. Aberrations in copy number levels in 4q34.1–34.3 and 10q26.13 chromosomal regions can result in disruption in the regulation of angiogenesis, which can rise the probability of vascularization of cancer tissues through novel vessels feeding cancer. Several theories explaining the roles of VEGFC and ATE1 singularly have been proposed in the last years for other types of cancer; however, to our knowledge, no interaction network has been illustrated as a BC evolution model. The advantages of such knowledge are quite evident: the complete list of steps characterizing BC can help researchers developing strategies to prevent the tumor following the known path till the degeneration of tissues. This can be translated in therapies enhancing self-repairing capabilities of DNA aimed at reducing the probability of specific-known next-to-come epigenetic events and, in a second step, at inducing self-repairing of damaged regions through drugs developed on purpose.

The potentialities of the present approach seem to be quite interesting. However, issues still remain in terms of the computational time needed by ILP algorithms. The computational complexity of these algorithms is still too high to think at feeding them with more than few tens of features. This aspect still limits their employment in the bioinformatics context that would probably obtain good benefits from their

use. Future directions for research could be found in the optimization of the approach used herein in order to keep computational costs low. Statistical evaluation and other methods described herein can be used in order to reduce the impact of the complexity issue. This remains an open question pushing the interest for further research in this fascinating field.

Materials and Methods

Specimens

In this study, we considered a cohort of 124 BC patients at different stages. Frozen tumor tissues were obtained from IRCCS "Giovanni Paolo II" of Bari. All specimens were collected under approved protocols from IRCCS with patient consent. The specimens' characteristics are provided in Table 1.

DNA extraction

Nucleic acids were extracted from tumor blocks as described in previous studies (*19*, *20*). Blocks were trimmed with a razor blade to remove normal tissues, and cryo-sections were obtained from either side of the block to ascertain that tumor cells comprised a significant part of the specimen. DNA was extracted using QUIamp tissue kits.

Array comparative genomic hybridization

In array comparative genomic hybridization (CGH), arrays of genomic BAC, P1, cosmid, or cDNA clones are used as the hybridization target in place of the metaphase chromosomes (*21–23*). The relative copy number is then measured at these specific loci by hybridization of fluorescently labeled test and reference DNAs as in conventional CGH (*19*). Since the clones used on the array contain sequence tags, their positions are accurately known relative to the genome sequence, and genes mapping within regions of copy number alteration can be readily identified using genome databases.

Imaging and analysis

Array CGH, imaging, and data acquisition were carried out using arrays of 2,464 genomic clones (BAC), each printed in triplicate (Hum Array 1.14 and Hum Array 2.0).

Data pre-processing

The output of a CGH array scanning has been converted in $\log_2$(R1/R2), where R1/R2 indicates the ratio of the two fluorescent tags; this is a common pre-processing of the data that tries to overcome the

Table 1 Summary of statistics for the series of data used in this study*

Property		All (n=124)	ER positive	ER negative
Age	Young ($\leq$ 45 years)	56	33	23
	Old ($\geq$ 70 years)	66	57	9
T status	T1	31	24	7
	T2	59	39	20
	T3	8	8	0
	T4	20	16	4
Differentiation	G1	15	13	2
	G2	57	45	12
	G3	35	18	17
	Missing	15		
PgR status	PgR positive	58	37	21
	PgR negative	65	53	12
Proliferation	MIB negative	18	17	1
	MIB positive	105	73	32

*The case set has been divided using common directions in the clinical field. Statistical properties of the discrimination are shown. ER, estrogen receptor; PgR, progestogen receptor; T1–4, breast cancer stage according to TNM classification; G1–3, histological grading according to TNM classification; MIB, the monoclonal antibody developed against the Ki-67 proliferation antigen.

bias introduced by the fact that lost and normal BACs are theoretically compressed in the interval [0, 1], and, on the other hand, amplifications can vary in the range [1, ∞). At this point, some missing values exist in the dataset (also indicated as NaN, that is, Not a Number); a decision about these values and the BACs they belong are needed. Some approaches for missing value handling tend to simply eliminate those features that contain missing values; this, obviously, inevitably leads to some loss of information. Another kind of approach consists in imputing missing values using other information; the most simple method imputes a missing value using the mean (or median) of the distribution of the single BAC to all the missing values; it is evident that if a single case out of all contains a value lost for all of the others, these methods will impute this single value to all of the cases leading to a strong bias in data. If the cases are two and each of the two belongs to one of the classes under investigation, it is clear that the mean imputation, in this case, will make powerful gene selection criteria like Wilcoxon test or student's t-test to be absolutely inadequate. For these reasons, we chose a hybrid approach to missing value imputation: we firstly removed all the BACs that were present in 33% of the cases. Then we used the collateral missing value estimation algorithm as described in Sehgal *et al* (*24*). As the final step, we applied a gene entropy filter (*23*) to the dataset and obtained a matrix of 124 by 2,218. This set of genes was used as input for the gene selection algorithm.

Gene selection

The feature selection stage is one of the most delicate steps in the whole microarray experimental pipeline. Many different approaches are documented in literature; one of the most recent contributions to this field of optimal feature set finding comes from Marghny and El-Semman (*25*). Other feasible approaches include sensitivity analysis by removing attributes, proportion correct use in rules, ratio of features between-category to within-category sums of squares, signal-to-noise scores in one-versus-rest or one-versus-all fashion, Kruskall-Wallis nonparametric test (ANOVA) and number of appearances in models (*1*, *26*, *27*). However, the scientific community seems to agree that the "optimal feature set" simply does not exist but, instead, it should be measured on the single classification approach and, in general, on the single experiment (*28*). For this reason, we developed a consensus scheme for attribute selection that takes advantage of three well established statistical methods, namely the student's t-test (Lilliefors test for normality of samples, $p < 0.01$), receiver operating characteristic, and entropy (Kullback-Liebler divergence). All of these techniques can be used to compile a ranking of the features that accounts for the power of a single attribute to discriminate between the output classes. All of the 2,464 BAC values for each of the 124 cases were processed and the outcome being T-stage (1–2 vs 3–4); using these algorithms, three rankings have been obtained. A new global ranking has been compiled using the three positions of each clone as an indicator of its discriminating power. This strategy has been employed in order to overcome the limitations of the single methods and to gain a deeper insight into the data structure and information distribution. In addition, as reported in Li *et al* (*29*), it should be considered that using a single viewpoint for relevance estimation can result in unbearable bias in results. Bonferroni adjustment has been employed to correct the statistics for multiple comparisons. The first 40 clones were selected for the following analysis stages.

Authors' contributions

PC and ST collected the datasets. PC, FM and ST conducted data analyses and prepared the manuscript. GM and AP assisted with manuscript preparation. VB and FM conceived the idea of using this approach. VB supervised the project and cowrote the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors have declared that no competing interests exist.

References

1. Golub, T.R., *et al.* 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286: 531-537.
2. Klein, G. 2002. Introduction: genetic and epigenetic contributions to tumor evolution. *Semin. Cancer Biol.* 12: 327-330.
3. Hill, R., *et al.* 2005. Heterogeneous tumor evolution initiated by loss of pRb function in a preclinical prostate cancer model. *Cancer Res.* 65: 10243-10254.

4. Bellomo, N. and Preziosi, L. 2000. Modelling and mathematical problems related to tumor evolution and its interaction with the immune system. *Math. Comput. Model.* 32: 413-452.
5. Knudson, A.G. Jr. 1971. Mutation and cancer: statistical study of retinoblastoma. *Proc. Natl. Acad. Sci. USA* 68: 820-823.
6. Nowell, P.C. 1976. The clonal evolution of tumor cell population. *Science* 194: 23-28.
7. Flach, P.A. and Lachiche, N. 2001. Confirmation-guided discovery of first-order rules with Tertius. *Mach. Learn.* 42: 61-95.
8. Muggleton, S. 1991. Inductive logic programming. *New Generat. Comput.* 8: 295-318.
9. Muggleton, S. and de Raedt, L. 1994. Inductive logic programming: theory and methods. *J. Logic Program.* 19-20: 629-679.
10. Farazi, P.A., *et al.* 2003. Differential impact of telomere dysfunction on initiation and progression of hepatocellular carcinoma. *Cancer Res.* 63: 5021-5027.
11. Wu, L.S.H. 2006. Construction of evolutionary tree models for nasopharyngeal carcinoma using comparative genomic hybridization data. *Cancer Genet. Cytogenet.* 168: 105-108.
12. Poon, T.C., *et al.* 2006. A tumor progression model for hepatocellular carcinoma: bioinformatic analysis of genomic data. *Gastroenterology* 131: 1262-1270.
13. Hung, C.J., *et al.* 2003. Expression of vascular endothelial growth factor-C in benign and malignant thyroid tumors. *J. Clin. Endocrinol. Metab.* 88: 3694-3699.
14. Kwon, Y.T., *et al.* 2002. An essential role of N-terminal arginylation in cardiovascular development. *Science* 297: 96-99.
15. Karakozova, M., *et al.* 2006. Arginylation of beta-actin regulates actin cytoskeleton and cell motility. *Science* 313: 192-196.
16. Kashina, A.S. 2006. Differential arginylation of actin isoforms: the mystery of the actin N-terminus. *Trends Cell Biol.* 16: 610-615.
17. Rai, R., *et al.* 2006. Molecular dissection of arginyltransferases guided by similarity to bacteria peptidoglycan synthases. *EMBO Rep.* 7: 800-805.
18. Rai, R. and Kashina A. 2005. Identification of mammalian arginyltransferases that modify a specific subset of protein substrates. *Proc. Natl. Acad. Sci. USA* 102: 10123-10128.
19. Pinkel, D., *et al.* 1998. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.* 20: 207-211.
20. Albertson, D.G., *et al.* 2000. Quantitative mapping of amplicon structure by array CGH identifies CYP24 as a candidate oncogene. *Nat. Genet.* 25: 144-146.
21. Solinas-Toldo, S., *et al.* 1997. Matrix-based comparative genomichybridization: biochips toscreen for genomic imbalances. *Genes Chromosomes Cancer* 20: 399-407.
22. Pollack, J.R., *et al.* 1999. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat. Genet.* 23: 41-46.
23. Snijders, A.M., *et al.* 2001. Assembly of microarrays for genome-wide measurement of DNA copy number. *Nat. Genet.* 29: 263-264.
24. Sehgal, M.S., *et al.* 2005. Collateral missing value imputation: a new robust missing value estimation algorithm for microarray data. *Bioinformatics* 21: 2417-2423.
25. Marghny, M.H. and El-Semman, I.E. 2005. Extracting logical classification rules with gene expression programming: microarray case study. In *Proceedings of the International Conference on Artificial Intelligence and Machine Learning (AIML 05)*, pp.11-16. Cairo, Egypt.
26. Gopalakrishnan, V., *et al.* 2006. Rule learning for disease-specific biomarker discovery from clinical proteomic mass spectra. *Lect. Notes Comput. Sci.* 3916: 93-105.
27. Statnikov, A., *et al.* 2004. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics* 21: 631-643.
28. Ein-Dor, L., *et al.* 2006. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics* 21: 171-178.
29. Li, J., *et al.* 2003. Discovery of significant rules for classifying cancer diagnosis data. *Bioinformatics* 19: ii93-102.

Supporting Online Material

http://oncologico.bari.it/laboratorio/bioinformaticsnews/supplementary_material_tertius_rules_acgh124.PDF