Institute of Biostatistics and Analyses
Masaryk University

Proceedings of the 8th Summer School on Computational Biology

From Analysis of Genomic Data to Clinical Applications – Case Studies

12–15 September 2012
Mikulov, Czech Republic

Editors:
Eva Budinská
Vlad Popovici

european social fund in the czech republic
EUROPEAN UNION
MINISTRY OF EDUCATION, YOUTH AND SPORTS
OP Education for Competitivness
UNIVERSITAS MASARYKIANA BRUNENSIS
INVESTMENTS IN EDUCATION DEVELOPMENT

Proceedings of the 8th Summer School on Computational Biology
From Analysis of Genomic Data to Clinical Applications – Case Studies

Editors: Eva Budinská, Vlad Popovici
Cover: Radim Šustr

Published by AKADEMICKÉ NAKLADATELSTVÍ CERM, s.r.o. Brno
Purkyňova 95a, 612 00 Brno
www.cerm.cz

Printed by FINAL TISK s.r.o. Olomučany
1st edition, 2012
100 prints

ISBN 978-80-7204-804-5

Contents

Foreword 5
Eva Budinská, Vlad Popovici

LECTURES

Recent progress in cancer research and targeted anticancer therapy 9
Stjepan Uldrijan

Gene expression-based classifiers 13
Vlad Popovici

Searching for robust and clinically relevant subtypes in higt density molecular data 31
Eva Budinská

Meta-analysis – taking the information further 41
Hana Imrichová, Eva Budinská

Tree of Life in a gappy genomic era 49
Natália Martínková

Multilocus phylogeny of Sciurini tree squirrels 55
Patrícia Pečnerová, Natália Martínková

Reconstructing phylogeny from patchy data of rodents 64
Jiří Moravec, Natália Martínková

COMPUTATIONAL BIOLOGY STUDENTS' ABSTRACTS

Genetic association studies 73
Lucie Brožová

Specification and monitoring of oscillatory properties in dynamical systems 77
Petr Dluhoš

Study of expression of genes specific for BRAF mutated colon tumours in early phases of tumour development.....80
Barbora Hanáková, supervisor: Mgr. Eva Budinská, Ph.D.

Spatial modelling of vegetation based on bioclimatic data.....84
Lenka Krupková, supervisor: Mgr. Klára Komprdová, Ph.D.

Modelling acidification of forest soils with the inclusion of uncertainties.....87
Petra Malcová

Statistical models for ecological assessment of reservoir using phytobenthos.....91
Lucie Panáčková

Regression diagnostic tools in survival analysis.....94
Ivana Svobodová

Foreword

The rapid pace of technological developments life sciences witnessed the last ten-fifteen years changed completely the research paradigm. Especially in genetics, we no longer analyze only a few genes or characteristics of few biological samples, but we are embarking on whole genome profiling of hundreds of specimens, in the hope of deciphering how life and species evolved, or how to treat cancer, for example. While in the beginning mathematics and informatics were simply tools used mostly for confirming or rejecting some biology-driven hypothesis, in the new standard of a novel research field of *Computational Biology* they play a central role, on equal foot with biology. In this highly interdisciplinary environment, researchers with various backgrounds have to find a common language to collaborate effectively.

This is the context in which the Faculty of Science of Masaryk University introduced in its curriculum the Computational Biology study programme, endorsed and supported by the Institute of Biostatistics and Analyses (IBA) of the Masaryk University. In line with its continuous efforts to keep abreast of latest trends in the field, IBA organizes a series of summer schools in Computational Biology, to encourage free and open exchanges and collaborations between professors, young scientists as well as students in computational biology and related domains. The informal format of these schools replaces the classical ex cathedra lectures with free discussions to which students are particularly welcome to contribute, their active participation constituting a substantial part of the summer school´s programme.

The 8th edition of the summer school in Computational Biology continues the tradition of inviting as lectures confirmed researchers specialized in various aspects of biological data analysis and interpretation. This year's programme places at its core the practical aspects of making discoveries from genomic data. Its vision is to provide an overview of theoretical aspects and examples (case studies) of typical applications which have the potential of changing our current understanding of biology, be it seen from a clinical or evolutionary perspective. The programme includes lectures and practical lessons presenting aspects of biomarker discovery, large-scale meta-analyses and evolutionary biology. We hope this specific field of application of computational biology will bring some new viewpoints and experience for all participants.

We are gratefully acknowledging the financial support of the Ministry of Education, Youth and Sports of the Czech Republic; project

CZ.1.07/2.2.00/07.0318, Interdisciplinary Development of Computational Biology Study Programme, where this summer school is organized.

On behalf of the programme and organizing committee,

Brno, August 19, 2012

Eva Budinská
Vlad Popovici

From Analysis of Genomic Data to Clinical Applications — Case Studies

Lectures

Recent progress in cancer research and targeted anticancer therapy

Stjepan Uldrijan[1,2]

[1] *Department of Biology, Faculty of Medicine, Masaryk University Brno; e-mail: uldrijan@med.muni.cz*
[2] *International Clinical Research Center, St. Anne´s University Hospital in Brno*

Abstract

Current treatment options for many cancers are still limited to standard chemotherapy or radiotherapy that are associated with serious side effects and often do not result in satisfactory patient responses. There is clearly a need for new, more efficient targeted anti-cancer therapies designed on the basis of our better understanding of biological processes taking place inside cancer cells on the molecular level. This lecture summarizes the principles involved in malignant transformation of cells and provides examples of a successful translation of new findings in the field of cancer biology into more efficient anticancer therapies. At the same time, the lecture suggests new areas of cancer research that could translate into cures for cancers that do not respond to currently available therapies.

Key words

Cancer biology, mutations, hallmarks of cancer, targeted therapy.

1. Introduction

Current treatment options for the majority of human cancer are limited to standard chemotherapy and radiotherapy and usually show variable efficacy depending on the tumor type and stage. Since DNA-damaging drugs and radiation cause a wide range of negative side effects in patients, scientists and medical professionals involved in cancer research are trying to find new, more efficient and at the same time less damaging molecularly targeted therapeutic approaches for as many tumor types as possible. Some of the most successful recently introduced anti-cancer therapies have been developed through deeper understanding at the molecular level of biological processes taking place in tumor cells. However, nearly all tissues in human body can give rise to cancer, leading to a large variety of cancer types with different behavior, including differences in the speed of growth, the ability of metastasize and in responses to treatment, which makes the identification of a universal cure for cancer virtually impossible. The quest for this universal cure is complicated even more by the fact that even tumors of the same type and origin can differ significantly in their genetic makeup. That is why current cancer research favors the idea of personalized approach to cancer therapy – finding new drugs targeting genetic defects in cancer cells found in a specific subtype of cancer. For this purpose, high-throughput whole-genome genetic analyses of large numbers of tumor samples of a particular cancer type will help us to fully comprehend the role of specific genetic changes in the development of individual types of cancer. While this is a long-term aim of current cancer research, some underlying principles common for all human cancers that govern the transformation of normal human cells into malignant cancers have already been identified and this knowledge has been used to rationally target the growth of tumor cells. These basic principles and several examples of their practical use in targeted cancer therapy will be discussed in this lecture.

2. Hallmarks of cancer

Hannahan and Weinberg have defined the hallmarks of cancer as acquired functional capabilities that allow cancer cells to survive, proliferate, and disseminate (Hanahan and Weinberg 2000). They include sustaining proliferative signaling, evading growth suppressors, resisting cell death, enabling replicative immortality, inducing angiogenesis, and activating invasion and metastasis. More recently, additional hallmarks have been added to the list: reprogramming of energy metabolism and evading destruction by immune cells (Hanahan and Weinberg 2011). All these hallmarks are acquired in different tumor types via distinct mechanisms and at various times during the course of multistep tumorigenesis. Their acquisition is made possible especially by the development of genomic instability in cancer cells, which generates random mutations, ranging from point mutations to large chromosomal rearrangements.

3. Examples illustrating the progress in targeted anticancer therapy

3.1. CML – Tyrosin kinase inhibitors

The majority of cases of human chronic myelogenous leukemia (CML) carry reciprocal translocation between chromosomes 9 and 22, which carry the *abl* and *bcr* genes, respectively. This specific genetic change leads to the formation of fused, hybrid genes encoding hybrid Bcr-Abl proteins with strong and deregulated growth-promoting activity. A small molecule inhibitor, imatinib mesylate (Gleevec), has been developed to inhibit the tyrosine kinase activity of the Bcr-Abl fusion protein in CML (Weinberg 2007). Imatinib and similar new drugs have allowed the majority of CML patients to survive a disease that was incurable only a decade ago.

3.2. Malignant melanoma – BRAF inhibitors

Malignant melanoma is responsible for about 80 % of all deaths from skin cancers (Miller and Mihm 2006). Although the large majority of people diagnosed with early melanoma are cured after surgical excision of the primary tumor, advanced metastatic disease is usually refractory to all current forms of systemic therapy and has a very poor prognosis. The 5-year survival rate is estimated at only 6%, with a median survival time of 6 months (Singh et al. 2008, Dahl and Guldberg 2007). Among cytotoxic agents, dacarbazine and its analogue temozolomide remain the chemotherapy of choice but they produce objective response in only 15 to 20% of patients and the median duration of the response is only 4 months.

Currently the most intensely tested targeted therapeutic strategy for malignant melanoma is the inhibition of the RAS/RAF/MEK/ERK mitogen-activated protein kinase (MAPK) pathway. This signal transduction pathway, which normally regulates cell proliferation and survival in response to extracellular stimuli, has been found constitutively activated in the majority of human melanomas, most commonly by activating mutations of *N-Ras* and *B-RAF* proto-oncogenes that are found in approx. 80% of malignant melanomas (Fecher et al. 2008). Sorafenib (BAY 43-9006), an oral inhibitor of BRAF kinase has demonstrated activity against melanoma in preclinical tests and it is currently the most promising drug in clinical trials for the targeted treatment of malignant melanoma, together with another BRAF inhibitor PLX4032.

3.3. Downregulated p53 pathway – Mdm2 inhibitors

Tumor suppressor p53 has a key role in cellular responses to various stress stimuli, including DNA damage, telomere erosion, ribosomal stress, hypoxia, and oncogene activation

(Vousden and Lane 2007). In the normal cellular environment, p53 protein levels are kept low, mainly by interactions with its major negative regulator Mdm2 that serves as an E3 ubiquitin ligase for p53 and targets p53 for degradation via 26S proteasome. In response to stress stimuli, p53 is stabilized and activates the expression of its target genes, leading to responses such as cell cycle arrest, senescence, or apoptosis. The growth suppressive function of p53 is commonly lost in human tumors, in about 50% of cancers by mutations leading to the loss of the ability of p53 to bind DNA and transactivate its target genes. In cancer cells that retain wild type p53 gene, the p53 pathway function is often abrogated by the overexpression of inhibitory proteins, such as protein E6 of human papillomavirus (HPV) (e.g. in cervical carcinomas), ubiquitin ligase Mdm2 (e.g. in sarcomas), or a related protein MdmX (e.g. in breast, colorectal or lung cancers) (Toledo and Wahl 2006). The stress response pathway regulated by p53 has gained considerable attention over the years and some small molecule drugs designed to modulate the activity of this pathway have already reached clinical testing, most notably the Mdm2 inhibitor Nutlin-3 (Brown et al. 2007).

3.4. Breast cancer - The concept of synthetic lethality

Synthetic lethality was first described in genetic studies on Drosophila, in which the loss of a certain combination of two different genes lead to a lethal phenotype, while loss of each of the genes separately had no effect on the viability of the flies. However, the majority of synthetically lethal combinations have been identified using genetic manipulations of yeast (Scherens and Goffeau 2004, Ooi et al. 2006, Nijman 2011). In comparison to yeast, the identification of synthetically lethal gene combinations in multicellular organisms including mammals is much more complicated, because of a much larger genome, higher copy number or increased number of variants of a certain gene and a certain level functional redundancy among related genes. Despite this, the concept of synthetic lethality has already entered clinical testing as a therapeutic approach for certain human cancers. Breast or ovarian carcinomas often exhibit defects in homologous DNA recombination caused by mutations in genes coding for BRCA1/2 proteins and this makes them extremely sensitive to PARP inhibitors that block the alternative DNA repair pathway - the base excision repair (BER). Combinations of PARP inhibitors with DNA damaging chemotherapy are currently tested in clinical trials in breast and ovarian cancers with very promising results (Farmer et al. 2005, Boss et al. 2010).

4. References

Boss DS, Beijnen JH, Schellens JH. 2010. Inducing synthetic lethality using PARP inhibitors. Current Clinical Pharmacology 5: 192-5.

Brown CJ, Lain S, Verma CS, Fersht AR, Lane DP. 2009. Awakening guardian angels: drugging the p53 pathway. Nature Reviews Cancer 12: 862-73.

Dahl C, Guldberg P. 2007. The genome and epigenome of malignant melanoma. APMIS 115: 1161-1176.

Farmer H, McCabe N, Lord CJ, Tutt AN, Johnson DA, Richardson TB, Santarosa M, Dillon KJ, Hickson I, Knights C, Martin NM, Jackson SP, Smith GC, Ashworth A. 2005. Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. Nature 434: 917-21.

Fecher LA, Amaravadi RK, Flaherty KT. 2008. The MAPK pathway in melanoma. Current Opinion in Oncology 20: 183-9.

Hanahan D., Weinberg RA. 2000. The hallmarks of cancer. Cell 100: 57-70.

Hanahan D., Weinberg RA. 2011. Hallmarks of cancer: the next generation. Cell 144: 646-74.

Miller AJ, Mihm MC Jr. 2006. Melanoma. New England Journal of Medicine 351: 51-65.

Nijman SM. 2011. Synthetic lethality: general principles, utility and detection using genetic screens in human cells. FEBS Letters 585: 1-6.

Ooi SL, Pan X, Peyser BD, Ye P, Meluh PB, Yuan DS, Irizarry RA, Bader JS, Spencer FA, Boeke JD. 2006. Global synthetic-lethality analysis and yeast functional profiling. Trends in Genetics 22: 56-63.

Scherens B, Goffeau A. 2004. The uses of genome-wide yeast mutant collections. Genome Biology 5: 229.

Singh M, Lin J, Hocker TL, Tsao H. 2008. Genetics of melanoma tumorigenesis. British Journal of Dermatology 158: 15-21.

Toledo F, Wahl GM. 2006. Regulating the p53 pathway: in vitro hypotheses, in vivo veritas. Nature Reviews Cancer 6: 909-23.

Vousden KH, Lane DP. 2007. p53 in health and disease. Nature Reviews Molecular and Cellular Biology 8: 275-83.

Weinberg RA. The Biology of Cancer. New York: Garland Science 2007. 796 p. ISBN 0-8153-4076-1.

This work was supported by The European Regional Development Fund - Project FNUSA-ICRC (CZ.1.05/1.1.00/02.0123).

Gene expression-based classifiers

Vlad Popovici[1,2]

[1] *Bioinformatics Core Facility, Swiss Institute of Bioinformatics, CH-1015 Lausanne, Switzerland; e-mail: vlad.popovici@isb-sib.ch*
[2] *Institute of Biostatistics and Analyses, Masaryk University, CZ-62500 Brno, Czech Republic*

Abstract

Whole genome profiling and decreasing costs of genome sequencing enable measuring the activity of tens of thousands of genes which can potentially be used for making predictions about patients' risk of relapse or response to a specific treatment. These predictions are based on mathematical models that combine the measurements from a selected set of genes into either a continuous score or a binary outcome. In order to build such models that can be used in clinical practice with real benefits for the patients, a rigorous methodological approach must be followed and the purpose of this chapter is to briefly describe some theoretical considerations and practical results in the field of gene expression-based classifiers.

Key words

Classifiers, biomarkers, performance estimation, model validation.

1. Introduction

The clinical practice has shown that many cancer treatments benefit only a small group of patients who received them. Lacking precise means of identifying the patients most likely to respond to a given treatment results in many patients being prescribed ineffective treatments, which puts a serious burden on them and on the health care systems. The *personalized medicine* addresses exactly this problem by trying to diagnose and treat a disease using information about patient's genes, proteins and environment. At the core of the diagnostic and treatment decisions are placed the *classifiers*, which are mathematical models assembling all the information into a system producing binary or multi-valued decisions. In this context, the new treatments are accompanied by diagnostic tests which are supposed to identify the most likely responders.

The problem of companion diagnostic tests is even more important in the case of *targeted therapies*, which are "drugs or other substances that block the growth and spread of cancer by interfering with specific molecules involved in tumor growth and progression" (NCI's Facts Sheet, http://www.cancer.gov/). As these drugs target specific molecular processes, such as cell growth signaling, angiogenesis, apoptosis, or stimulate the immune response, highly specific tests are needed to identify the right patient population.

1.1. What is a classifier in the context of genomic data

There are various names under which the classifiers appear in the literature related to gene expression-based diagnostics and prognostics. They may be called "(multigene) expression signature" or "(multigene) biomarkers" or simply "risk predictors/scores". In general, we talk about a classifier when we have in mind a model which produces a crisp decision (be it

binary or multi-level). While a score can be converted into a decision (see further on in this chapter), and so "score" and "classifier" terms could be used interchangeably in some context, a gene expression signature is usually not enough to specify a classifier. A gene expression signature refers more to the genes selected to be specific to some phenotype, but it normally does not specify the way these genes should be combined to predict the phenotype in question. Also the term "biomarker" could be misleading, since it may also refer to some markers that can be mechanistically linked to a disease activity. In conclusion, we prefer the term "gene-based classifier" by which we mean a prediction model which combines the gene expressions (and maybe other variables) in a model. This classifier may have a score as an intermediate step towards decision.

2. Classifiers

Without any loss of generality, we will consider in the following the case of *binary classifiers*, constructed on continuous variables and we will denote the two alternatives (called *classes*) by "-1" and "+1". Let $f: \mathbb{R}^p \to \mathbb{R}$ be a real-valued function which will map a vector $\mathrm{x} \in \mathbb{R}^p$ to a continuous *score*. A score $s = f(\mathrm{x})$ is converted into a *binary class label* by $h(s) = \mathrm{sign}(s)$. Some classifiers will directly produce the binary label (e.g. the basic Top Scoring Pairs algorithm, described later in this chapter), while others will firstly produce a score. As the labels are easily obtained from the scores and since using continuous functions is more convenient for modeling, we will generally focus on finding the function f rather than h. We can then state the problem of learning a classification rule to be the task of finding a real-valued function $f \in \mathcal{F}$ that maps each point of the input space (here considered to be $\mathbb{R}^p$) to a score that, after thresholding, will produce a binary label which will not differ in too many cases from the true label. This formulation is too vague to be of any practical utility as long as we do not specify

- which is the function space $\mathcal{F}$ in which we search for the solution;
- what we mean by 'differ', and
- how many misclassified cases is 'too many' for a classifier to be considered good.

The choice of the function space $\mathcal{F}$ is the first decision a data modeler has to take and, in most cases, it means defining a parametric form for the score function f. Let the parameters on which f depends be denoted by a r-dimensional parameter vector $\omega \in \Omega \subseteq \mathbb{R}^r$. The problem of *training* a classifier becomes an optimization problem in which one has to find the optimal vector ω^* such that the *expected risk of misclassification (expected prediction error)* is minimized:

$$\omega^* = \arg\max_{\omega} \int L\big(y, f(\mathrm{x})\big)\, dP(\mathrm{x}, y) \ ,$$

where L is a *loss function* penalizing the discrepancies between the predicted label $f(\mathrm{x})$ and the true label y. The integral is taken with respect to the probability density function P which is generating the population $\{(\mathrm{x}, y) | \mathrm{x} \in \mathbb{R}^p,\ y = \pm 1\}$. As the probability function is usually not available, the risk is estimated from a finite *training set (sample)* given as a pair of sets $X_n^t = \{\mathrm{x}_i, i = 1, \dots, n\} \subset \mathbb{R}^p$ and $Y_n^t = \{y_i = \pm 1,\ i = 1, \dots, n\}$ of points draw independently and identically distributed from the underlying probability. In this case, the prediction error can only be estimated from the finite sample, thus the estimation will become dependent on the particular training set. This observation justifies the introduction of various error estimation techniques, some briefly described in this chapter.

The loss function is of central importance in defining the form of the classifier and several ways of penalizing the errors have been proposed in the literature (Hastie et al, 2009; Duda et al, 2001). Here we will consider only the case of *squared error loss*,

$$L\big(y, f(\mathrm{x})\big) = \frac{1}{2}\big(y - f(\mathrm{x})\big)^2$$

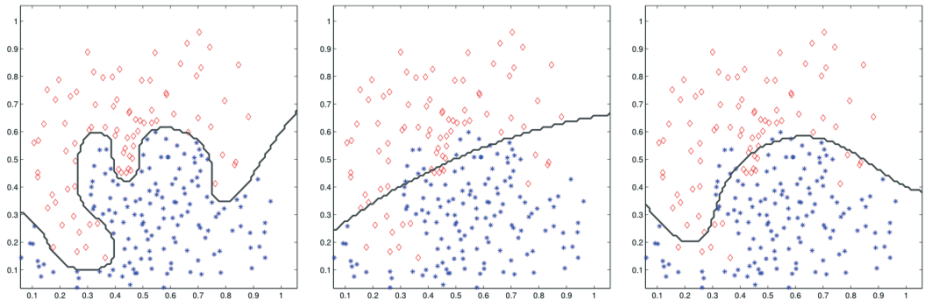
which is, by far, the most commonly used. In the case of a risk of misclassification estimated from a finite sample, we talk about *empirical risk (of misclassification)* and we estimate it by its mean value over the given sample:

$$\frac{1}{n}\sum_{i=1}^{n} L(y_i, f(\mathrm{x}_i)) \quad ,$$

which for squared error loss is simply the usual mean squared error, $\frac{1}{n}\sum_i \big(y_i - f(\mathrm{x}_i)\big)^2$.

Figure 1 depicts a possible scenario for a binary classification problem in the case of $p = 2$, with the solid line representing the *classification boundary* (i.e. the separation between the two classes), defined by the equation $f(\mathrm{x}) = 0$. Let us analyze the three proposed solutions: in the first panel, the classifier perfectly separates the two classes, while the other two solutions are simpler (smoother) classification functions, which misclassify some points. In practice, it turns out that a function that perfectly separates the training set will usually perform poorly on unseen data, i.e. the prediction will have high *variance* on different samples drawn from the same underlying distribution *P* as the training data. We say in this case that the first function *overfits* the training set. On the other hand, a too simplistic explanation as the one given by the second classifier will never be able to satisfactory fit the training data, i.e. the model chosen has a large *bias*. In this case we say that the model *underfits* the training set. The central problem of machine learning is to find that right tradeoff between underfitting and overfitting, that will generate functions f able to generalize well, i.e. their performance remains good on unseen data. We will further detail what we mean by good performance of a classifier. This problem is also known as bias-variance dilemma in classical statistics.

Figure 1. Three possible scenarios for a trained classifier: different degrees of *regularization* lead to different solutions, with various performances.



To conclude this introductory section, we note that the genomic applications of classifiers face a specific problem of fitting models in very high dimensional spaces (in general, $p \gg n$). Because of the high number of degrees of freedom of the learning problem, one can always find a classifier that perfectly fits the training set, for any possible labeling. Or, to put it in other terms, higher is the dimensionality of the space slower is the convergence of the estimators (of the parameters) – phenomenon called *curse of dimensionality*. It follows that

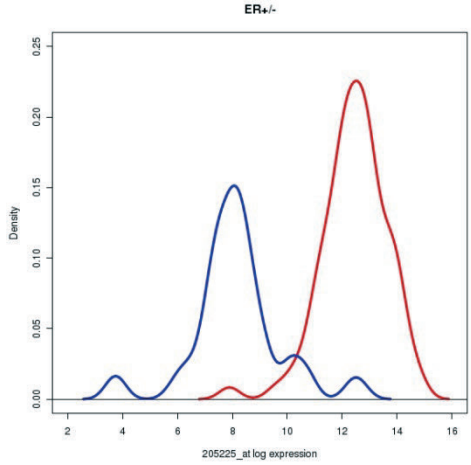
one has either to use a very large learning set or to constrain the form of the classifier such that it fits only the most salient characteristics of the two classes. The first solution is not practically possible, so only the second approach remains feasible. Luckily, the variables (genes) are not all independent and a large proportion of them are usually not important for the classification task. This explains why, despite of the unfavorable settings, for many applications one can find proper classifiers with reasonable performance.

2.1. Bayesian decision theory

Let us consider for a moment the best case scenario in which the classification problem is completely specified by the probability functions:

- $P(y = +1)$ and $P(y = -1)$ are called *prior probabilities (priors)* and give the probability of either of the classes, when no other information is available. In general, for a binary classification problem, one excludes the possibility of observing any other class but one of the two ($y \in \{\pm 1\}$), so $P(y = 1) = 1 - P(y = -1)$.
- *class-conditional density functions*, $p(\mathrm{x}|y = 1)$ and $p(\mathrm{x}|y = -1)$, for $\mathrm{x} \in \mathbb{R}^p$, the probability density function of x, given that its label is "+1" (or "-1").

Figure 2. Class conditional density functions for ESR1 gene expression as measured by one probeset: $p(205225\_\mathrm{at}\,|y = "ER+")$ and $p(205225\_\mathrm{at}\,|y = "ER-")$. The two classes are estrogen-positive (ER+, red line) and estrogen-negative (ER-, blue line).



Using Bayes' rule, it is easy to obtain the *posterior* probability

$$P(y = \pm 1|\mathrm{x}) = \frac{p(\mathrm{x}|y = \pm 1)\, P(y = \pm 1)}{p(\mathrm{x})}.$$

This shows that by observing the vector x (called evidence) and using information about priors and class conditional densities – called *likelihood* – one can obtain the posterior probability that the observed instance belongs to one of the classes. It follows naturally that, for minimizing the risk of misclassification one must assign x to the class with maximum posteriori probability (*Bayes decision rule*):

$$f(\mathrm{x}) = \begin{cases} -1, & P(y=-1|\mathrm{x}) > P(y=+1|\mathrm{x}) \\ +1, & P(y=-1|\mathrm{x}) \leq P(y=+1|\mathrm{x}) \end{cases}$$

It is sometimes convenient to consider this rule in terms of *log-ratio*: assign x to class "+1" if

$$\log \frac{P(y=+1|\mathrm{x})}{P(y=-1|\mathrm{x})} \geq 0,$$

and to class "-1" otherwise.

The Bayesian decision is optimal in the sense that it minimizes the probability of error, but it requires full information about priors and class-conditional densities to be available. However, this is not the case in real applications, and a plethora of approaches have been proposed to deal with more realistic scenarios. One can try, for example, to consider a parametric model for the probabilities (e.g. linear discriminant analysis, naïve Bayes classifier, etc. etc.) or to use nonparametric estimators of the densities.

2.2. Linear discriminants

Suppose that the class-conditional densities are multivariate Gaussians,

$$p(\mathrm{x}|y=\pm 1) = \frac{1}{(2\pi)^{p/2}\left|\Sigma_{\pm 1}\right|^{1/2}} e^{-\frac{1}{2}(\mathrm{x}-\mu_{\pm 1})^T \Sigma_{\pm 1}^{-1}(\mathrm{x}-\mu_{\pm 1})}$$

where $\mu_{\pm 1}$ are the mean vectors and $\Sigma_{\pm 1}$ are the covariance matrices of the two respective classes ($|\cdot|$ is the determinant operator). If the two classes have equal covariance matrices, $\Sigma_{-1} = \Sigma_{+1} = \Sigma$, the log-ratio of the posteriors becomes

$$\log \frac{P(y=+1|\mathrm{x})}{P(y=-1|\mathrm{x})} = \log \frac{P(y=+1)}{P(y=-1)} - \frac{1}{2}(\mu_{-1}+\mu_{+1})^T \Sigma^{-1}(\mu_{+1}-\mu_{-1}) + \mathrm{x}^T \Sigma^{-1}(\mu_{+1}-\mu_{-1}).$$

The values for the class means and the covariance matrix have to be estimated from the training set, using the usual estimators. The priors are estimated by the class frequencies, $\hat{P}(y=+1) = n_{+1}/n$, and $\hat{P}(y=-1) = n_{-1}/n$, where $n_{+/-1}$ are the number of elements in each class. The above equation shows that the decision boundary between classes is a linear function of x (for equal covariance matrices). By introducing the discriminant functions

$$\delta_{\pm 1}(\mathrm{x}) = \mathrm{x}^T \Sigma^{-1} \mu_{\pm 1} - \frac{1}{2}\mu_{\pm 1}^T \Sigma \mu_{\pm 1} + \log P(y=\pm 1)$$

the decision rule $y = \arg\max_{k=\pm 1} \delta_k(\mathrm{x})$ is equivalent to comparing the log-ratios of the posteriors with 0. As a final remark, we note that $d(\mathrm{x},\mu) = (\mathrm{x}-\mu)^T \Sigma (\mathrm{x}-\mu)$ is called *Mahalanobis distance* from x to μ (which becomes Euclidean distance if the covariance matrix is the unit matrix) and that, under equal covariances assumption, the decision rule assigns x to the class whose centroid (μ) is the closest in the sense of this metric.

The squared error loss function, mentioned in the introduction of this chapter, is intimately linked to LDA classifier as this can be derived from a linear regression model, where we fit a linear model to the label variable, considered this time a continuous variable.

2.3. Nearest neighbor and related classifiers

The intuition behind the nearest neighbor and related methods is that an observation should be assigned to the class containing other similar observations. The nearest neighbor classifiers employ a *voting scheme* for deciding the class membership of a sample $\mathrm{x} \in \mathbb{R}^p$. The predicted (estimated) label is

$$\hat{y} = \text{sign}\left(\sum_{x_i \in N_k(x)} y_i\right)$$

where $N_k(\text{x})$ is a neighborhood of k closest points to x. In other words, the predicted label $\hat{y}$ is the most common label among the k points in the neighborhood. If $\hat{y} = 0$ it means that the point lies on the decision boundary and it has equal number of points from each class in its neighborhood.

The notion of neighborhood implies the existence of a metric which, at its turn, is closely related to the notion of *similarity*, in the sense that more similar observations are closer to each other than observations less similar. In the case of $\mathbb{R}^p$, the natural metric is the Euclidean distance, but this is not necessarily the best choice. For genomic applications, in which the observations may be corrupted by high levels of noise, one may consider alternative distances, for example

- *correlation distance*: $d_{corr}(\text{x,z}) = 1 - \rho(\text{x,z})$
- *cosine distance*: $d_{cos}(\text{x,z}) = 1 - \frac{\langle \text{x,z} \rangle}{\|\text{x}\|\|\text{z}\|}$, where $\langle \cdot,\cdot \rangle$ denotes the scalar product of two vectors, and $\|\cdot\|$ the L_2 norm, respectively.

The parameter $k \geq 1$ has to be optimized for each problem, usually by cross-validation (see later on in this chapter). Smaller values will lead to a better fit of the training set, but may have an adverse effect on the generalization properties of the classifier. Also, there is a direct link between the parameter k and the smoothness of the decision boundary.

Instead of considering all the points in the data set in the decision rule, one may choose to select only a few "representative" patterns from each class and to compute the distances only to these points in order to classify a new observation. This is similar to LDA decision where, as we have seen above, one computes the Mahalanobis distance to the centers of the two classes and uses this information to classify the new observation. However, many other strategies of choosing the "centers" of the classes (like averaging all class members, or taking their median, for example) and distances to these centers can be employed, each leading to a slightly modified version of the algorithm. This class of *nearest centroid* classifiers is commonly employed in genomic applications because, despite not being necessarily the best in term of performance, it generally leads to simple classification rules that are readily interpretable and have reasonable performance.

2.4. Top scoring pairs

Top scoring pairs (TSPs) (Geman et al, 2004) are simple two-genes binary classifiers, in which the prediction of the class label is based solely on the relative ranking of the expression levels of the two genes. The rank--based approach to classification ensures a higher degree of robustness to technical variations and makes the rule easily portable across platforms. Also, the direct comparison of the expression level of the genes is easily interpretable in the clinical context, making the TSPs attractive for medical tests.

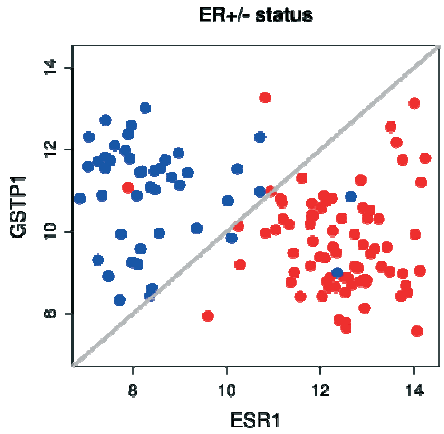
Let again $\text{x} = [x_i]_{i=1,\ldots,p} \in \mathbb{R}^p$ be a vector of measurements (e.g. gene expression) representing a sample and let the corresponding class label be $y = \pm 1$. Then, for all pairs of variables *i* and *j*, a score is computed,

$$s_{i,j} = P\big(x_i < x_j \big| y = 1\big) - P\big(x_i < x_j \big| y = -1\big), \;\; 1 \leq i, j \leq p$$

where P are conditional probabilities and the corresponding decision rule is: if $x_i < x_j$ then predict $y = 1$, otherwise $y = -1$. The pair with the highest score or the top k pairs are then considered for the final model (Geman et al, 2004; Tan et al, 2005).

Remarkably, this method does not require the optimization of any parameter and does not depend on any threshold. Figure 3 shows an example of a TSP predicting the estrogen receptor status. The decision boundary (in grey) is always a line with a slope of 1.

Figure 3. Predicting estrogen receptor status: if GSTP1 < ESR1, then the sample is considered ER+ (red dots), otherwise ER- (blue dots).



3. Performance parameters and performance estimation

In the context of clinical applications, a classifier is seen as a test and its continuous value $f(x)$ is called *score*. This score is discretized into two (binary tests) or more categories by using a number of thresholds (or cut-offs) and a prediction about the patient is made based on the predicted category. For example, a test can be used to predict if a patient has a given disease (binary test), or to which of a number of risk groups he/she belongs (e.g. low, medium and high risk groups – categorical tests). By convention, we will say that an individual which is predicted to have the disease to be positive for the test.

Several categories of medical tests are more common:

- *diagnostic* tests are designed to detect the 'diseased' condition in a patient;
- *prognostics* tests try to predict an outcome of interest, like 'recurrence' vs. 'no-recurrence';
- *predictive* tests are used to detect which patients may/may not respond to a treatment; and
- *screening* tests are usually applied to a large population of normally healthy individuals in which the disease has low prevalence, and are usually followed by other confirmatory tests.

Each of these tests is designed to work in specific settings. For example, we require a screening test to detect all (or, say 99%) of all diseased cases (must be sensitive), even if it will produce a relatively high rate of false alarms (false positives). In contrast, a diagnostic

test must be sensitive and with low false positive rates. On the other hand, as we have seen from the Bayes decision theory, the prior distributions influence the final decision. A screening test is used in a population where the positive cases have a low prevalence: for example, in 2009 breast cancer in UK had an age-standardized incidence of 124.4 cases in 100 000 women, so we can set a prior $P(y = disease) = 0.125$. A diagnostic test which is applied to confirm a screening test will work on a population with a much higher incidence of the disease, so a possible prior would be $P(y = disease) = 0.75$. The screening and diagnostic tests could be the same, with the only difference being the value of the threshold for the score, above which we call a patient diseased. And this threshold is optimized based on the prior probabilities.

In the following, we will briefly review some of the performance parameters that are used for characterizing the classifiers. For a comprehensive treatment of the subject in the context of clinical applications, see (Pepe, 2003).

3.1. Threshold-dependent performance parameters

We will use the following convention for calling the classes:

- *true label* (disease status) is denoted by D:

$$D = \begin{cases} -1, & \text{if non-diseased} \\ 1, & \text{if diseased} \end{cases}$$

- *predicted label* is denoted by Y:

$$Y = \begin{cases} -1, & \text{if negative for the test} \\ 1, & \text{if positive for the test} \end{cases}$$

A continuous score $f(\mathrm{x})$ is converted into a prediction by $\mathrm{sign}(f(\mathrm{x}) - \theta)$, where θ is a threshold.

For a given observation x one of the following 4 situation may arise:

- $D = 1, Y = 1$: *true positive* – the prediction and the true label are both indicating a diseased case
- $D = -1, Y = -1$: *true negative* – the prediction and the true label are both indicating a non-diseased (healthy) case
- $D = 1, Y = -1$: *false negative* – the test fails to detect the disease status
- $D = -1, Y = 1$: *false positive* – the test predicts as diseased a healthy case

In assessing the performance of a classifier/test one is interested in estimating the probabilities of each of the above four events to occur. The estimation is done based on the respective frequencies in a test set. One usually constructs a *confusion matrix* containing counts of the observed occurrences (Table 1), from which the probabilities are estimated.

Table 1. The confusion matrix and the associated probabilities.

Predicted labels	True labels (gold standard) $D=-1$	$D=1$	Marginal probabilities
$Y=-1$	True negatives $P(Y=-1\|D=-1)$	False negatives $P(Y=-1\|D=1)$	$P(Y=-1)$
$Y=1$	False positives $P(Y=1\|D=-1)$	True positives $P(Y=1\|D=1)$	$P(Y=1)$
Marginal probabilities (priors)	$P(D=-1)$	$P(D=1)$ (prevalence)	

The following performance parameters are some of the most commonly used criteria for judging a diagnostic test:

- *disease-centric* measures the performance predicting the disease: The *true positive/negative fractions (TPF, FPF)*:

  $$\text{TPF} = P(Y=1|D=1), \text{FPF} = P(Y=1|D=-1)$$

  They are both needed to characterize the test and they are dependent on the chosen threshold. If one knows the disease prevalence ($P(D=1)$), then the probability of error can be estimated by

  $$P(Y \neq D) = P(D=1)(1-\text{TPF}) + \big(1-P(D=1)\big)\text{FPF}$$

  A perfect test will have $\text{TPF}=1$ and $\text{FPF}=0$. The TPF is also called *sensitivity*, while $1-\text{FPF}$ is called *specificity*. In the clinical testing literature, the two latter terms are more common than the first ones.

- *predicted values* are used to quantify the clinical value of a test (the likelihood of disease when the test is positive): The *positive/negative predicted values (PPV, NPV)* are defined as

  $$\text{PPV} = P(D=1|D=1), \text{NPV} = P(D=-1|Y=-1)$$

  A perfect test will have $\text{PPV}=\text{NPV}=1$, while one totally uninformative, $\text{PPV} = P(D=1)$ and $\text{NPV} = P(D=-1) = 1-P(D=1)$.

There is a simple connection between the two groups of measures and its derivation is left as an exercise to the reader.

Since the estimators for the above measures are random variables from a Bernoulli trial, one can compute *confidence intervals (CI)*, using any of the proposed methods (e.g. normal approximation, Wilson score, Agresti-Coull, and others (Newcombe, 1998)). Whatever method is used, the confidence intervals (usually 95% CIs) must be reported for a full characterization of the test.

As a final remark, we note that the Cis obtained based on binomial distribution refer to each of the measures individually and do not provide a *confidence region* for the joint distribution of the pairs (TPF,FPF) or (PPV,NPV). To obtain such confidence region, one can use the following result:

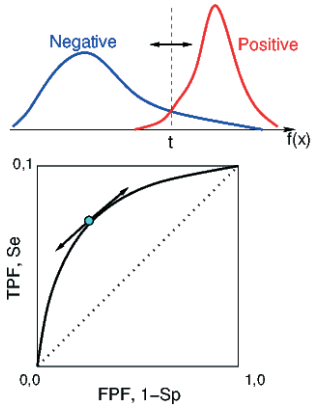
Proposition. If (P_{low}, P_{up}) and $Q_{low}, Q_{up})$ are $1-\alpha^*$ univariate confidence intervals for two binomial random variables P and Q, then the rectangle $\left(P_{low}, P_{up}\right) \times (Q_{low}, Q_{up})$ is a $(1-\alpha)$ confidence region for (P, Q), where $\alpha = 1-(1-\alpha^*)^2$.

For example, from two 95% univariate confidence intervals, one can construct a 90.25% confidence region for the joint variable.

3.2. Threshold-independent performance parameters

We have already noted that the performance measures described in the previous section depend on the chosen value of the threshold θ, and therefore we call them point estimates. However, these tests (classifiers) may need to work in different contexts, where one may want to select a different operating regimen (trade-off between sensitivity and specificity, or PPV and NPV). Moreover, when comparing two tests with different operating regimens, it is difficult to draw any conclusion. It is clear that we need a characterization of the test which is independent of the threshold. The *receiver operating characteristic (ROC) curve* serves exactly this purpose.

Figure 4. Varying the threshold *t* above which a score $f(\mathrm{x})$ leads to a positive test, generates a ROC curve in the (FPF, TPF) space.



By letting the TPF and FPF varying with the threshold,

$$\mathrm{TPF}(\theta) = P(f(\mathrm{x}) \geq \theta | D = -1)$$

$$\mathrm{FPF}(\theta) = P(f(\mathrm{x}) \geq \theta | D = 1)$$

we obtain the definition of the ROC curve:

$$ROC = \left\{\left(\mathrm{FPF}(\theta), \mathrm{TPF}(\theta)\right) \;\middle|\; \forall \theta \in \mathbb{R}\right\}$$

It is easy to see that the ROC function is monotone increasing and that it is invariant to strictly increasing transformation of the scores. The parametric form of the curve is given by

$$ROC = \left\{\left(\alpha, \mathrm{TPF}\left(\mathrm{FPF}^{-1}(\alpha)\right)\right) \mid \forall \alpha \in (0,1)\right\}$$

A summary of the ROC curve is obtained by taking the area under the curve (AUC):

$$AUC = \int_0^1 \mathrm{ROC}(\theta)\, d\theta$$

AUC is lower-bounded by 0.5 (corresponding to a totally uninformative test) and upper-bounded by 1. It can also be seen as the Mann-Whitney-Wilcoxon U-statistic: $\mathrm{AUC} = P(Y_{D=1} > Y_{D=-1})$, i.e. the probability of correctly ordering a random pair of cases.

3.3. Performance estimation

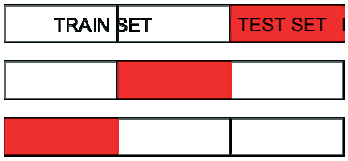
Once a classifier is trained, one has to estimate its performance on unseen data. Lacking access to the full data collection on which the classifier will be applied, one will have to rely on statistical estimates of the performance. The easiest estimate would be the one obtained by applying the classifier on the same data used for training it (plug-in estimate). Except for a few rare cases, this estimate will be optimistically biased, i.e. will underestimate the error rate. Furthermore, relying on the plug-in estimate will more often than not lead to overfitting the training set, i.e. one will find the parameters such that the classifier will have minimum error rate on the training set, but it will perform poorly on new data. The morale is that the estimation of performance has to be done on an independent data set, completely different than the one used for building the classifier.

One possible option would be to randomly split the data available into two disjoint subsets, one used for building the model and one for estimating its performance (*split sample validation* or *holdout validation)*. While appealing, this method has at least two drawbacks: it does not use the available data in an optimal way and the training set is reduced drastically in comparison with the original sample size. However, the true validation of a classifier, diagnostic test remains its long run application on unseen data.

In order to better use the training data, several *resampling methods* have been proposed, among which: the k-fold cross-validation, Monte Carlo cross-validation, leave one out cross-validation, bootstrapping, etc. They all have in common the idea of repeatedly randomly partitioning the available into a training set and a validation set. The training set thus obtained is used for full model construction (including feature selection, meta-parameter optimization, model selection, etc.), while the validation set is used for obtaining intermediate estimates of the performance. At the end of the procedure, the intermediate estimates are aggregated into a final value (usually be averaging, but more sophisticated estimates can be used – see for example the .632 estimator below) and a measure of variability of the estimate is also computed (variance, standard error, confidence intervals). These methods differ in the strategy they use for partitioning the data. We will briefly describe some of them here, while for others the reader is referred to (Duda et al, 2001; Hastie et al, 2009).

- *k-fold cross validation* splits the data into k partitions and uses each of them in turn as validation set. Typical values for k are 5 and 10 and the choice represents usually a trade-off between a reasonable training set size and the computational burden, as the procedure is repeated k times. Note that any two models built in this setting share k-2 folds as training data. This means that the predictions are not totally independent so the variance of the estimates is usually underestimated. An improved performance estimation is obtained by repeating the k-fold cross validation on randomly shuffled versions of the original set (the so called *repeated k-fold cross validation*). The final estimate of the performance (e.g. error rate, sensitivity, specificity, etc.) is the average of the intermediate estimates.

Figure 5. A 3-fold cross-validation scheme: each of the folds is used once and only once as validation set (the red block).



- *leave-one-out cross-validation* is an extreme case of k-fold cross-validation for the case k=1. The training/testing steps are repeated *n* times, where *n* is the sample size.
- *Monte Carlo cross-validation*: repeatedly splits randomly the data set into a training and validation set. For example, it retains 2/3 of the data in training and 1/3 for validation. The procedure is, in fact, a sequence of split-sample validations applied on random permutations of the data. Because of the random split of the data, the procedure does not ensure that all points are used for training and validation.
- *bootstrapping*, in contrast with the above methods, resamples *with replacement* from the original data set, generating new training sets (bootstraps) of the same size *n*. It means that the new training sets may contain duplicated training examples, while other samples are not included. On average, the bootstraps contain 0.632 of the original set. The procedure is repeated *B* times.

  The .632 estimator of the error rate (or other performance measure) is given by

$$\hat{E}_{.632} = 0.368\,\hat{E}_0 + 0.632\frac{1}{B}\sum_{b=1}^{B}\hat{E}_b$$

  where $\hat{E}_0$ is the plug-in error rate on the full training set and $\hat{E}_b$ are the error rate obtained at repetition *b* by applying the classifier on the left out data. The empirical distribution of $\hat{E}_b$ can be used for estimating the confidence intervals (for example, the 0.025 and 0.975 quantiles of this distribution are good estimates for the lower and upper limits of the 95% confidence interval).

All these resampling procedures for performance estimation can be implemented to preserve the proportions of the classes from the original data set. In this case, they are called *stratified* since, indeed, the sampling takes place within strata (levels) of the class label variable.

4. Guidelines for gene-based classifier development

Developing gene-based classifiers poses several specific problems, in addition to the "classical" issues that arise when building predictive models. Some of the specific issues are methodological, while others relate to the utility and relevance of the classifiers built.

4.1.Methodological issues

During the last decade thousands of new gene-based classifiers have been published, covering a large palette of applications. The US Food and Drug Administration, which is responsible for approving new diagnostic tests for medical applications, set up a series of projects to investigate the reproducibility and reliability of decision models built on gene expression data. These projects, gathered under the acronym of MAQC (MicroArray Quality Control) have shown that the technology is mature enough to be used in clinical practice.

The second phase (MAQC-II) dealt specifically with classification models (Shi et al, 2010) and put forward a number of recommendations, some of which are mentioned below.

As mentioned before, the data points lies in a high dimensional space where the number of dimensions greatly outnumbers the data set cardinality ($n \ll p$). This makes the problem to be ill-posed, in the sense that, theoretically at least, there may be an infinite number of solutions to a classification problem. This is why building a classifier requires a proper feature (variable) selection before training the model per se, but the methods for performing feature selection are not discussed here. We only mention that the feature selection can be done either independently or jointly with training the classifier (may be embed into the process of classifier training, as in the case of penalized logistic regression, for example), but in any case it is a mandatory step.

In the early phases of development of a new classifier, one usually tries many different algorithms before narrowing the selection to a few of them. The initial set of classifiers to be tried should be rich enough such that a suitable model can be found. MAQC-II has shown that in most cases the simpler methods perform as well as the more sophisticated ones on gene expression data. In this project, more than 30,000 models have been assessed and the conclusion was that the major factor impacting the performance of the models is the problem difficulty and not the complexity of the algorithms thrown at it (Shi at al, 2010). Once the initial exploratory phase completed, the list of candidate models should be short (2-3 models). These candidates should be evaluated on new data and the final model selected. The final model can then be trained and its performance estimated (either by resampling methods or on other independent data). This approach requires a fairly large amount of data but will likely produce a robust model that will not be overfitted to the training data.

The performance estimation is usually the most prone to methodological errors task in building a classifier. In theory, *all the steps performed from raw data to final model* must be included in the cross-validation (or whatever resampling method) loop. However, this is not always feasible: for example, microarray data normalization usually inspects the whole batch of raw samples for producing the normalized data. This means that any input vector for the classifier will be influenced by the information from other vectors in the data set, so the data normalization step has to be performed inside the cross-validation. On the other hand, the normalization step can be quite computationally demanding and repeating it at each iteration will slow down the process of model assessment. While this issue is not well studied in the literature, the common consensus is that the performance estimates are marginally impacted by the inclusion or not of the data normalization step in the cross-validation. That is why the overwhelming majority of studies leave the normalization outside the cross-validation. Nevertheless, apart from the normalization step, all the other processing steps must be included in the cross-validation (ideally, even the model selection step – if a model is selected at the end). Failing to obey this rule will lead to an optimistically biased estimation of performance (Varma and Simon, 2006). By far, the incorrect performance estimation for prediction models is the most common error (Dupuy et al, 2007).

Finally, a question that still lacks a definite answer refers to the samples size needed for developing a gene-expression classifier. Sample size estimation can be done under some parametric assumptions: for example (Dobbin and Simon, 2005) assume a normal multivariate distribution of the classes and derive mathematical formulae for computing the sample size for classifier development and validation. Assumption-free approaches exist and relies on simulations: (Popovici et al, 2010) shows how using the learning curves can be used to estimate if increasing the sample size would bring any benefit for classifier training and what would be the required sample size to achieve a predefined performance.

4.2. Clinical utility and relevance

The final goal of gene based classifiers is to answer a clinical or biology research need, so they have to compete with the current predictive models used in the respective fields. Thus, for the development and validation of clinically-relevant genomic tests a number of key stages must be successfully fulfilled (Simon, 2006):

- identify an important therapeutic decision which would need improvement;
- the target patient population should be homogeneous enough and treatment uniform, so that the results would be therapeutically relevant. Also, the economic considerations should not be overlooked: the treatment options and costs of misclassification should be such that the resulting classifier/test would be likely to be used in clinical practice (the test itself would incur some costs as well);
- develop the classifier and perform internal validation to assess whether the classifier appears to be sufficiently accurate relative to standard prognostic factors currently used. This means that initial analysis should prove the superiority (performance and/or costs at equal performance) of the new test with respect to current practice;
- translate the classifier to a platform likely to be used in practice. For example, a classifier relying on the use of several (in the order of tens) genes, even though it could be develop from microarray data, it is more likely that it implementation on qPCR would be more appealing to the clinicians/laboratories;
- demonstrate reproducibility of the results;
- independent validation of the complete test in prospective clinical trials.

5. Examples of gene-based classifiers

A simple search on PubMed (www.pubmed.com) portal for scientific literature lists hundreds of papers proposing new gene expression classifiers (sometimes called biomarkers), reflecting the importance these tools gained in the biomedical research. It would be a futile and inherently subjective attempt to list here "the most representative" results in the field. Therefore we will limit ourselves to mention just three such classifiers and some of their applications, each of them having something particular that makes them to stand out of the crowd.

5.1. Golub's ALL vs AML classifier

Golub's classifier (Golub et al, 1999) represents one of the first classifiers built in the early days of the microarrays. It was designed to distinguish between acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML), and was addressing the need for a standardized test to establish the diagnostic. Their training set consisted of $n = 38$ cases (27 ALL, 11 AML) profiled on an early Affymetrix chip ($p = 6817$ genes). They identified 50 genes correlated with the class distinction (based on a signal-to-noise ratio measure) and combined the genes into a score by "weighted vote" (i.e. linear combination of genes' expression values). And then they validated their predictor on an independent collection of 34 samples. By today's standards, this represents and easy problem, nevertheless the merit of this first system was to prove that building classifiers on gene expression is not only feasible

but could solve important diagnostic problems. The fact that their classifier relied on know oncogenes (like c-MYB, HOXA9) strengthen the confidence in such decision systems.

5.2. Compound covariate predictor

In (Radmacher et al, 2002) a generalization of Golub's classifier was proposed. Again, for a specimen i a score is computed as a weighted sum of the expression values of a number of genes,

$$s_i = \sum_j t_j \, x_{ij}$$

where the weights t_j are the signed t-statistics measuring the association of gene j with the class to be predicted. The sign is indicating if the gene is positively or negatively associated with the class. The score is then compared to a threshold computed as the average of the mean scores of each class. This *compound covariate predictor* is prototypical for large number of classifiers based on gene expression. It has the appealing property of being easily understandable as each gene contributes to the score proportional to its fold change between the two classes.

5.3. Top scoring pairs

The final example of gene based classifier is represented by Geman's Top Scoring Pairs (TSP) classifier (Geman et al, 2004), described in section 2.4. The striking feature of this classifier is its simplicity: for easier classification problems, it suffices to compare the expression levels of only two variables (genes) for taking a decision. However, as this is seldom enough for most of the problems, extensions of this algorithm have been proposed in which the top pairs are combined by majority vote (Tan et al, 2005) or by weighted combinations (Popovici et al, 2011). Despite its apparent simplicity, the classifier performs remarkably well on a large number of problems. Moreover, as the decision is taken by comparing the relative order of two genes, the classifier is extremely robust to noise and translates well from one platform to another.

Recently, this classifier was used to build a predictive model for identifying the colorectal cancer patients harboring a BRAF mutation (Popovici et al, 2012). In this study, the authors used the TSP to build a 64-gene-based classifier (32 pairs) to distinguish the BRAF mutant patients from those BRAF wild type and KRAS wild type. The training set consisted in 431 cases (of which 47 were BRAF mutants). Despite the highly imbalanced settings, the classifier's estimated performance (using repeated 5-fold cross-validation) was extremely good (sensitivity 95.8% and specificity 86.5%). The proper use of cross-validation procedure led to an accurate estimation of the performance, as the independent validation has shown: on three external data sets, the aggregated performance was: sensitivity 96.0% and specificity 86.24%. The classifier has demonstrated its good robustness, as the external validation sets were originating from different microarray platforms than the training set.

While the original purpose of the classifier was to predict the BRAF mutant patients, when applied to KRAS mutant population (which was not part of the training set) it segregated it into two subpopulations with clearly different gene expression patterns (on selected differentially expressed genes) – Figure 6.

Figure 6. Heatmap showing the different expression patterns within the KRAS mutant population, between BRAF-mutant-like patients (those predicted by the classifier, marked in red in the right column) and the rest of KRAS mutants.

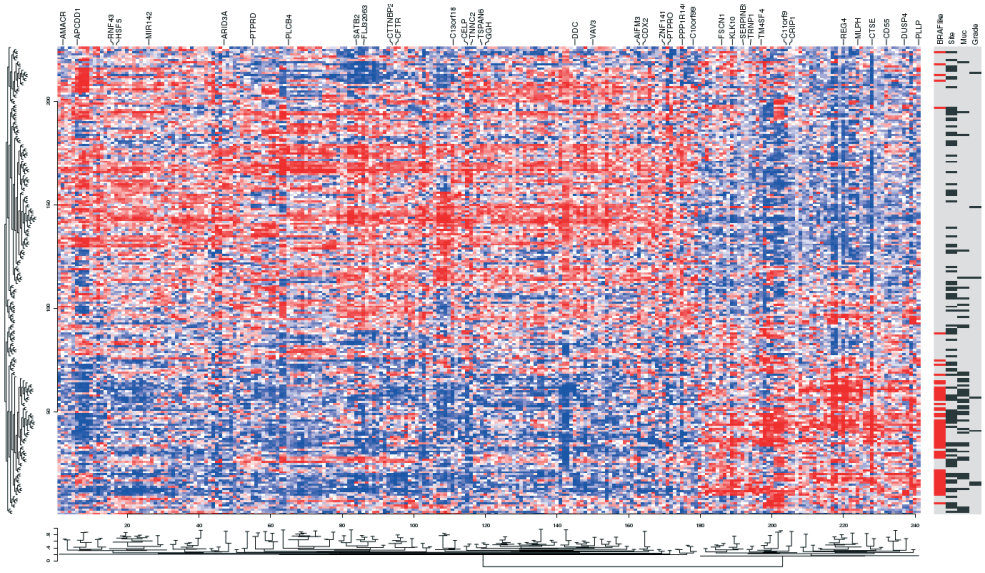
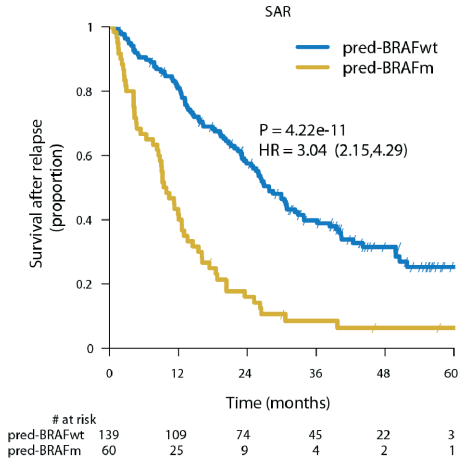


Figure 7. Survival after relapse: patients predicted to be "BRAF mutant" form a high risk group, with a median survival time of about 12 months.



The classifier had also strong prognostic value, i.e. it predicted the high risk patients. For examples, Figure 7 shows the Kaplan-Meier curves for the two populations predicted by the classifier (pred-BRAFm stands for "predicted BRAF mutant", while pred-BRAFwt for "predicted BRAF wild type"). This discovery is of clinical relevance, since it identifies a larger population at risk than initially considered by the clinical practice. Also, it opens new interventional avenues which would target specific pathways active only in this "BRAF-mutant-like" population. Finally, it is of importance also for the design of clinical trials since it clearly shows that the KRAS mutant population is not homogeneous and extra stratification factors should be taken into account.

6. Some concluding remarks

In this chapter we tried to briefly present a number of key concepts for understanding the classifiers in general and the specific issues arising from their application in the context of gene expression data. While for optimal application of classification algorithms intimate knowledge of the theory underlying their development is needed, for making good use of them a more superficial understanding of the principles of rigorous classifier development is enough. What remains extremely important is to understand the risks resulting from improper validation and performance estimation: the classifiers will never perform as expected.

7. References

Duda RO, Hart PE, Stork DG. 2001. Pattern classification. 2nd edition. John Wiley and Sons

Dupuy A, Simon R. 2007. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. Journal of National Cancer Institute 99:147-157

Geman D, D'Avignon C, Naiman DQ, Winslow RL. 2004. Classifying gene expression profiles from pairwise mRNA comparisons. Statistical Applications in Genetics and Molecular Biology 3(1): Article no. 19.

Golub TR, Slonim DK, Tamayo P, Huard C, et al. 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286: 531-537.

Hastie T, Tibshirani R, Friedman J. 2009. The elements of statistical learning. 2nd edition. Springer Verlag.

Newcombe RG. 1998. Two-sided confidence intervals for the single proportion: comparison of seven methods. Statistics in Medicine 17:857-872.

Pepe MS. 2003. The statistical evaluation of medical tests for classification and prediction. Oxford University Press.

Popovici V, Chen W, Gallas BG, Hatzis C, Shi W, Samuelson FW, Nikolsky Y, Tsyganova M, Ishkin A, Nikolskaya T, Hess KR, Valero V, Booser D, Delorenzi M, Hortobagyi G, Shi L, Symmans WF, Pusztai L. 2010. Effect of training sample size and classification difficulty on the accuracy of genomic predictors. Breast Cancer Research 12(1): R5.

Popovici V, Budinska E, Delorenzi M. 2011. Rgtsp: a generalized top scoring pairs package for class prediction. Bioinformatics 27(12):1729-1730.

Popovici V, Budinska E, Tejpar S, Weinrich S, Estrella H, Hodgson G, Van Cutsem E, Xie T, Bosman FT, Roth AD, Delorenzi M. 2012. Identification of a poor-prognosis BRAF-mutant-like population of patients with colon cancer. Journal of Clinical Oncology 30:1288-1295.

Radmacher MD, McShane LM, Simon R. 2002. A paradigm for class prediction using gene expressino profiles. Journal of Computational Biology 9(3):505-511.

Shi L, MAQC consortium. 2010. The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. Nature Biotechnology 28(8): 827-838.

Simon R. 2006. Roadmap for developing and validating therapeutically relevant genomic classifiers. Journal of Clinical Oncology 23:7332-7341.

Tan AC, Naiman DQ, Xu L, Winslow RL, Geman D. 2005. Simple decision rules for classifying human cancers from gene expression profiles. Bioinformatics 21(10):3896-3904.

Varma S, Simon R. 2006. Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics 7: Article no. 91.

Searching for robust and clinically relevant subtypes in high density molecular data

Eva Budinská[1,2,3]

[1] *Institute of Biostatistics and Analyses, Masaryk University; e-mail: budinska@iba.muni.cz*
[2] *Swiss Institute of Bioinformatics, Lausanne, Switzerland*
[3] *Masaryk Memorial Cancer Institute, Brno*

Abstract

Identification of clinically relevant molecular subtypes became an important tool in elucidating tumor biology. Robustness of the analytical approach employed for this task and consequently the robustness of derived subtypes is of major concern, if further experiments are to be conducted to confirm derived hypotheses. Here, we discuss multiple novel techniques for the control of robustness in cluster analysis designed for analysis of high-density molecular data.

Key words

Gene expression, molecular subtyping, consensus clustering, dynamic tree cut

1. Introduction

Subtyping based on high density molecular data, such as microarrays, aims at identifying groups of samples with similar molecular patterns. These can be for instance similar patterns of gene expression, microRNA expression or methylation. Finding molecular subtypes is very relevant mainly in medicine, especially in diseases that appear homogenous histo-pathologically, yet give a very heterogeneous response in terms of treatment outcome or survival. Recently, a lot of effort is dedicated to molecular subtyping of different cancers. Breast cancer for instance, was the first one where gene expression subtyping was applied, revealing a set of groups, that serve until now a basis for treatment consideration (Perou et al., 2002). Molecular subtypes help to elucidate the underlying biological mechanisms responsible for heterogeneity in tumour behaviour and help to focus the research on the subtype specific drugs targets, with hope to optimize treatment and ensure better prognosis of the given cancer as a whole. In order to make molecular subtypes clinically relevant, many additional analyses elucidating the biological, clinical and prognostic inference of subtypes are needed. The analysis becomes a fairly complex process involving different data-mining and statistical tools, together with thorough bio-medical interpretation of results.

Hereby, we will focus on the most important part of the subtyping procedure – robust clustering.

2. Example dataset

Throughout this article, we will use two datasets:

golub dataset - a microarray derived gene expression dataset available in R package `multtest` under the name `golub`, comprising 38 samples of three groups of acute leukaemia (AML – acute myeloid leukaemia, ALL-B – acute lymphoid leukaemia B cell

type and ALL-T – acute lymphoid leukaemia T-cell type) and gene expression values of 3051 genes. This was the first dataset used to demonstrate the use of gene expression data in cancer studies (Golub et al., 1999).

random dataset - a matrix of 1000 features and 100 samples, randomly sampled from normal distribution with 0 mean value and standard deviation equal to 1. This dataset will serve an example of a dataset without particular inner structure.

3. Robust clustering

Clustering – or, so called unsupervised learning - is the analytical approach used for subtype derivation. The main objective of clustering is to find distinct, preferably non-overlapping subpopulations within the large population of interest, members of which share similar pattern. Different basic clustering techniques exist and can be divided into model-based and distance-based methods. The model based methods use parametric assumptions on data distribution and often provide probabilities of cluster assignment. The distance-based methods are based on a similarity measure and can be further split into hierarchical and non-hierarchical, according to the algorithm they apply in order to group the samples. The detailed description of these methods and discussion on the choice of metrics is beyond the scope of this article and can be found elsewhere (Budinska et al., 2009).

In large genomic studies, hierarchical clustering is a particularly preferred method, because of its pattern visualization advantage. Often, not only clusters of samples, but also clusters of features – molecular entities - that underpin biological differences are of importance. Heatmap – a colored two dimensional plot with rows and columns representing samples and genes, ordered according to the hierarchical clustering dendrogram is one of the most often published type of figure in the field of large-scale molecular data (with the exception of DNA sequencing).

It is well known that the choice of clustering algorithm and metrics affects the final results, because clustering algorithms are biased towards partitions in accordance with their own clustering criterion. Moreover, clustering algorithms are designed to provide a data partition, even in non-existence of such a pattern, and the significance of these results must be assessed ad-hoc. While the clustering algorithms and corresponding metrics can be selected a-priori, based on the data type, our experience or published recommendations, two main issues are still to be addressed: i) the determination of the number of clusters and ii) the assessment of the confidence of the selection of number of clusters and cluster assignment for individual samples. Missing the external measure of class assignment (ground truth), the evaluation of clusters is based solely on internal validation measures, estimating the quality based on the intrinsic data values.

These issues are of particular importance in the data analysis of high density molecular data, which suffer from the curse of dimensionality problem. The small number of samples (tens to hundreds) and relatively huge number of molecular features (thousands, tens of thousands) makes clustering techniques susceptible to over-fitting, due to the sensitivity to noise, which is in these data much more abundant. This highly affects the robustness of the clustering to the sampling variability.

Resampling of the original dataset is away to simulate sampling variability. Although the idea of resampling in clustering is not new (Jain and Moreau, 1988), in the case of more noisy high-density molecular data, the preference is to avoid sampling with replacement, because replicated values can be artificially considered a separate cluster (Monti et al.,

2003). Multiple methods have been recently suggested to address these problems in the concept of microarray data analysis, mainly based on repeated resampling and consequent re-clustering of the original dataset, in order to study the behavior of the results when data is disturbed. This approach is simulating possible differences between different datasets, presumably resulting in a more robust result (for a review, see e.g. Handl et al., 2005).

For example a prediction-based resampling method *Clest* was designed (Dudoit and Fridlyand, 2002) in order to robustly estimate the number of clusters, showing the superiority of its performance in microarray data over six other methods, including more conventional such as *Silhouette* (Kaufman, Rousseeuw, 1990), or more recent such as *gap* (Tibshirani et al., 2001). However, this method does not solve the problem of the assessment of the confidence of cluster assignment for individual samples. A new method assessing both problems – *consensus clustering* (Monti et al., 2003) - was suggested and was successfully applied in different cancer subtyping analyses. In a comparative study (Giancarlo et al., 2008) this method was also evaluated the best method in terms of performance and algorithm independency. We will dedicate the following subsection to the description of this method.

3.1. Consensus clustering

Is a resampling and re-clustering based method designed to represent the *consensus* across multiple runs of a clustering algorithm (Monti et al., 2003), in order to:

- determine the number of clusters in the data and to assess the stability of the discovered clusters
- represent the consensus over multiple runs of a clustering algorithm with random restart, so as to account for its sensitivity to the initial conditions.

In addition, it serves a visualization tool for the evaluation cluster number, membership, and boundaries.

The basic principle is to disturb the structure of the original $N \times P$ data matrix by random selection of a subset of samples and/or features. The new dataset is then consequently clustered, given the selected clustering algorithm, similarity measure and number of clusters or tree cut height. This resampling and clustering is repeated L times. In the l-th run, the cluster membership of samples is recorded and two $N \times N$ matrices are created:

- *connectivity matrix* $C^{(l)}$ that stores for each pair of samples *i,j* the information whether they were clustered together, e.g. $C_{ij}^{(l)} = 1$ if sample *i* and *j* belong to the same cluster, 0 otherwise
- *indicator matrix* $I^{(l)}$ that stores for each pair *i,j* the information whether they were both selected in the resampling, e.g. $I_{ij}^{(l)} = 1$ if sample *i* and *j* were in the same selection, 0 otherwise

After all l runs, the *consensus matrix* M is calculated by dividing the number of times two features were found together in the same cluster by the number of times that they have been selected together in the sampling subsets. A consensus matrix is therefore a N × N matrix that stores for each pair of items the weighted proportion of clustering runs in which the two items were clustered together:

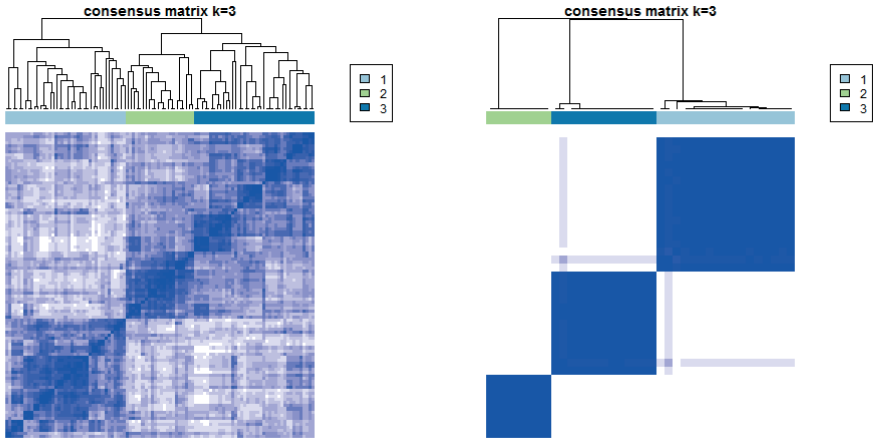
$$M_{ij} = \frac{\sum_{l=1}^{L} C_{ij}^{(l)}}{\sum_{l=1}^{L} I_{ij}^{(l)}}$$

The idea behind this approach is that samples that are frequently found in the same cluster represent more reliable cluster members than those who cluster together less frequently, being more sensible to the random noise and changes in feature selection. Each entry of the consensus matrix is a *consensus index* of a given pair of samples, with values from 0 (no consensus, samples were never members of the same cluster) to 1 (perfect consensus, clusters were members of the same cluster 100% times). Consensus matrix M represents a robust similarity measure and 1-M is a distance matrix, that can be used as an entry to hierarchical clustering in order to obtain a final robust clustering. Figure 1 demonstrates a result of hierarchical clustering applied on consensus matrix on our example data. While the consensus matrix of the *random* dataset is very unstructured, a very clear three-class structure is visible for the *golub* dataset.

The consensus matrix between samples can be directly used to define statistics of stability of clusters and cluster sample assignments. If I_k be a set of indices of samples belonging to cluster k, the consensus measure of a cluster k - *cluster consensus* - can be defined as an average consensus between all pairs of samples belonging to the same cluster:

$$m^k = \frac{1}{N_l(N_l - 1)/2} \sum_{\substack{i,j \in I_k \\ i<j}} M_{ij}$$

Figure 1. Heatmap representation of the consensus matrix for the *random* dataset – left and the *golub* dataset – right, for three clusters. The colour ranges from white representing 0 consensus to bright blue, representing the consensus of 1.



The corresponding *sample consensus* for each sample s_i and cluster l can be defined as:

$$m_i^k = \frac{1}{N_l - 1\{s_i \in I_k\}} \sum_{\substack{j \in I_l \\ j \neq i}} M_{ij} ,$$

where $1\{s_i \in I_k\}$ is the indicator function that equals 1 if $\{s_i \in I_k\}$ is true, 0 otherwise. The sample consensus is the average consensus index of the sample to all members of the cluster. Both measures can be used to identify outliers – either clusters with relatively low consensus, suggesting remaining heterogeneity in the cluster, or samples, that could be considered outliers because of very small consensus to any other sample in the dataset.

Consensus matrix can be also used to estimate the optimal number of clusters. For details, see section 2.2.

3.1.1. Other consensus clustering techniques

Multiple variations of the consensus clustering method exist and are a natural extension of the original algorithm.

Method called *merged consensus clustering* (Swift et al., 2004), in contrast to the method of Monti et al., creates the consensus matrix as a function of runs of consensus clustering with multiple different algorithms. This should eliminate the possible negative effect a single algorithm, which might not be suitable for the particular type of data.

Weighted clustering (Deohdar and Ghosh, 2006) builds on the idea that the clusterings produced within a consensus clustering procedure are not necessarily of the same quality. If an external metrics of quality exists, one should be able to integrate this in order to weigh the contributions of each clustering to the final consensus matrix, which is then calculated as

$$M_{ij} = \sum_{k=1}^{K} w_k C_{ij}^{(k)} ,$$

where w_k is weight of the particular clustering. This method also uses different clustering algorithms and different distance measures.

For the comparison of different consensus clustering algorithms, see for example (Goder and Filkov, 2008).

3.1.2. R-packages for consensus clustering

Two major packages are available in R for consensus clustering. The package `ConsensusClusterPlus` (Wilkerson, 2011) provides all the algorithms and metrics as described in (Monti et al., 2003). It is a part of Bioconductor repository and can be installed directly from R console using command:

```
>source("http://bioconductor.org/biocLite.R")
>biocLite("ConsensusClusterPlus")
```

Package `clusterCons` implements the merged clustering of (Swift et al., 2004). It can be installed directly from R console by using the `install.packages()`command.

3.2. Determining the number of subtypes

In this section, two methods for determining the number of clusters are discussed. Both were developed specially for microarray data analysis and hierarchical clustering algorithm.

3.2.1. Consensus measure

Consensus matrix – as described in section 2 - can be also used to estimate the optimal number of clusters. If consensus clustering is run for different cluster number values *k=1..K*, the decision criteria can be based on the calculation of for example the average intra-cluster consensus for each *k*. (Monti et al., 2003) propose another measure - the *empirical cumulative distribution* (CDF)

$$CDF^{x} = \frac{\sum_{i<j} 1\{M_{ij} \le x\}}{N(N-1)/2},$$

which compares the distribution of histograms of entries of consensus matrix *M* for each *k*. If clustering with *k* clusters represents a perfect partition, histogram of consensus matrix entries will consist of two bins over 0 (no consensus at all between samples from different clusters) and 1 (perfect consensus between clusters from different samples). The optimal number of clusters can then be decided by computing the area under CDF curve and by examining its relative change between different *k* (*delta area*). The CDF measure, however, is applicable mainly for hierarchical clustering, for which the method was designed. Figure 2 shows examples of histograms for k=3 and k=6 and CDF and delta area for the golub data. While histogram of consensus measures for the three cluster structure (heatmap on Figure 1 right) reveals indeed majority of values on 0 or 1, six cluster structure has a substantially decreased number of perfect consensus between samples and increased number of values between 0-1 suggesting instability of this number of clusters. The delta area plot shows that increasing number of clusters from 2 to 3, the area under CDF gains around 0.36, while further increasing the number of clusters to 5 has no real impact on the area under CDF change and therefore the estimated value of k would be 3 or 4 subtypes.

3.2.2. Dynamic Tree Cut

As already mentioned, hierarchical clustering has its particular importance in genomic data analysis. In comparison to other clustering techniques, clusters are defined ad-hoc, by cutting the branches of the hierarchically structured similarity tree – dendrogram – the output of this clustering – on a fixed height. All the branches below this cut are preserved and represent final clusters. The major disadvantage of this static cut approach is that often, different clusters are present on different cut heights – naturally presenting more or less similar groups of samples, and cutting low in order to obtain a cluster with high internal similarity results in the loss of structure of clusters with relatively lower similarity.

To address this problem, a set of novel dynamic branch cutting methods for detecting clusters in a dendrogram of hierarchical clustering was recently proposed (Langfelder et al, 2007). In this approach, clusters are being defined depending on their shape. The huge advantage is that the system of cluster determination is flexible – a set of parameters can be used to control the resulting cut – such as for instance cut height, minimal cluster size or minimal intra cluster. First method called *Dynamic Tree* - this is a flexible extension of the static cut, works solely with the structure of the dendrogram. The second method *DynamicHybrid* dynamically crawls the dendrogram in the bottom-up direction and after defining clusters offers a possibility of additional assignment of the unassigned samples to the closest core clusters defined in the first step, if the requirements on cluster internal similarity are met. The description of both algorithms is fairly complex and I do strongly

recommend the reader to consult the original paper for more details. Dynamic Tree Cut methods are implemented in R package `dynamicTreeCut`.

An example of comparison of a static and dynamic cut of dendrogram is demonstrated on the golub dataset in Figure 3. Both static cut and DynamicHybrid algorithm (represented by function `cutreeHybrid`) were run with cut height of 1.2. Minimal cluster size selected for cutreeHybrid was 3 and 5. While the static cut on the selected height identifies 3 clusters, cutHybrid with minimal cluster size of 5 identifies four major clusters. Decreasing the cluster size to 3 identifies further, yet still consistent splits.

Figure 2. Example of CDF derivation and selection of number of clusters on golub dataset. Consensus CDF and delta area plot are shown for k varying from 2 to 10.

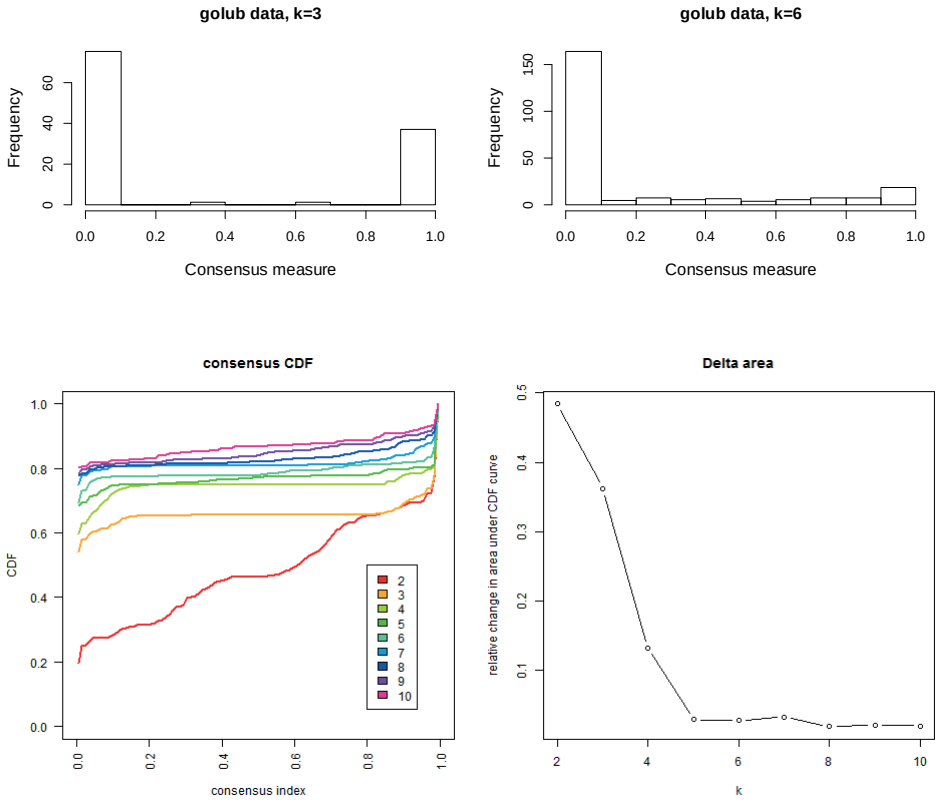
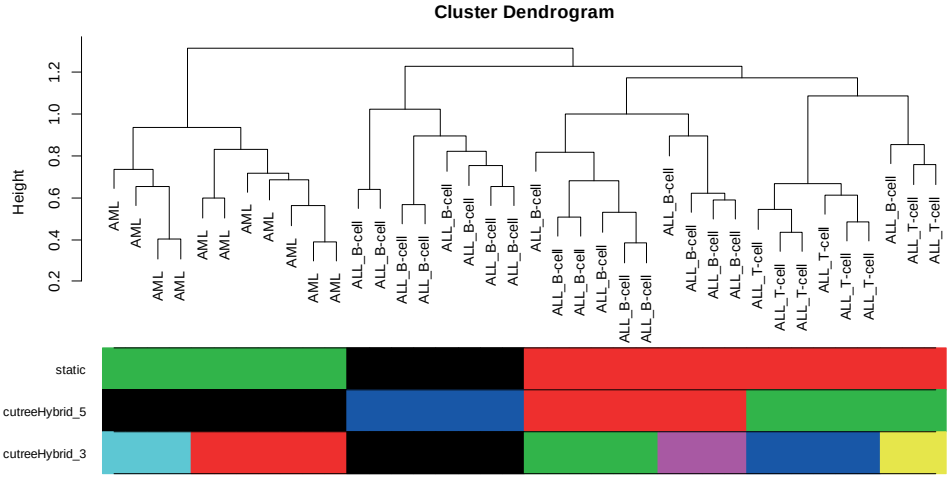


Figure 3. Comparison of static and DynamicHybrid cut (as output by `cutreeHybrid` function of the R package `dynamicTreeCut`)on the dendrogram from hierarchical clustering with average linking algorithm and correlation-based distance between samples of golub data.



4. Other analytical challenges

Robustness of findings is one of the most important aspects of the applied research, and is indispensable for the clinical relevance. In order to call the subtypes robust, it is vital that the patterns defining the groups we find are not specific for a particular dataset, but can be found in other similar populations. We say that subtypes must be validated. However, the validation in a de-novo developed subtyping system has somewhat different meaning and is a much less evident analytical task than in the construction of classifiers. This is because it is not obvious to validate a pattern without existing objective class label (the ground truth). Without the ground truth, the validation can be done only indirectly, by the assessment of subtype specific differences in population characteristics that were not used for their construction. Different survival experience or clinically relevant variables are examples of such characteristics.

Often, a development of a subtype classifier is necessary in order to make the results applicable for the practice. Preferably, such a classifier will be accurate and robust to different technological platforms used to derive data and will be able to classify one sample only. This classifier can also serve to call subtypes in the validation set. However, in a complex analysis of subtyping, many decisions on types of methods and choice of parameters must be made. Although some can be subjected to sensitivity analysis exploring the effect of different choices on clustering results (such as similarity metrics or clustering algorithms), it is almost impossible to perform such an analysis for all considered parameters and algorithmic choices, due to the complexity of the problem. For this reason, simple application of the classifier on the validation set and consequent comparison of external characteristics of training vs. validation subtypes is not the optimal solution. Better solution would be to reproduce the subtyping on the validation set, using the same methods and parameters as selected for the training set.

5. Concluding remarks

We have seen a selection of state-of-the-art approaches for robust clustering in the molecular subtyping. However, the field is evolving very quickly and reader is strongly encouraged to search for the methodological improvements and critically review all the information provided with respect to the nature of the particular data analysed.

Some concepts remain, though, the same. The robustness and reproducibility in clinical research is indispensable. One should never search for the final and unchangeable answer – which is almost impossible to achieve because of the nature of biology and technological limitations - but rather focus on the extraction of the most essential information from the data that are available. In this respect, application of consensus clustering base methods seems inevitable, although in case of hierarchical clustering, one might consider to use rather dynamic Tree Cut for cluster assignment, as it allows for identification of core samples, without forcing the less representative samples to be assigned a cluster membership.

6. References

Budinska E, Bortlicek Z. 2009. E-learning E-learning in analysis of genomic and proteomic data. URL: http://telemedicina.med.muni.cz/genomic-proteomic-analysis/index-en.php.

Deodhar M, Ghosh J. 2006. Weighted Consensus Clustering for Microarray Data Analysis. In: Dagli, CG et al. (ed): Intelligent Engineering Systems through Artificial Neural Networks, Volume 16, ACMA.

Dudoit S, Fridlyand J. 2002. A prediction-based resampling method for estimating the number of clusters in a dataset. Genome Biology 3(7): Article no. 0036.1

Goder A and Filkov V. Consensus Clustering Algorithms: Comparison and Refinement. 2008 Proceedings of the Ninth Workshop on Algorithm Engineering and Experiments (ALENEX) — San Francisco, 19 January 2008. Society for Industrial and Applied Mathematics.

Golub TR, Slonim DK, Tamayo P, Huard C, et al. 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286: 531-537.

Giancarlo R, Scaturro D, Utro F. 2008. Computational cluster validation for microarray data analysis: experimental assessment of Clest, Consensus Clustering, Figure of Merit, Gap Statistics and Model Explorer. BMC Bioinformatics, 9: Article no. 462

Handl J, Knowles J, Kell DB. 2005. Computational cluster validation in post-genomic data analysis. Bioinformatics 21 (15): 3201-3212.

Jain AK. and Moreau J. 1988. Bootstrap techniques in cluster analysis. Pattern Recognition 20: 547–568

Kaufman L, Rousseeuw PJ. 1990. Finding Groups in Data: An Introduction to Cluster Analysis. New York: Wiley.

Langfelder P, Zhang B, Horvath S. 2007. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. Bioinformatics 24(5):719-720

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. Aquatic Microbial Ecology 30: 83–89.

Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lønning PE, Børresen-Dale AL, Brown PO, Botstein D. 2000. Molecular portraits of human breast tumours. Nature 406: 747–752.

R Core Team. 2012. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Simpson TI, Armstrong JD and Jarman AP. 2010 Merged consensus clustering to assess and improve class discovery with microarray data. BMC Bioinformatics, 11: Article no. 590.

Tibshirani, R., Walther, G. and Hastie, T. 2001. Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63: 411–423

Wilkerson M. 2011. ConsensusClusterPlus: ConsensusClusterPlus. R package version 1.8.0.

Meta-analysis – taking the information further

Hana Imrichová[1], Eva Budinská[1,2,3]

[1] *Institute of Biostatistics and Analyses, Masaryk University; e-mail: budinska@iba.muni.cz*
[2] *Swiss Institute of Bioinformatics, Lausanne, Switzerland*
[3] *Masaryk Memorial Cancer Institute, Brno*

Abstract

In this chapter we will provide a brief overview of the opportunities offered by the meta-analytical approach to gene expression data analysis. Meta-analysis allows the combination of data or intermediate statistical results from multiple studies and thus it results in more robust conclusions. After introducing the basic ideas of meta-analysis, we will describe their application in the context of a large analysis that led to the identification of surprising connections between two different solid tumour cancers – breast and colorectal carcinomas.

Key words:

Meta-analysis, gene expression data, cancer molecular subtype, gene-module, meta-gene

1. Introduction

DNA microarrays are high-throughput technologies often used in biology for measuring the expression of thousands of genes in a single experiment (so called "gene expression profiling") (Somasundaram et al., 2002). These expression profiles can be used for identification of differentially expressed genes between two phenotypes (for example cancer and normal tissue) or to perform more advanced analyses e.g. studies of the similarities at molecular level between various types of cancer. An increasing number of studies use DNA microarrays for gene expression profiling and the resulting data collections become accessible to the scientific community. Meta-analysis is the ideal tool for the more efficient use of these data sets and combining their samples thus increase the sample size of the analysis. The inferences made in the framework of such analyses are expected to be more robust than from any analysis of single data sets. Here we will exemplify such a large scale meta-analyses within a project aiming at revealing the connections between two types of cancers by comparing co-expressed genes.

2. Case study

Cancers in general are very heterogeneous and this high degree of heterogeneity can be observed even within a single cancer type. That is why two patients with the same diagnosis can have different prognosis or responses to therapy. Breast cancer molecular subtypes are well characterized and it is of interest to know, whether this knowledge can be used in explaining heterogeneity of other cancers, such as colorectal cancer. To this end, we will apply a meta-analytical approach to a large collection of expression-clinical-prognostic studies in breast cancer.

Two sets of data are used for illustrative purposes of using of the meta-analytical approach:

- The first set consists of 54 gene modules (groups of genes with highly correlated gene expression profiles) that were defined in colorectal cancer. These modules are representing biological processes and their expression patterns define subtypes of colorectal cancer. They were obtained as a result of a previous gene expression study in colorectal cancer (Budinská et al., 2012). The file contains only names and IDs of genes and numbers of corresponding gene modules.
- The second set of data comprises 11 publicly available gene expression datasets from breast cancer studies. These datasets had been $\log_2$-normalized as were used in a previous study (Haibe-Kains et al, 2012). Each dataset contains information about gene expression for each patient (2245 patients with breast cancer in total). In addition, clinical and survival data are available for many of these patients.

3. Meta-analysis

Meta-analysis provides the framework and the method for combining results (test statistics, effect sizes, p-values) from different studies. The hope is to gain statistical power from the increased sample size and, in the meantime, to reach more robust conclusions. Instead of combining the results, one may attempt to pool the data into a single larger data sets. This second approach has first to address the problem of removing the batch effects, which are typical in the case of gene expression data sets, which originate from various laboratories and microarray platforms.

We will first describe this second approach (section 3.1) and then we will delve into meta-analytical approach. All analyses can be performed in R/Bioconductor - language for statistical computing (R Development Core Team, 2010).

3.1. Removing batch effect between datasets

Before comparison datasets from different studies, one has to solve the problem of a batch effect to be able to combine microarray datasets in order to increase the statistical power, often in the framework of sample clustering. To some extent, batch effect can be removed by centering the (meta)genes to their respective mean/median values in each dataset or, more special methods can be applied, such as the one named `ComBat` – an R function that implements a method used for adjusting batch effects in microarray expression data using empirical Bayes methods (Johnson and Li, 2006).

3.2. Data dimensionality reduction

In the analysis of genomic data, it is common to reduce the dimensionality of thousands genes by creating several co-expressed modules - groups of genes with correlated expression. While correlated gene expression measures provide redundant information, gene modules who are often representing biological processes are more reproducible than individual genes across different microarray platforms. Gene module is represented often by a single representative value – called ‘module score’ or ‘metagene’. This single value is computed for each patient for instance as a median or an average of values of gene expression belonging to genes within each gene module. By employing metagenes, the dimensionality of data can be drastically reduced while retaining simplicity of interpretation, in contrast to data reduction by principal component analysis.

3.3. Correlating across datasets

Assessment of the internal gene module correlation structure in other datasets, before performing any additional analyses on summarized meta-gene values, is of particular interest. The simplest way would be to score each module with respect to its internal gene-gene correlations by a dataset-wise average, median or lower quartile. However, more robust measure can be obtained by Z-transforming the correlations coefficients across these datasets.

In our example, it was of interest to know which of the colorectal cancer gene modules are also highly correlated in breast cancer. For this purpose, we combined individual correlations into a meta-correlation defined as negative of Fisher Z-transformed (inverse hyperbolic tangent transformed) Pearson's correlation.

The computation of the Pearson correlation r_{ijk} for each pair of gene (*i*, *j*) is performed in each study *k*, followed by a transformation of *r* using Fisher's method:

$$z_{ijk} = \tanh^{-1}(r_{ijk}) \quad . \tag{1}$$

In the next step, the z-scores are combined across all datasets using a meta-analysis. This combined correlation (meta-correlation) is used as a measure of similarity between pairs of genes *i* and *j*:

$$z_{ij} = \sum_{k=1}^{K} \frac{z_{ijk}}{\sqrt{K_{ij}}}, \tag{2}$$

where K_{ij} is the number of datasets where genes *i* and *j* are present.

The quality criterion for significantly correlated modules (for example a minimum of 75 % significant gene-gene meta-correlations in the module) can be derived from p-values obtained from nonparametric permutation test – 10 000 random modules consisting of *n* random genes are generated for each gene module consisted of *n* genes. The lower quartile Q1 of the tested colorectal cancer module with *n* genes is compared to the lower quartile of the distribution derived in non-parametric permutation test:

$$p\text{-}value = x + 1 / 10\,000 + 1, \tag{3}$$

where x is the number of the Q1 paired meta-correlations in random modules, for whose the value of meta-correlation is greater than or equal than the Q1 paired meta-correlation of the corresponding colorectal cancer module.

In our example, modules with $p < 0.001$ after Bonferonni correction were determined as significantly correlated.

3.4. Hierarchical clustering

As described in detail in previous chapter, hierarchical clustering is frequently used method for cancer subtyping, because it has the added advantage of providing a visualisation representation of the results. The similarity between samples can be calculated either on the whole set of genes, or on meta-genes, reducing thus a dimensionality of the data. However, when combining multiple datasets, the same problem as adressed above arises – how to calculate similarity across datasets? Again, meta-analytical approaches needs to be used.

In our example, we used *nclust2* – a function of R package *nclust* (Wirapati – personal communication) which enables hierarchical agglomerative clustering of both rows (patients)

and columns (genes) across multiple datasets, implementing the meta-analytical correlation as described in previous subsection for clustering of genes.

In the case of sample clustering, the expression profiles are compared both within datasets and among all datasets. Therefore it is necessary to know whether measures of similarity have biological support (not technical – it would lead to clustering of samples from the same datasets). Here, batch effect removal methods as described in section 3.1. can be applied.

The result of clustering can be visualized in a form of a heatmap, e.g. using *coldmap* (also a function of R package *nclust*). The *coldmap* displays the result of clustering in a color coded heatmap of gene expression matrix median-centered (meta)gene expression values, but dendrogram of samples is split per dataset for visualization of the consistency of pattern across datasets. In addition, clinical information of each patient and each dataset can be added as a color-coded side panel.

New clusters of samples can be determined by application of dynamic pruning method described in previous chapter - *cutreeHybrid* – a function of R package *dynamicTreeCut*, which is specifically designed for hierarchical clustering of microarray data (Langfelder et al., 2008).

3.5. Hypothesis testing

Hypothesis testing plays an important role in the analysis of genomic data. It may serve for identification of differences in continuous variables between groups, such as (meta)gene expression. Under the null hypothesis, there is no difference between groups and rejection of this hypothesis flags a gene as a potentially biologically important.

When pooling the data from several sources, the same criteria are applied as already mentioned above. Either the analysis is performed in each dataset separately and resulting p-values or effect-sizes are prone to meta-summarization methods, or – in case of pooling data in order to increase the statistical power – a method for batch effect removal must be applied. Assuming no residual batch effect, one can use known parametric or nonparametric statistical tests for identifying the differentially expressed (meta)genes. The usual parametric test assumes Gaussian distribution for the groups and test whether the means of the groups are equal. The t-test serves for comparing two groups and the analysis of variance (ANOVA) can be used to mutually compare more than two groups. In an assumption-free test, one uses nonparametric tests comparing the medians of tested groups. We can use Wilcoxon rank sum test and Kruskal-Wallis test comparing two and more than two groups, respectively.

Statistical testing in genomics results in multiple hypothesis problem, because tens to thousands of (meta)genes can be tested simultaneously. Therefore the chance of observation of false positive results increases and in consequence it is necessary to correct the level at which the null hypothesis is rejected. Standard methods include False Discovery Rate (FDR) correction or more conservative Bonferroni correction for adjusting p-values.

Pearson's Chi-squared or Fisher's exact test can be used for categorical variables (for testing of the association between cancer subtypes and clinical variables). The Fisher’s exact test is recommended when expected cell counts are less than 5.

In our example, due to the non-normality of the majority of variables tested, nonparametric statistical tests were used for testing for significantly differentially expressed metagenes among groups stratified according to levels of clinical variables and markers. Associations between metagenes and binary clinical variables were assessed by Wilcoxon rank sum test. Kruskal-Wallis test was used in the case of variables with more than 2 levels. Adjustment for multiple hypothesis testing was performed by conservative Bonferroni correction.

Associations between new molecular subtypes and known clinical and demographic variables and molecular markers were tested using Pearson's Chi-squared test. All statistical tests were two-sided and results were considered significant at $p < 0.05$.

3.6. Survival analysis

Survival analysis is very valuable in genomic analysis. This is particularly true when new groups of patients are being derived, since the differences in survival would be indicative of clinical relevance of the identified groups. For this, one needs survival data corresponding to patients from gene expression datasets – for example RFS (*relapse free survival*), OS (*overall survival*), or DMFS (*distant metastasis free survival*). Then Kaplan-Meier estimates of survival (or other method) can be used for estimating each group's survival probability functions.

Pairwise differences in survival experience between groups of patients can be assessed using log-rank test (if survival functions do not intersect) or Gehan-Wilcoxon test (otherwise). Because of possible difference among populations one needs to adjust for the effect of the dataset – e.g. in the Cox proportional hazards model by applying the strata parameter of the `coxph` function of the `survival` package in R.

4. Case study results and discussion

Applying methods described above within a large meta-analysis on our example data (54 colorectal cancer modules and 11 gene expression datasets from breast cancer studies), we obtained the following results (Imrichová, 2012):

The analysis of meta-correlations revealed that 28 of 54 gene expression modules as defined in colorectal cancer are significantly correlated also in breast cancer datasets, which means that these 28 modules can be used for classification of breast cancer too. Metagenes of these 28 modules are clustered into higher level structures corresponding to molecular functions identically as in (Budinská et al, 2012). These groups of metagenes are connected with proliferation, immune response and EMT (epithelial-mesenchymal transition).

Statistical testing showed that these significantly correlated metagenes are statistically significantly different in their expression among groups defined by majority of clinical variables and breast cancer markers. A parallel biological significance (a difference of medians of $\log_2$ expression of a metagene within each group ≥ 1) was shown only for three metagenes connected with proliferation in the relation to variables histological grade and intrinsic subtypes. This discovery is not surprising, because histological grade relates to proliferation and very often is used as a surrogate instead of proliferative index Ki67 in IHC classification (Tamimi et al., 2012). These proliferative metagenes also showed lower expression within luminal A intrinsic subtype – the subtype that is characterized by low proliferation (Wirapati et al., 2008).

The sample hierarchical clustering based on expression patterns of 28 colorectal cancer metagenes was performed and new five breast cancer subtypes was determined by *dynamic tree cut* (Langfelder et al., 2008). The result is shown in the form of previously mentioned *coldmap* on Figure 1. Luminal A tumours (that constitute a major part of new subtypes 2 and 3) were best separated. This subtype was separated probably because of metagenes related to proliferation, because luminal A (as the only one of intrinsic subtypes) is characteristic of low expression of proliferation AURKA module that is used by breast cancer classifiers (Wirapati et al., 2008; Desmedt et al., 2008; Haibe-Kains et al., 2012).

Colorectal cancer modules do not contain any gene of ERBB2 module that is used by breast cancer classifiers to determine HER2+ intrinsic subtype, therefore it was not surprising that these colorectal cancer metagenes did not separate HER2+ tumours from the rest of highly proliferating intrinsic subtypes (basal-like and luminal B). On the other hand, colorectal cancer modules of immune cluster (whose genes are almost not found in breast cancer classifiers) divided intrinsic HER2+, basal-like and luminal B subtypes into two groups (creating new subtypes 1 and 4) according to expression of these modules. The new subtype 5 is composed mainly of luminal B tumours and is characteristic by low expression of metagenes related to immune response.

It is worth noting that an association was observed between the new subtype classification and the tumour size, status of ER, PGR, HER2 and histological grade. But nodal status and age at diagnosis were independent of the new molecular subtypes. This resulted from Pearson's Chi-squared test.

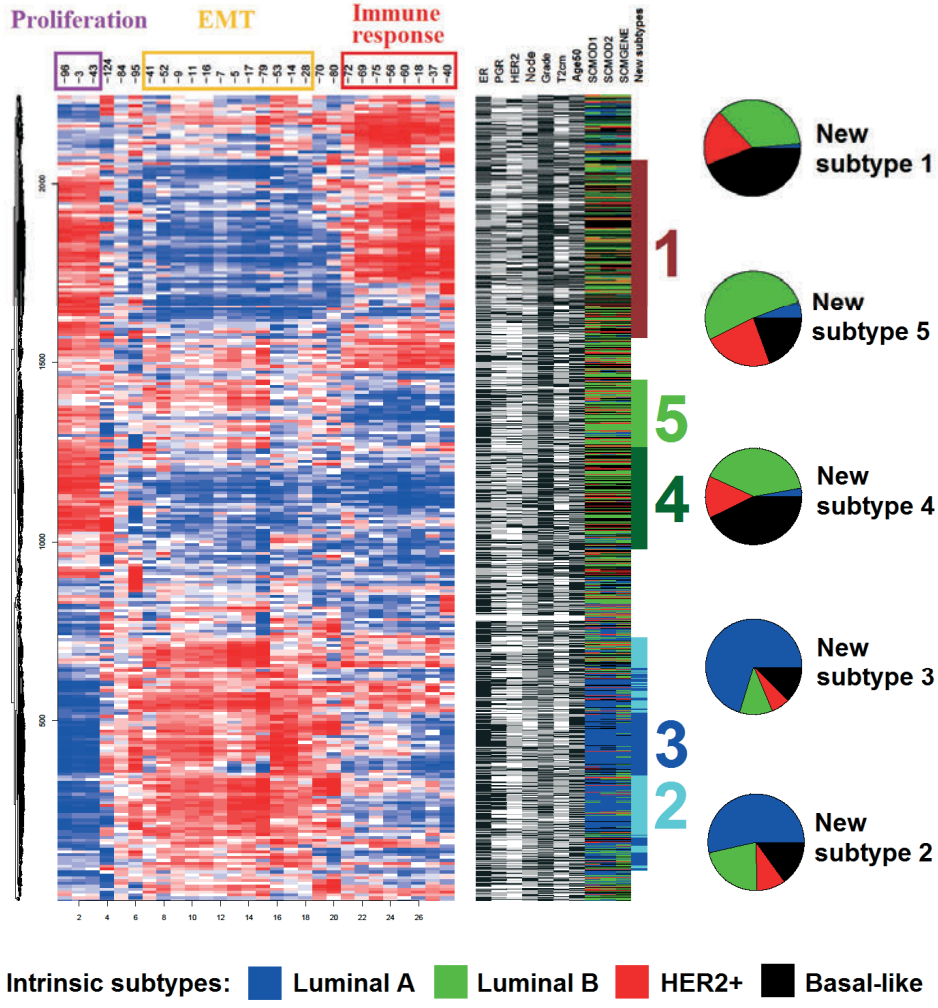
Survival analysis revealed that the new subtypes are significantly different with regard to OS, DMFS and RFS. Patients with new subtypes 1, 2 and 3 had good survival, while patients with subtype 4 and 5 had poor survival. Very interesting is the significantly different survival between subtypes 1 and 4 that are very similar by their composition of intrinsic subtypes. These two subtypes vary in the expression of their immune metagenes. The subtype 1 has these metagenes active contrary to subtype 4. An important role of the immune metagenes for patient survival was subsequently confirmed by the next testing.

The main conclusion of the findings summarized above obtained by a meta-analytical approach is that gene modules representing biological processes (proliferation, EMT and immune response) and defining colorectal cancer subtypes can be used to derive breast cancer groups that are significantly different from the intrinsic subtypes and differ in the distribution of clinical variables and survival time. The detection of these connections can help understanding the heterogeneity of colorectal cancer. The future research of colorectal carcinoma therapy could follow the same way as research of breast cancer treatment. For example, the same therapy could be used for some of subtypes of colorectal carcinoma with similar expression patterns of gene modules as some of the new subtypes of breast cancer derived by these modules. It was encouraging to find that some of the hypotheses we derived in this analysis have been already published or alluded to in the literature.

5. Conclusions

Many gene expression datasets have been published recently. A meta-analytical approach enables the mathematical combination of two or more datasets in order to improve the reliability of the results. This analytical exercise often requires clever switching between purely meta-analytical approaches that combine results from individual datasets (p-values of effect sizes), and batch effect removal techniques where the increase of statistical power is needed. We have shown above both approaches and their usefulness in an assessment of internal correlation structure of gene modules, in the analysis of differentially expressed (meta)genes, clustering of genes and samples or statistical testing of the significance of (meta)genes or groups of patients with regard to clinical variables and markers or carrying out a survival analysis.

Figure 1. A coldmap created on 11 publicly available gene expression datasets from breast cancer studies. Columns are metagenes, rows are patients. The coldmap shows increased (red) or decreased (blue) gene expression of metagenes. The bars on the right show clinical information of each patient and classification to intrinsic breast cancer subtypes. Pie diagrams demonstrate composition of the new breast cancer subtypes by the intrinsic breast cancer subtypes.



6. References

Budinska E, Popovici V, Tejpar S, Lapique N, Sikora KO, Di Narzo AF, Hodgson JG, Weinrich S, Bosman F, Roth A, Delorenzi M. 2012. Identification and validation of gene expression subtypes in a large set of colorectal cancer samples. ASCO Annual Meeting 2012, 1-5 June 2012, Chicago, USA . Journal of Clinical Oncology 30(suppl): Abstract no. 3511.

Desmedt C, Haibe-Kains B, Wirapati P, Buyse M, Larsimont D, Bontempi G, Delorenzi M, Piccart M, Sotiriou C. 2008. Biological Processes Associated with Breast Cancer Clinical Outcome Depend on the Molecular Subtypes. Clinical Cancer Research 14: 5158-5165.

Haibe-Kains B, Desmedt C, Loi S, Culhane AC, Bontempi G, Quackenbush J, Sotiriou C. 2012. A Three-Gene Model to Robustly Identify Breast Cancer Molecular Subtypes. JNCI Journal of the National Cancer Institute 104: 311-325.

Imrichová H. Comparison of subtypes of colorectal carcinoma with breast cancer subtypes by correlation of their gene expression profiles. 2012. Masaryk University, Faculty of Science. Diploma thesis.

Johnson WE, Li C. 2006. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 8: 118-127.

Langfelder P, Zhang B, Horvath S. 2008. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. Bioinformatics 24: 719-720.

R Development Core Team. 2010. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Somasundaram K, Mungamuri SK, Wajapeyee N. 2002. DNA Microarray Technology and its applications in Cancer Biology. Applied Genomics and Proteomics 1: 209-218.

Tamimi RM, Colditz GA, Hazra A, Baer HJ, Hankinson SE, Rosner B, Marotti J, Connolly JL, Schnitt SJ, Collins LC. 2012. Traditional breast cancer risk factors in relation to molecular subtypes of breast cancer. Breast Cancer Research and Treatment 131: 159-167.

Wirapati P, Sotiriou Ch, Kunkel S, Farmer P, Pradervand S, Haibe-Kains B, Desmedt Ch, Ignatiadis M, Sengstag T, Schutz F, Goldstein DR, Piccart M, Delorenzi M. 2008. Meta-analysis of gene-expression profiles in breast cancer: toward a unified understanding of breast cancer sub-typing and prognosis signatures. Breast Cancer Research 10: R65.

Tree of Life in a gappy genomic era

Natália Martínková[1,2]

[1] *Institute of Biostatistics and Analyses, Masaryk University, Brno;*
e-mail: martinkova@ivb.cz
[2] *Institute of Vertebrate Biology, Academy of Sciences of the Czech Republic, v.v.i., Brno*

Abstract

Increasing volume of publicly available DNA sequence data enables comprehensive studies that address integrative questions. For these projects, bioinformatic analysis requires advanced methods and computational infrastructure. I present the character of DNA sequence matrices for multilocus datasets, which contain large portions of missing data. A condition critical for analysis of multilocus data is that datasets for all loci or genes need to have partially overlapping taxon sets. The work-flow for analysing such data differs between supermatrix and supertree estimation of species trees. In the supermatrix approach, aligned sequences for all genes are concatenated and the species tree is estimated directly from a partitioned matrix. In the supertree approach, gene sequence alignments are used for inference of gene trees. Those are then combined into a species supertree. Smaller projects could benefit from utilising all available information in the supermatrix. Larger projects should rely on supertree methods for computational optimisation.

Key words

DNA sequence evolution, phylogeny, multilocus genotyping, supermatrix, supertree.

1. Introduction

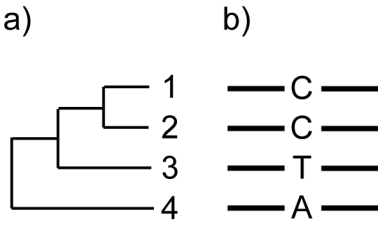
Tree of Life and similar initiatives intend to map biodiversity across life forms and to reconstruct evolutionary relationships between them (Tree of Life Web, Encyclopedia of Life, Barcoding Life, BioLib). The information in them reflects current knowledge from phylogenetic studies. Results of those, phylogenetic trees and their analyses, are in turn input directly into the phylogeny databases (TreeBASE, dryad) with direct links to their respective scientific articles. The phylogenies originate from studies that mostly utilise DNA sequence data, but frequently also morphological, karyological, behavioural or ecological markers. Nevertheless, DNA sequence data provides an easy and cheap way how to obtain abundance of reliable markers with relatively simple modes of evolution. Existence of depositories for such data (GenBank, dryad) further exacerbates their usage. The principle of usefulness of DNA sequences lies in the way in which variability forms and is maintained in DNA. The molecule consists of four basic nucleotides, adenine, thymine, cytosine and guanine, and their sequence determines the encoded genetic information. The sequences differentiate by a clearly defined exchange of one nucleotide base for another. If this change occurs in cells that would pass to the next generation, the mutation can persist and even further evolve. Other mutations that affect DNA include loss or gain of one to many nucleotides, called insertions and deletions, or indels, and rearrangement of longer sequence fragments between different positions in the genome such as changes in gene order or chromosomal rearrangements.

The reconstruction of the Tree of Life requires both large- and fine-scale resolution to achieve the goal of complete knowledge of life. I will show the options for construction of the Tree of Life based on DNA sequence data that address this.

2. Data for the evolution of life

From the perspective of phylogeny reconstruction, substitutions are the markers of choice. Indels and genomic rearrangement represent structural changes in DNA that are analysed similarly as morphological markers. That is: without the need for explicit assumptions of what is one evolutionary step (e.g. deletion of one nucleotide or the whole missing segment?) and therefore based on similarity. Substitutions have an advantage that their evolution can be modelled by accounting for multiple mutations in a single position. Comparison of sequences 1 and 4 from Figure 1, without information from all sequences 1-4, would lead to a conclusion that a single mutation, C→A, separates the two sequences. With complete knowledge of the system, we see that more closely related sequences 1 and 2 share the same nucleotide in the given position. Their sister sequence 3 is separated by one mutation, because it has a T whereas sequence 4 has an A. Thus, the nucleotide position mutated twice throughout history of this system.

Figure 1. Multiple mutations in a nucleotide base position in DNA sequences. a) Phylogeny of four taxa. b) Alignment of their DNA sequences, one base position displayed. See text for explanation



The older the studied system, the more likely is the occurrence of multiple changes in a single nucleotide position. Such multiple mutations, called multiple hits, can be modelled with a substitution model. In effect, the substitution model enlarges genetic distances for more distantly related sequences to accommodate the chance of multiple hits in certain positions in the DNA sequence if those evolved over longer time periods.

3. Multilocus phylogeny work-flow

3.1. DNA sequence alignment

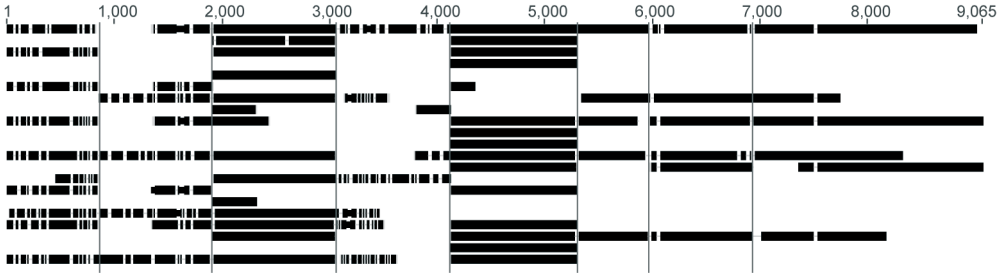
Each nucleotide position in a DNA sequence is a marker that carries information about evolution of the sequence. Sufficient variability in the dataset together with correct assessment of homology in specific positions – DNA sequence alignment – help reconstruct accurate and precise phylogenies. Ideally, the Tree of Life.

Such a situation is purely hypothetical with our current knowledge and capabilities. The reason, unrelenting to date, lies in the need to find homology of nucleotides in the DNA sequence across biodiversity. Intensive research leads to development of new methods such as alignment-free composition vector phylogenies (Xu and Hao, 2009), *k*-tuple estimation of

genetic distances (Reyes-Prieto et al., 2011) or anchor-based phylogenies (Vishnoi et al., 2010) that are independent of the homology estimation in DNA sequences. However, these methods are intended for whole-genome comparisons utilising datasets orders of magnitudes larger than those now considered ambitious that are based on DNA sequence alignment. With large amount of sequence data, the alignment-free methods can reduce the information present in the DNA sequence in a method-specific way and still retain robustness to infer a reliable phylogenetic tree. While the number of whole genome sequences rapidly increases, it is still inferior to the number of taxa with shorter segments of sequenced genomes. Alignment-based phylogenies are thus still a leading-edge approach in reconstruction of the Tree of Life.

With that comes the need for aligning sequences in such a way that each column in the resulting matrix would represent a position in the genome that shares the same evolutionary pathway with all other compared sequences and can thus be analysed as a single marker. It can be considered a discrete variable with rates of change from one state to another given by the substitution model. A gene that would be universally present in life forms and had maintained the same function since the appearance of the earliest split that survived to date most likely does not exist. Reconstruction of the Tree of Life must then rely on partial datasets for sufficiently related organisms to facilitate correct assessment of sequence homology and be based on multiple genes for different taxa sets that would enable overlap in represented taxa and combination of the information (Figure 2).

Figure 2. Alignment of eight genes of Sciurini tree squirrels (Pečnerová and Martínková, 2012). Each line represents a species, DNA sequence data is indicated with black horizontal bars, genes are separated with grey vertical lines, white-space represents missing gene segments and alignment gaps

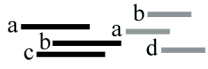
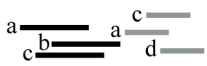
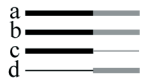
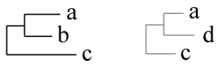
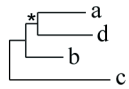


There are two approaches how to combine multilocus data from sets of sequences with overlapping taxon content (Table 1). One is a concatenation of alignments for individual genes, called a supermatrix. The other is a combination of trees reconstructed for each gene separately, called a supertree.

In the supermatrix, one creates a long sequence where each species has DNA sequence data in at least one gene and a symbol for missing data (?, -, N, depending on software requirements for subsequent analyses) at each position for all other gene segments (Figure 2). To ascertain that homologic positions with common evolutionary history are ordered in columns, DNA sequences must be aligned first for each gene separately (Table 1). Otherwise, the alignment algorithm might wrongly assume homology for sequence regions that are similar to each other through shared functional constraints on their gene products rather than shared evolutionary history. Gene alignments are then concatenated – fit back-to-back one after another with missing data symbols input for species without data for the respective genes.

The resulting matrix contains all information that is available in the dataset, albeit with great gaps. In case studies presented in this volume (Martínková and Moravec, in press; Pečnerová and Martínková, 2012), the supermatrices had 73% and 67% of missing data. In other words, as few as 27% and 33% of cells in the supermatrices contained data, which can be expressed also as coverage density as low as 0.27 and 0.33, respectively. Yet, both studies present robust results with resolution of relationships that were unclear in previous studies based on smaller datasets even if those had higher coverage densities.

Table 1. Work-flow of supermatrix and supertree reconstruction of multilocus phylogenies

Step	Supermatrix	Supertree
DNA sequence retrieval		
Gene alignments		
Estimation of substitution model	both genes separately	both genes separately
Concatenation of the gene alignments into a supermatrix		
Gene tree inference		
Estimation of species tree		

3.2. Phylogeny reconstruction

3.2.1. Sequence partitioning

After concatenation, the supermatrix is ready for a phylogenetic analysis with one important caveat. The sequence is a chimeric construct of genetic information from often very different regions of the genome. These might face different evolutionary pressures, which lead to different persistence of mutations in DNA. Mutations in protein-coding regions would pass through generations only if they were not lethal. Similar mutations in non-coding regions or in genes that occur in the genome in multiple copies would not have the same limitation. Mutations in them could pass to the next generation without marked hindrance. When viewed on a sequence, regions without strict selection against some changes would appear to have higher evolutionary rates. This is usually not a result of more frequent mutations in such regions, but of easier survival of all mutations to the next generation.

Therefore, each gene must be treated separately in a partition. Such partitions, blocks of the supermatrix, are coded prior to the analysis (Figure 2). The reason why alternative selective pressures matter in phylogeny reconstruction is that they affect rates of substitutions. Rate matrix and rate heterogeneity distribution modelling (substitution model) in turn affect likelihood of the phylogenetic tree. As the tree likelihood is the only measure used to select the best tree to represent relationships between studied taxa, its correct estimation is

important – a.k.a. correct substitution model is important. Supermatrix is therefore analysed in partitions, where each partition represents a sequence with common evolutionary rates, not necessarily a single gene.

3.2.2. Gene and species trees

In the supertree construction, gene alignments are subject to separate phylogenetic analyses first, without the concatenation of the sequences (Table 1). Each gene tree then contains relationships of only those taxa, for which DNA sequence information was available, with resolution that could be estimated for each gene separately.

Combination of gene trees depends on the used algorithm. In Table 1, the supertree shows a polychotomy – incomplete resolution of a node – with taxa *a*, *b* and *d* displaying unresolved relationships. This is frequent in supertrees and can be resolved most clearly by obtaining a gene sequence for one species in the group that is not represented in another gene tree. The black gene tree shows that sequences *a* and *b* are more closely related to each other than either is to *c*. In the grey gene tree, *a* forms a group with *d* and both are again more distant from *c* than from each other. Based on the gene trees, there is no information on relationship between *b* and *d*. Most supertree construction methods use information on topology and ignore branch lengths in gene trees (see Pečnerová and Martínková in this volume for an overview). Their result is then a cladogram, a tree that demonstrates topology and its branch lengths are meaningless.

Supermatrix analysis generates a phylogeny. Here, branch lengths represent accumulation of evolutionary change over time. Using all information in the dataset, the analysis could resolve that *a* is probably more closely related to *d* than to *b*. The node defining this and marked with an asterisk in Table 1 would be resolved in a dataset of two genes if one of the genes would be much more informative for phylogeny reconstruction than the other. For more complex datasets with more genes, the relationships become more complicated to infer, but even limited information from each partition improves resolution of the species tree.

4. Supermatrix vs. supertree

The supermatrix approach is often superior to the supertree approach in accuracy and resolution. However, it becomes computationally very intensive very quickly, because the number of possible resolved trees rapidly increases with increasing number of analysed sequences. Where relationships between the four taxa in Table 1 could be resolved in total in 15 rooted trees, the Sciurini tree squirrels (Pečnerová and Martínková, 2012) could form about 2.2×10^{20} trees and Arvicolini rodents (Martínková and Moravec, in press) 2.8×10^{126} trees. Exploring the complete tree space becomes unfeasible for more than about ten sequences. With increasing number of sequences the computational demands increase too quickly even for heuristic searches in maximum likelihood analyses or optimised Markov chains Monte Carlo in Bayesian phylogeny inference (Moravec and Martínková in this volume). Additionally, assessment of an inspected tree during the search is done with tree likelihood, which is estimated by calculating probabilities of ancestral states at all tree nodes using the substitution model and for each alignment position. This further slows the analysis as calculation of tree likelihood for very long sequences takes time that becomes non-negligible. Supertrees are vastly more advantageous in such cases. Their construction is fast and as the gene trees contain only a fraction of inspected biodiversity, they are easier to obtain. The balance for choosing between supermatrix or supertree estimation of species trees is therefore decided in the only currency that has a true value in life – time.

5. References

Martínková N, Moravec J. Multilocus phylogeny of arvicoline voles (Arvicolini, Rodentia) shows small tree terrace size. Folia Zoologica, in press.

Pečnerová P, Martínková N. 2012. Evolutionary history of tree squirrels (Rodentia, Sciurini) based on multilocus phylogeny reconstruction. Zoologica Scripta 41: 211–219.

Reyes-Prieto F, García-Chéquer AJ, Jaimes-Díaz H, Casique-Almazán J, Espinosa-Lara JM, Palma-Orozco R, Méndez-Tenorio A, Maldonado-Rodríguez R, Beattie KL. 2011. LifePrint: a novel k-tuple distance method for construction of phylogenetic trees. Advances and Applications in Bioinformatics and Chemistry 4: 13–27.

Vishnoi A, Roy R, Prasad HK, Bhattacharya A. 2010. Anchor-based whole genome phylogeny (ABWGP): A tool for inferring evolutionary relationship among closely related microorganisms. PLoS ONE 5: e14159.

Xu Z, Hao B. 2009. CVTree update: a newly designed phylogenetic study platform using composition vectors and whole genomes. Nucleic Acids Research 37(Web Server issue): W174–W178.

Multilocus phylogeny of Sciurini tree squirrels

Patrícia Pečnerová[1], Natália Martínková[2,3]

[1] *Department of Botany and Zoology, Faculty of Science, Masaryk University, Brno; e-mail: pata.pecnerova@mail.muni.cz*

[2] *Institute of Biostatistics and Analyses, Masaryk University, Brno; e-mail: martinkova@ivb.cz*

[3] *Institute of Vertebrate Biology, Academy of Sciences of the Czech Republic, v.v.i., Brno*

Abstract

Phylogenetic relationships inside the tribe Sciurini produce conflict between older morphological research and modern molecular studies. We provided a detailed phylogenetic analysis by incorporating eight loci and various methods of data processing. We used prevailing and user-friendly software packages (Geneious, BioEdit, MrBayes, ModelTest). Evolutionary history of Sciurini squirrels was examined by means of Bayesian inference of concatenated data set and six supertree construction methods. The concatenated data and superstrees generated by SuperTriplets, modified MinCut, standard MRP and *veto* supertree (without source tree correction) yielded similar results with taxa grouped according to their zoogeographic distribution. The genus *Tamiasciurus* formed a separate evolutionary lineage at the base of our trees and the other taxa gradually diverged into Palaearctic/Indomalayan, Nearctic and Neotropical groups. The other used methods, MinCut, Purvis-MRP and *veto* (with source tree correction) showed deviations from this pattern.

Key words

Phylogeny, supermatrix, supertree, Sciurini, squirrels

1. Introduction

Introducing sequences from multiple loci helps enhance accuracy of phylogenetic analyses. Wiens (1998) showed that even adding character sets with missing data alters phylogenetic accurancy. However, profits of higher data content descend with increasing ratio of missing data.

Two strategies are used for phylogenetic inference based on multilocus data - supermatrix and supertree approach. Supermatrix analysis concatenates all characters of all taxa into a single matrix, which works as a template for tree construction. Supertree analysis combines information of source trees, built independently from source data.

There are diverse preferences in usage of supermatrices and supertrees to estimate phylogeny (de Queiroz and Gatesy, 2007; Sanderson, 1998; Bininda-Emonds, 2004). Supertrees have been widely used to reconstruct large phylogenies. According to Bininda-Emonds (2004), deficiency of compatible data disabled comparable extant of supermatrix analyses. On the other hand, supertree methods are criticised for losing information during integraton of source data (de Queiroz and Gatesy, 2007). Comprehensive phylogenetic analysis was applied to the group of Sciurini tree squirrels (Pečnerová and Martínková, 2012). We analysed eight loci and sequences were processed by both approaches of phylogenetic

analysis, supermatrix and supertree. This way, we could compare outcomes of these two strategies. Six methods were used for supertree construction. Our results showed similar phylogenetic tendencies in trees generated from concatenated dataset and three methods of supertree reconstruction.

2. Materials and Methods

2.1. Source data

We used the large collection of sequences of Sciurini squirrels available in GenBank, the open-access sequence database (Benson et al., 2012). For this multilocus analysis, loci, which contained sequences for at least five different species were chosen. In total, we included eight genes in the study - four mitochondrial (12S rRNA, 16S rRNA, d-loop, *mt-cyb*) and four nuclear (*irbp*, *c-myc* exon 2, *c-myc* exon 3, *rag1*). Species composition in gene data sets was partially overlapping. Overall, our dataset consisted of two outgroup species and 19 Sciurini species of all currently recognized genera (*Microsciurus*, *Rheithrosciurus*, *Sciurus*, *Syntheosciurus*, *Tamiasciurus*).

2.2. Alignments

Sequences were aligned in Geneious software v4.7 (Biomatters Ltd., Auckland, New Zeland; Drummond et al., 2009). Alignments were executed in *fasta* format. As alignment in *nexus* format is needed in Bayesian analysis, we transformed *fasta* files to *nexus* files in BioEdit program v7.1.3 (Hall, 1999). For examples of the sequence formats see Martínková (2008).

2.3. Bayesian inference of phylogeny

Bayesian inference of phylogeny was used for estimation of phylogenetic relationships. We analysed each locus separately in MrBayes v3.1.2 (Huelsenbeck and Ronquist, 2001). Bayesian analysis in phylogeny is based on the posterior probability of a tree, which indicates the probability that the calculated tree is correct (Huelsenbeck et al., 2001).

MrBayes employs Markov chain Monte Carlo (MCMC) in determination of posterior probabilities. According to Cummings et al. (2003), MCMC can be described as "an algorithm-led trip through parameter space, where parameter space is defined in terms of topology, branch lengths, substitution rates, and other parameters". Each suggested parameter modification can be accepted or rejected depending on the change in likelihood (e.g. Moravec and Martínková, 2012).

In this study, we used the standard values of posterior probability considered significantly supported, which means values equal or higher than 0.95. Substitution models were determined according to Bayesian information criterion in ModelTest v3.7 (Posada and Crandall, 1998).

Commands for analysis in MrBayes had following form:

```
#NEXUS
begin mrbayes;
log start filename=filename.log replace;
set autoclose=yes nowarn=yes;
lset nst=6 rates=gamma;
```

```
mcmcp ngen=2000000 relburnin=yes burninfrac=0.3
samplefreq=1000 printfreq=500 nchains=5 nruns=2 temp=0.1
swapfreq=3 nswaps=1;

mcmc;

sump burnin=600;

sumt displaygeq=0.5 burnin=600;

log stop;

end;
```

Command `lset` defines the substitution model. In this case, `nst=6` indicates the GTR model with rate matrix with six parameters, and `rates=gamma` means that substitution rates vary between positions in the DNA sequence and the variation is modelled with a Γ distribution.

Command `mcmcp` adjusts parameters of the MCMC. To optimize chain convergence, we ran five Markov chains Monte Carlo (`nchains=5`) for 2 million generations (`ngen=2000000`), sampling trees every 1000th generation (`samplefreq=1000`). Chain heating parameter was set to 0.1 (`temp=0.1`), one chain swap was attempted every 3rd generation (`swapfreq=3`). Relative burn-in was used (`relburnin=yes`) and the burn-in fraction was 30% (`burninfrac=0.3`).

2.4. Supermatrix approach

Supermatrix analysis (also known as total evidence or combined analysis) is based on a matrix, consisting of all character data from all included taxa. All characters are then analysed simultaneously (de Queiroz and Gatesy, 2007). The advantage of the supermatrix approach is the preservation of data and utilization of all evidence. In comparison, in the supertree analysis some part of the character information is lost through combining data sets (Moravec and Martínková, 2012).

2.5. Supertree approach

Supertrees result as a conjunction of source trees. Gene trees are estimated independently and consequently the information from these gene trees is combined by various methods of supertree reconstruction. Supertree approach facilitated construction of large phylogenies and research of evolutionary history for higher taxonomic units.

Table 1. Methods of supertree construction used in this study

Method	Program	Program reference
Standard MRP	r8s	Sanderson, 2003
	PAUP*	Swofford, 2002
Purvis-MRP	r8s	Sanderson, 2003
	PAUP*	Swofford, 2003
MinCut	Supertree	Page, 2002
Modified MinCut	Supertree	Page, 2002
SuperTriplets	SuperTriplets	Ranwez et al., 2010
Veto	PhySIC_IST	Scornavacca et al., 2008

We used six methods of supertree reconstruction: standard matrix representation with parsimony (MRP; Baum, 1992; Ragan, 1992), matrix representation with parsimony modified by Purvis (Purvis-MRP; Purvis, 1995), MinCut (Semple and Steel, 2000), modified MinCut (Page, 2002), SuperTriplets (Ranwez et al., 2010) and *veto* supertree method (Scornavacca et al., 2008; Table 1).

2.5.1. Standard matrix representation with parsimony and Purvis modification

Standard MRP and Purvis-MRP are methods based on translating phylogenetic trees to a binary matrix. According to Baum and Ragan (2004) each tree is considered as "a hierarchically ordered collection of nodes". Information about nodes on a source tree is expressed in the form of additive binary coding. Accordingly, nodes are treated as binary matrix elements and the tree topology is represented by the matrix. Matrices of all source trees are integrated to form a single matrix of binary elements (Baum and Ragan, 2004). The composite matrix is subsequently analyzed with parsimony.

Modification of MRP by Purvis (1995) is aimed at elimination of redundant data, which is brought into analysis by additive binary coding. Purvis suggested adapted coding, which scores all taxa beyond sister relationship to examined node as missing, instead of 0 (Bininda-Emonds and Sanderson, 2001).

In this work, we used program r8s v1.70 (Sanderson, 2003) for matrix construction and PAUP* v4b10 (Sinauer Associates, Inc., Sunderland, MA; Swofford, 2002) for maximum parsimony analysis.

Data for r8s and PAUP* analysis should be in *nexus* format.

Data entry for r8s was in following form:

```
#nexus
begin trees;
tree 1 = (A:0.104,B:0.043,...);
tree 2 = (A:0.176,D:0.087,...);
tree 3 = (A:0.042,B:0.003,...);
end;

begin r8s;
mrp method=baum;
mrp method=purvis;
end;
```

Data entry for PAUP* was in following form:

```
#nexus
begin paup;
set criterion=parsimony increase=no maxtrees=10000;
hsearch nreps=10 addseq=random swap=tbr;
```

```
contree all/ majrule=yes percent=50 treefile=filename.con;
end;
```

We adjusted paramters to 10 heuristic search replicates (`nreps=10`), TBR branch swapping algorithm (`swap=tbr`), 10000 swapped trees as maximum number (`maxtrees=10000`) and final tree constructed as 50% majority rule consensus (`majrule=yes percent=50`).

2.5.2. MinCut and modified MinCut

These methods translate source trees into a graph. In the graph, all branches have weights, which are determined according to the number of their appearances in source trees (Page, 2002). Consequently, branches with too low weight are eliminated from the graph. This elimination is termed as minimum cut, as it cuts branches with minimum weights. The supertree is formed from the graph after the minimum cut. MinCut is one of the methods that run in polynomial time, hence it is quick to compute even for very large datasets.

Main point of modified MinCut method (Page, 2002) is to prevent the uncontradicted branches from source data to be cut. This adaptation also helps to solve the influence of size of source trees.

MinCut and modified MinCut analysis were performed in the Supertree software (Page, 2002).

2.5.3. SuperTriplets

SuperTriplets (Ranwez et al., 2010) is another method running in polynomial time. This algorithm calculates the triplet median supertree. Analysis starts with partitioning the source trees into simple triplets. Triplets have weights depending on the number of source trees including it. Weights help assess which triplet is the most frequent among source trees. Another step is agglomerative procedure generating the initial supertree. SuperTriplets then utilises swapping subtrees to check if any other supertree represents source trees better until any better tree cannot be found (Ranwez et al., 2010).

We used SuperTriplets program available online (Ranwez et al., 2010).

2.5.4. Veto method

Other types of supertree construction methods are *veto* methods. These are based on a simple condition that the supertree has to be consistent with all source trees. In other words, supertree cannot contain a clade contradicted by any of the source trees. As a result, *veto* supertrees tend to be unresolved, with higher number of multifurcations. The supertree is constructed in a stepwise manner by gradually adding leaves to the initial tree of two nodes (Brinkmeyer et al., 2010).

We constructed *veto* supertree in PhySIC_IST (Scornavacca et al., 2008). In comparison with the original PhySIC program, this modification tries to reduce the abundance of unresolved relationships by elimination of conflicting taxa. Taxa producing conflicts during supertree construction are discarded to avoid excessively unresolved phylogenies. PhySIC_IST includes an additional tool, source tree correction, which edits source trees in conflicting regions before analysis. We applied the version with and without source tree correction.

3. Case study

Sciurini squirrels represent tree squirrels inhabiting forests of Eurasia, North America and South America. These sciuromorph rodents constitute 37 species distributed into five genera: *Microsciurus, Rheithrosciurus, Sciurus, Syntheosciurus* and *Tamiasciurus* (Wilson and Reeder, 2005). In Eurasia, only four of the taxa can be found. The highest diversity rates are in Central and South America, which Sciurini entered only three million years ago (Mercer and Roth, 2003).

Origin of the tribe Sciurini is in the Northern Hemisphere according to the fossil record (Emry and Thorington, 1982; Emry et al., 2005) but it is still unclear if Sciurini squirrels diverged in Eurasia or North America. Oshida et al. (2009) proposed that the genus *Sciurus* originated in Eurasia based on a phylogenetic analysis of eight Sciurini taxa.

The presented study utilised all recently available genetic data to estimate phylogenetic relationships inside the Sciurini group (Pečnerová and Martínková, 2012). We applied two alternative approaches, supermatrix and supertree approach. By means of such a complex analysis, we confirmed the paraphyly of *Sciurus* and disagreement between current taxonomy and phylogeny. At the same time, resolved phylogenetic history helped us better understand the evolutionary pathways of Sciurini tree squirrels.

3.1. Gene trees

In total, 9065 base pairs (bp) of eight loci and 19 species were analyzed in this study. The gene trees contained at least five species, and the largest dataset had DNA sequences for 15 species (Pečnerová and Martínková, 2012).

Most of the gene trees shared a single pattern with Sciurini split into two evolutionary lineages, first containing all *Tamiasciurus* species and the second including representatives of all other genera. Inside the latter group, the Old World taxa occupied basal positions and the New World clade was monophyletic according to five gene trees. Relationships inside of the New World group were poorly resolved, but two trees supported monophyly of Neotropical taxa (Pečnerová and Martínková, 2012).

3.2. Supermatrix results

Bayesian analysis of the concatenated data set yielded a tree with similar trends as described for gene trees.

3.3. Supertree results

Despite deviations in supertrees generated by different methods of supertree construction, there were the same tendencies (Pečnerová and Martínková, 2012). *Tamiasciurus* at the base of the tree, followed by a group of Palearctic/Indomalayan taxa and the monophylum of Nearctic and Neotropical taxa. Differences between supertrees consisted of a diverse placement of species inside the Palearctic/Indomalayan cluster and the Nearctic/Neotropical cluster. Supertrees produced by SuperTriplets, modified MinCut, standard MRP and *veto* without source tree correction were the most consistent with this pattern and most resembled each other. Due to the character of *veto* method, some species were excluded from the *veto* supertrees.

3.4. Relationships of Sciurini tree squirrels

Both approaches of phylogenetic analysis revealed similar tendencies, corresponding with the biogeographic distribution of Sciurini squirrels (Pečnerová and Martínková, 2012; Figure

1). Taxa were grouped according to zoogeographic regions they inhabit, with gradual divergence of species from Eurasia, North and Central America and South America. In terms of their current taxonomy, the genus *Tamiasciurus* was located at the base of tree at supermatrix and supertree approaches. Genera *Rheithrosciurus*, *Microsciurus* and *Syntheosciurus* grouped with species from the genus *Sciurus*.

Figure 1. Schematic representation of Sciurini tree squirrel phylogeny with branch tips located at approximate centroids of geographic distribution of the species

Interestingly, *Sciurus* spread through North America to tropical regions of Central and South America, where it diversified into many species, generating the peak of its diversity in the south. Sciurini tree squirrels entered South America only three million years ago. This high diversity in such recently formed group might be produced by high diversification rate as proposed by Roth and Mercer (2008).

Herewith, we demonstrated that both supermatrix and supertree approaches are consistent in assessing phylogenetic relationships from datasets with large content of missing data with good resolution. The supertrees that differed from other results use methods that are in general intended for removing conflict in source data. In our data, this paradoxically introduced patterns in the trees that were unique for each method. This might indicate that for datasets with roughly consistent signal in gene trees, utilising maximum amount of information and thorough analysis might be the optimal approach.

4. References

Baum BR. 1992. Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. Taxon 41: 3–10.

Baum BR, Ragan MA. 2004. The MRP method. In: Bininda-Emonds ORP (ed.) Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life. Dordrecht: Kluwer Academic, 2004. 564 p. ISBN: 978-1-4020-2328-6.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2010. GenBank. Nucleic Acids Research 38: D46–51.

Bininda-Emonds ORP, Sanderson M. 2001. Assessment of the accuracy of matrix representation with parsimony analysis supertree construction. Systematic Biology 50: 565–579.

Bininda-Emonds ORP. 2004. The evolution of supertrees. Trends in Ecololgy and Evolution 19: 315–322.

Brinkmeyer M, Griebel T, Böcker S. 2010. Polynomial supertree methods revisited. In: Dijkstra TMH, Tsivtsivadze E, Marchiori E, Heskes T (eds.) Proceedings of the 5th IAPR international conference on Pattern recognition in bioinformatics. Berlin: Springer, 2010. 442 p. ISBN 978-3-642-16000-4.

Cummings MP, Handley SA, Myers DS, Reed DL, Rokas A, Winka K. 2003. Comparing bootstrap and posterior probability values in the four-taxon case. Systematic Biology 52: 477–487.

Drummond AJ, Ashton B, Cheung M, Heled J, Kearse M, Moir R, Stones-Havas S, Thierer T, Wilson A. 2009. Geneious v4.7. Available from http://www.geneious.com/.

Emry RJ, Thorington Jr., RW. 1982. Descriptive and comparative osteology of the oldest fossil squirrel, Protosciurus (Rodentia, Sciuridae). Smithsonian Contributions to Paleobiology 47: 1–35.

Emry RJ, Korth WW, Bell MA. 2005. A tree squirrel (Rodentia, Sciuridae, Sciurini) from the Late Miocene (Clarendonian) of Nevada. Journal of Vertebrate Paleontology 25: 228–235.

Hall TA. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symposium Series 41: 95–98.

Huelsenbeck JP, Ronquist F. 2001. MrBayes: Bayesian inference of phylogeny. Bioinformatics 17: 754–755.

Huelsenbeck JP, Ronquist F, Nielsen R, Bollback JP. 2001. Bayesian Inference of Phylogeny and Its Impact on Evolutionary Biology. Science 294: 2310–2314.

Martínková N. 2008. Tutorial in phylogenetic analyses. In: Dušek L, Haruštiaková D, Martínková N. (eds.) Proceedings of the 4th International Summer School on Computational Biology: Statistical Methods for Genetic and Molecular Data. Brno: Masaryk University, 2008. 122 p. ISBN 978-80-210-4793-8.

Mercer JM, Roth VL. 2003. The effects of Cenozoic global change on squirrel phylogeny. Science 299: 1568–1572.

Moravec J, Martínková N. 2012. Reconstructing phylogeny from patch data in rodents. In: Proceedings of the 8th Summer School on Computational Biology. Brno: Akademické nakladatelství CERM, 2012.

Oshida T, Arslan A, Noda M. 2009. Phylogenetic relationships among the Old World Sciurus squirrels. Folia Zoologica 58: 14–25.

Page RDM. 2002. Modified mincut supertrees. In: Guigó R, Gusfield D. (eds.) Proceedings of the Second International Workshop on Algorithms in Bioinformatics, WABI 2002, LNCS 2452. Berlin: Springer, 2002. 564 p. ISBN 978-3540442110.

Posada D, Crandall KA. 1998. Modeltest: testing the model of DNA substitution. Bioinformatics 9: 817–818.

Pečnerová P, Martínková N. 2012. Evolutionary history of tree squirrels (Rodentia, Sciurini) based on multilocus phylogeny reconstruction. Zoologica Scripta 41: 211–219.

Purvis A. 1995: A modification to Baum and Ragan's method for combining phylogenetic trees. Systematic Biology 44: 251–255.

de Queiroz A, Gatesy J. 2007. The supermatrix approach to systematics. Trends in Ecology and Evolution 22: 34–41.

Ragan MA. 1992 Phylogenetic inference based on matrix representation of trees. Molecular Phylogenetics and Evolution 1: 53–58.

Ranwez V, Criscuolo A, Douzery EJP. 2010. Supertriplets: A triplet-based supertree approach to phylogenomics. Bioinformatics 26: i115–i123.

Roth VL, Mercer JM. 2008. Differing rates of macroevolutionary diversification in arboreal squirrels. Current Science 95: 857–861.

Sanderson MJ. 1998. Phylogenetic supertrees: assembling the trees of life. Trends in Ecology and Evolution 13: 105–109.

Sanderson MJ. 2003. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. Bioinformatics 19: 301–302.

Semple C, Steel M. 2000. A supertree method for rooted trees. Discrete Applied Mathematics 105: 147–158.

Scornavacca C, Berry V, Lefort V, Douzery EJP, Ranwez V. 2008: PhySIC_IST: cleaning source trees to infer more informative supertrees. BMC Bioinformatics 9: 413.

Swofford DL. 2002. PAUP*: Phylogenetic Analysis Using Parsimony (and Other Methods), Version 4.0b10. Sinauer, Sunderland, Massachusetts.

Wiens JJ. 1998. Does adding characters with missing data increase or decrease phylogenetic accuracy? Systematic Biololgy 47: 625–640.

Wilson DE, Reeder DM. (eds.). Mammal Species of the World. A Taxonomic and Geographic Reference (3rd ed). Baltimore: John Hopkins University Press, 2005. 2000 p. ISBN: 978-0-8018-8221-0.

Reconstructing phylogeny from patchy data of rodents

Jiří Moravec[1], Natália Martínková[1,2]

[1] *Masaryk Memorial Cancer Institute, Brno; e-mail: smith@gmail.com*
[2] *Department of Geology, Faculty of Science, Charles University in Prague; e-mail: newman@natur.cuni.cz*

Abstract

To reveal phylogeny of sparsely sequenced taxa, standard methods could not be successfully used due to patchy character of data and new methods had to be developed. We summarize such methods and present their functionality on phylogeny of Arvicolini voles. Analyzing tree space with terraces, we have found that supermatrix approach is superior to supertree approach in extracting signal from data and determining a resolved and well-supported phylogeny. The most widely used program for Bayesian phylogeny inference fails to determine the correct lengths of branches in a large supermatrix with a lot of missing data, it still successfully determines the true tree topology.

Key words

Phylogeny, supermatrix, supertree, arvicolini, Bayesian inference

1. Introduction

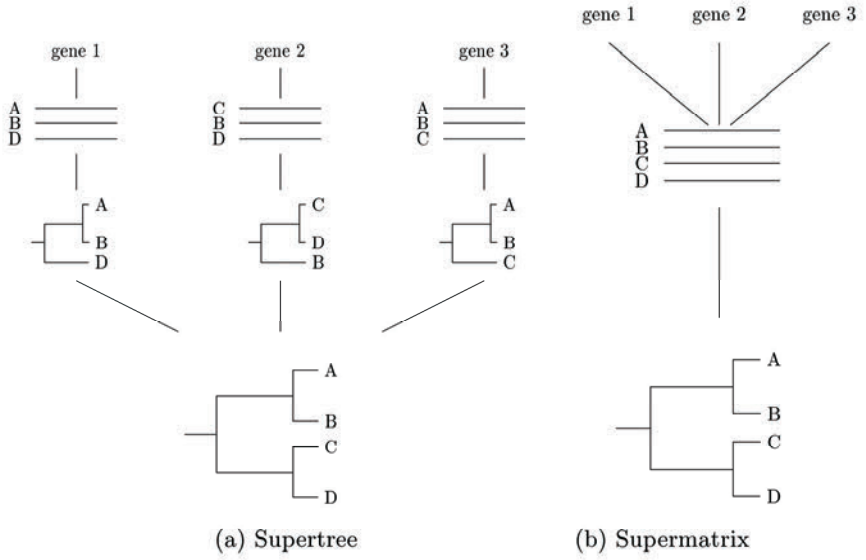
DNA sequence analysis greatly advanced with founding and world-wide usage of sequence databases GenBank, EBI and DDBJ. With free access to sequences, many new analyses were possible on a much greater scale. This brought new type of an incompleteness problem. Sequences used now more likely originated from different studies with different aims, with few genes sequenced for some species and many for others. Standard phylogenetic analysis of single a gene could either use only a subset of available data or it could not produce a resolved tree. One way to solve this problem is to obtain new sequences - this is an expensive and laborious approach might be even impossible sometimes in some cases. On top of that, genes are functional units that are adapted to serve a specific function. To maintain the function of gene products under similar selective pressure, gene products would be more similar to each other and the genes converge. Thus, their phylogeny may not match the true tree and analyses of multiple genes are necessary. New data obtaining methods were developed. Through usage of those methods, new research topics could be addressed despite seemingly bare data matrix. Hereby, we present a multilocus phylogeny of Arvicolini voles inferred from a data matrix of eight genes that contained 72.8% missing data (Martínková and Moravec, in press).

2. Building trees from multiple loci

Generally, there are two ways how to build a tree from multiple loci, using a supermatrix (de Queiroz and Gatesy, 2007) or a supertree (Bininda-Emonds, 2004). Supertree approach is a type of metaanalysis. Trees are constructed separately into gene trees compared and assembled into the species tree. It does not work directly with characters, but with whole trees. This makes the method ideal for building supertrees from published trees, be it either

trees built from molecular data, trees built from morphological data or combined trees and other supertrees.

Figure 1. Diagram of Supertree and Supermatrix approaches work-flow.



The alternative supermatrix approach directly works with characters from all loci. Sequences are aligned separately for each locus and their alignment is concatenated, creating a large data matrix on which standard phylogeny methods can be applied. Although computationally intensive, this approach can amplify weak signal from multiple loci to solve complicated relationships.

The supermatrix approach has several advantages over the supertree approach. In a supertree, some information is lost, when data are summarized into gene trees. Information from each tree has same weight, although original data might not support this. Estimating weight of each gene tree *ad-hoc* could then be inaccurate. Serious disadvantage of supermatrix analysis lies in its high computational difficulty. Data consisting of several hundreds of species and dozens of genes would require enormous computational capacity to process. Future of such giant projects, such a Tree of Life project (http://tolweb.org), thus consists in combining those two methods. With divide and conquer approach (Bininda-Emonds, 2010), giant datasets are partitioned into several smaller supermatrixes, each consisting of multiple loci, and, computed trees are then summarized with the supertree approach, exploiting supermatrix precision and supertree lower computational needs.

3. Computing from supermatrix

The most robust methods of computing phylogenies from the supermatrix are Maximum Likelihood (ML) and Bayesian inference (BI). Although they are mathematically quite similar, the likelihood function is a component of the Bayes theorem, their implementations differ greatly.

In summary, ML algorithm starts with a starting tree, be it random, user input or computed by fast and dirty method, and searches similar topologies by exchanging subtrees and maximizing tree likelihood with respect to nuisance parameters, such as branch length. The best tree is then the tree with the largest likelihood (Schmidt and Haeseler 2009).

The BI computes probability of a tree, a scenario, a hypothesis or values of parameters directly with the Bayes theorem (Ronquist et al., 2009). Let τ be a vector of parameters (topology, branch lengths, substitution rates etc.) defining the tree under a certain model *M*, *T* be tree/parameter space and *D* data. Posterior probability of the tree τ given data *D* given model *D* is calculated as:

$$\overbrace{f(\tau|D,M)}^{posterior} = \frac{\overbrace{f(D|\tau,M)}^{likelihood}\overbrace{f(\tau|M)}^{prior}}{\int_T f(D|\tau,M)f(\tau|M)d\tau} \quad (1)$$

Because the parameter space is continuous, it is sampled by Markov Chain Monte Carlo (MCMC), a Metropolis-Hasting variant (Ronquist et al., 2009), which is sometimes improved with Metropolis-coupling (forming MCMCMC or MC^3) (Huelsenbeck and Ronquist, 2003). MCMC searches the parameter space, making small steps and accepting them if the new position has higher probability than the current one, otherwise it remain in place. The algorithm could be described as follows:

1. Start with a random tree τ_i
2. Draw new a state τ_j from arbitrary chosen symmetric proposal distribution $Q(\tau_j|\ \tau_i)$.
3. Probability *R* of accepting new the state is:

$$R = \min\left[\frac{f(D|\tau_j)}{f(D|\tau_i)} \times \frac{f(\tau_j)}{f(\tau_i)} \times \frac{Q(\tau_i|\tau_j)}{Q(\tau_j|\tau_i)}\right] \quad (2)$$

1. Generate a uniformly distributed random variable *U* from interval (0,1). If $R > U$, accept new the state.
2. Repeat from the step 2.

Each accepted state becomes a sample. After a sufficient number of samples, the sample distribution should be a good estimate of the sampled distribution. Metropolis-coupling improves computational speed and lowers risk, that the chain will stay in a local optimum, by running several MCMC with different probability of accepting the new state, called temperature, swapping it each *n* states between chains. Chains with higher temperature would accept the new state even if it is worse than the current state, which gives the chain ability to escape local optima and to search the parameter space quickly. Colder chains are more likely to stay and search local space for potential global optimum. Several runs are usually made to ensure that the stationary distribution was reached. Worth mentioning, that in complicated models with many parameters, BI is highly computationally intensive.

Apart from a likelihood, which comes directly from data, the Bayes theorem includes another important component, a prior. Prior is information about a set of hypotheses or values of parameters we have before analyzing data. It is represented with probability density function for a continuous prior or probability mass function for a discrete priors and sampled with MCMC. In BI uninformative priors are used, so likelihood will easily dominate them.

In the past few years, when analyses of large supermatrix problems became more common, a new problem emerged. The default exponential prior used in most widely used BI phylogeny program MrBayes was shown to be too informative. Additionally, the default starting branch length is set to be 0.1, which causes MCMC to spend too much time in prior-elevated local optima. This results in longer branches across the whole tree. Fortunately, even in analyses with branch lengths problems, the BI usually infers the true topology - the long branch problem (Marshall, 2010). It was suggested, that partitioning branch lengths between external branches (carrying tree leaves - taxa) and internal branches should solve this problem (Yang and Rannala, 2005) and a new compound GammaDirichlet prior was suggested (Rannala et al., 2012) and implemented (Zhang et al., 2012). Prior serves as important optimizing factor and new priors are constructed to better demonstrate our ignorance about current dataset but hold enough information about very process of origin of data, molecular evolution.

4. Assessing tree support

One of the major questions after using an analytic method is: *How good my results are?* The most popular method for distance, maximum parsimony and ML methods is the bootstrap.

For BI the support for bipartitions is given by their posterior probability.

Bootstrapping relies on analyses with resampled data. The data are resampled by randomly picking columns from the data matrix allowing repetition, until the same length as original data matrix is reached. The bootstrap trees are constructed from the resampled data matrices. Bipartitions of the original tree are then evaluated with the percentage of the bootstrap trees that contain bipartitions of the same clade, creating the bootstrap score. Bootstrap score over 70% is considered significant. Original claim for 95%, assessed from confidence interval (Felsenstein, 1985), proved to be exaggerated because of conservative character of bootstrapping (Hillis and Bull, 1993). Three interpretations of bootstrap are possible: *type I. error*, *repeatability* and *accuracy*; with *accuracy* being the most popular (Soltis and Soltis, 2003).

In BI with posterior joint probability already sampled, posterior probability of each bipartition can be computed as proportion of sampled trees that have the same bipartition as optimal tree (Ronquist and Mark, 2009).

Note that the Bayesian posterior probability is not comparable between different trees or analyses, as it is conditional on data and model (equation 1). The same is true for the bootstrap values, although the model and data condition is not explicitly stated.

There is another way of assessing robustness of analysis that was developed for multilocus analyses with large amount of missing data. Let's have an optimal, fully resolved tree topology from ML or BI analysis of multiple loci. Then, restricting this tree to taxa from each locus we get a subset tree defined by this locus. All trees generated by subset trees form terrace (Sanderson et al., 2011) and have the same ML or BI score as optimal tree. If the best

tree belongs to a small terrace, to a small group of trees with identical optimality score, or is even unique, data provided significant signal and analysis was able to identify it.

5. Case study

Arvicoline voles (Rodentia, Arvicolini) represent a group of rodents that is often used as a model for molecular genetic studies (e.g. Martínková and Moravec, in press). Extensive DNA sequence data is available for multiple genes and different taxa sets. Arvicoline voles first appeared in the fossil record in Pleistocen boundary, and over along the last 2 millions years, they diverged into one of the most speciose mammalian group. Their morphological similarity even for distantly related species and rapid karyotype rearrangement hinder classification. The phylogeny reconstructed from a single gene sequence dataset would be limited by the very rapid diversification that could be represented as incomplete lineage sorting or lack of resolution due to the short time. Multilocus phylogeny is necessary. Patchy character of sequenced species and genes makes them an ideal example for presented methods.

5.1. Material and Methods

Sequences of eight loci for a total of 74 species of voles from the Arvicolini tribe were downloaded from GenBank, and sequences for each locus were aligned using Geneious 5.4 (Drummond et al. 2011). Alignments were concatenated into a supermatrix containing 72.8% missing data. Supermatrix was analyzed with BI implemented in MrBayes 3.1 (Ronquist and Huelsenbeck, 2003) and with ML implemented in RAxML 7.2 (Stamatakis, 2006) Sinvididual gene alignments were analyzed with BI, and the computed trees were combined into supertree with the SuperTriplets package (Ranwez et al., 2010). Robustness of analyses was estimated with terraces using the PhyloTerrace package (Sanderson et al., 2011).

5.2. Results

Inferred supermatrix phylogeny was well supported for BI and ML analyses. Both concurred on similar tree topologies. Significantly, phylogeny was better resolved than in previous studies, and we were able to estimate phylogenetic relationshops of undersampled taxa. Further, the terrace analysis of the BI tree topology placed the optimal tree on a unique terrace.

The SuperTriplet supertree agreed with the BI and ML analyses on basic relationship. Its terrace analysis found 15 alternative topologies with the same tree likelihood. This is still a very small terrace. Small terraces are retrieved probably due to the fact that cytochrome *b* (*cyb*) was present for 68 species and served as a scaffolding.

Long branch problem manifested here as well, with BI credible interval of branch length being 7.89-15.12 and ML estimate equal to 3.86. We have tried to optimize the analysis with modified MrBayes with compound GammaDirichlet prior (Zang and Rannala, 2012), but stationary distribution was not reached.

5.3 Conclusion

We have described methods for the inference of large phylogeny problems with high amount of missing data and demonstrated their utilization on phylogeny of Arvicolini voles. Understanding the presented methods is crucial for optimizing their parameters and achieving sensible results. Although emerging negative attributes of complicated analyses designs could not be predicted at the moment, common practice of using several different

methods and simulation studies should reduce these problems and enable verification of results.

6. References

Bininda-Emonds ORP. 2004. The evolution of supertrees. Trends in Ecology and Evolution 19: 315–322.

Bininda-Emonds ORP. 2010. The future of supetrees, bridging gap with supermatrices. Paleodiversity 3: 99–106.

Drummond AJ, Ashton B, Buxton S, Cheung M, Cooper A, Duran C, Field M, Heled J, Kearse M, Markowitz S, Moir R, Stones-Havas S, Sturrock S, Thierer T, Wilson A. 2011. Geneious v5.4, Available from http://www.geneious.com/

Felsenstein J. 1985. Confidence Limits on Phylogenies: An Approach Using the Bootstrap. Evolution 39: 783–791.

Hillis DM, Bul JJ. 1993. An empirical test of bootstraping as a method for assessing confidence in phylogenetic analysis. Systematic Biology 42: 182–192.

Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17: 754–755.

Marshall DC. 2010. Cryptic Failure of Partitioned Bayesian Phylogenetic Analyses: Lost in the Land of Long Trees. Systematic Biology 59: 108–117.

Martínková N, Moravec J. Multilocus phylogeny of arvicoline voles (Arvicolini, Rodentia) shows small tree terrace size. Folia Zoologica, in press.

de Queiroz A, Gatesy J. 2007. The supermatrix approach to systematics. Trends in Ecology and Evolution, 22: 34–41.

Rannala B, Zhu T, Yang Z. 2012. Tail Paradox, Partial Identifiability, and Influential Priors in Bayesian Branch Length Inference. Molecular Biology and Evolution 29: 325–335.

Ranwez V, Criscuolo A, Douzery EJP. 2010. Supertriplets: a triplet-based supertree approach to phylogenomics. Bioinformatics 26: 115–123.

Ronquist F, van der Mark P., Huelsenbeck JP. 2009. Bayesian phylogenetic analysis using MRBAYES. In: Lemey P., Salemi M., Vandamme AM. (ed.) The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing. New York: Cambridge University Press 723 p. ISBN 978-0521877107.

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19: 1572–1574.

Sanderson MJ, McMahon MM, Steel M. 2011. Terraces in Phylogenetic Tree Space. Science 333: 448–450.

Schmidt HA., von Haeseler A. 2009. Phylogenetic inference using maximum likelihood methods. In: Lemey P., Salemi M., Vandamme AM. (ed.) The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing. New York: Cambridge University Press 723 p. ISBN 978-0521877107.

Soltis PS, Soltis DE. 2003. Applying the Bootstrap in Phylogeny Reconstruction. Statistical Science 18: 256–267.

Stamatakis A. 2006. RaxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22: 2688–2690.

Yang Z, Rannala B. 2005. Branch-Length Prior Influence Bayesian Posterior Probability of Phylogeny. Systematic Biology 54: 455–470.

Zhang C, Rannala B, Yang Z. 2012. Robustness of Coumpound Dirichlet Priors for Bayesian Inference of Branch Lengths. Systematic Biology, eprint.

From Analysis of Genomic Data to Clinical Applications — Case Studies

Computational Biology Students' Abstracts

Genetic association studies

Lucie Brožová

Faculty of Science, Masaryk University, Brno

Abstract

The goal of genetic association studies is to indentify DNA variants which contribute to risk of disease. In my work I focus on population association studies in which unrelated individuals of case-control disease state are typed at a number of single nucleotide polymorphisms (SNPs). The aim of this study was to create the review of statistical analyses testing association based on single SNP and consequently move on to the tests of multiple SNP. I applied studied methods to test association between SNPs and result of direct microscopy (DM) in Murgese horses that detect presence of disease-causing parasites. Analysis results point at SNPs that could cause positive result of DM. Unfortunately, statistical significance was not confirmed for most of them.

Key words

Single nucleotide polymorphism, case-control design, association, interaction

1. Introduction

Genetic association studies aim to detect an association between polymorphisms and a trait. I concern over SNP as a type of polymorphism, and case-control state as a kind of trait. The cases have studied disease and the controls are independently sampled from the general population without disease. SNP is defined as a single base pair change in the DNA sequence with frequency of at least 5% within population. The advantage of genetic association studies is that it can detect genetic variants with small effect on disease outcome and thus it enables to understand the etiology of complex diseases (Risch and Merikangas, 1996).

In practical part of my work I tested association between SNP and positive (case) or negative (control) result of DM in Murgese horses that detect presence of parasites Theileria equi and Babesia caballi. These parasites damage blood cells and cause disease piroplasmosis (Homer et al., 2000). Clinical signs of the disease are highly variable.

2. Methods

2.1. Data quality control

I evaluated the validity of Hardy-Weinberg equilibrium (HWE) for horses with negative result of DM. Deviations from HWE in the population can indicate inbreeding, stratification and error in genotyping for example (Balding, 2006; Lewis, 2002). I counted linkage disequilibrium (LD) which represents correlation between SNPs. High correlation could present a problem for multivariate analysis when preferring SNP, which is in LD with the causal one.

2.2. Tests of association: single SNP

Testing whether there is an association between genotypes and result of DM is equivalent to test the dependence in contingency table (Balding, 2006). For this testing I applied Fisher's

exact test which detects all departures from the independence. It is reasonable to assume that horses having genotype of two copies of at-risk alleles are more likely to be positive in DM than horses having genotype with one or none at-risk allele. Cochran-Armitage trend test (CATT) tests this trend in relative risks (Armitage, 1955). It assigns score to genotype which can be changed to correspond to different genetic models. Since I do not know the true genetic model of inheritance, I prefer robust tests. I applied MAX3 accounting p-value for maximum of test statistics CATT for recessive, additive and dominant models (Freidlin et al., 2002). In case of complex diseases, SNPs contribute to disease roughly additive which is why I applied CATT for additive model (Balding, 2006). These tests detect marginal effect of SNP on result of DM. I visualized the results of tests by Q-Q graph of p-values with 95% confidence interval (CI). If we observe p-values in this CI, we can say they come from uniform distribution, and thus we cannot reject the null hypothesis of no association (Quesenberry and Hales, 1980). Correction for multiple testing was provided by Fisher's combination and Bonferroni correction. For the SNPs with detected influence on the result of DM I analyzed dominant, recessive and overdominant model of inheritance for the minor allele. I summed genotype counts in the contingency table and applied Fisher's exact test.

2.3. Tests of association: multiple SNPs

SNPs can influence the result of DM only in combination with genotype in different loci. In the next step I applied two non-parametric methods who test interaction between SNPs.

2.3.1. "Set association"

"Set association" combines information from multiple SNPs but rely on marginal effect of SNPs (Ott and Hoh, 2003). This method creates sums from the most extreme test statistics CATT with score for the additive model of inheritance. Based on permutations p-values for these sums are accounted. As a number of SNP in sum increase, the p-values for sums tend to decrease. When we add to the sum SNPs that do not influence result of DM, p-values tend to increase. The global p-value is accounted for minimum of p-values.

2.3.2. Multifactor dimensionality reduction (MDR)

MDR is based on 10-fold cross validation, which divides the data into 10 equal parts (Moore, 2004). 9/10 of the data is used to develop a model (combination of SNPs) and the rest 1/10 of the data is used to evaluate its predictive ability. MDR assigns each combination of genotype to high risk or low risk group and that is how it reduces the multifactor dimensionality. High risk is assigned when the ratio of horses with positive and negative DM exceeds the total ratio. This new variable is evaluated for its ability to classify (training accuracy) and predict result (testing accuracy) of DM. The software just evaluates specific model on the basis of balanced accuracy, which represents mean of sensitivity and specificity. To minimize the impact of the current distribution of data I applied MDR ten times and averaged out the results.

3. Results

Initially, I excluded one SNP because of low allele frequency. I checked validity of HWE for horses with negative result of DM and excluded high LD between studied SNP. Analysis was performed on the complete dataset and on the dataset with horses older than two years. This limit was based on ROC curve analysis and aimed to exclude the horses who had not met with parasite yet and might be affected by maternal antibodies. For tests on single SNP I did not detect statistically significant results so I evaluated SNPs with p-value less than 0.05 as

interesting. Table 1 summarizes results of the analysis. “Set association” detected minimum of p-value for the second sum into which contribute test statistics of MAFK BseRI and IL12p40. In case of dataset for horses older than two years, the method had detected the smallest p-value for the first sum, TLR2 BsaHI. Statistical significance was not confirmed. MDR detected interaction between MAFK BseRI, TLR3 (HpyCH4III) and IL12p40 for complete dataset, and TLR2 BsaHI and TLR3 (HpyCH4III) for the dataset with horses older than two years. There are only statistically significant results of testing genetic model of inheritance in the last column of table 1.

Table 1. SNPs with detected effect on result of DM

	Fisher’s exact test/ CATT/ MAX3	„Set association“	MDR	Model of inheritance
Complete dataset (N=101)				
PRKAR1B BmgBI	✓			Recessive
MAFK BseRI	✓	✓	✓	Dominant
TLR3 (HpyCH4III)			✓	
IL12p40	✓	✓	✓	
Dataset for horses older than two years (N=76)				
TLR2 BsaHI	✓	✓	✓	
TLR3 (HpyCH4III)			✓	

4. Conclusion

For testing of single SNP effect, I cannot tell which one of tests performs better. Results of Fisher’s exact test are consistent with results of MAX3. We can see from Table 1 effect of TLR3 (HpyCH4III) was detected only during analysis of interaction and effect of PRKAR1B BmgBI only during analysis of marginal effects. In conclusion, I have to emphasise that results of analyse are influenced by small sample size (N=101). SNP from Table 1 should enter into further studies with sufficient sample size where their influence would be confirmed or rejected. Insignificance of the results can also be caused by the strict correction for multiple testing. In the future I would like to study a correction based on Bayesian static. These methods correct p-values on the basis of prior probability of true association of SNP with disease.

5. References

Armitage P. 1955. Tests for linear trend in proportions and frequencies. Biometrics 11: 375-386.

Balding DJ. 2006. A tutorial on statistical methods for population association studies. Nature Reviews Genetics 7: 781-791.

Freidlin B, Zheng G, Zhaohai L, Gastwirth JL. 2002. Trend tests for case-control studies of genetic markers: power, sample size and robustness. Human Heredity 53: 146-152.

Homer JM, Aguilar-Delfin I, Telford III. SR, Krause P, Persing DH. 2000. Babesiosis. Clinical Microbiology Reviews 13: 451-469.

Lewis CM. 2002. Genetic association studies: Design, analysis and interpretation. Briefings in Bioinformatics 3: 146-153.

Moore JH. 2004. Computational analysis of gene-gene interactions using multifactor dimensionality reduction. Expert Review Molecular Diagnosis 4: 795-803.

Ott J, Hoh J. 2003. Set association analysis of SNP case-control and microarray data. Journal of Computational Biology 10: 569-574.

Quesenberry CP, Hales C. 1980. Concentration bands for uniformity plots. Journal of Statistical Computation and Simulation 11: 41-53.

Risch N, Merikangas K. 1996. The future of genetic studies of complex human diseases. Science 273: 1516-1517.

Specification and monitoring of oscillatory properties in dynamical systems

Petr Dluhoš

Faculty of Informatics, Masaryk University; e-mail: 269281@mail.muni.cz

Abstract

Formal techniques for analysis and verification of systems are increasingly used in natural and systems sciences. Application of this approach in the new areas requires new methods. The master's thesis focuses on reasoning about complex biological signals. Existing approaches based on temporal logics are reviewed and a new temporal logic extending the Signal Temporal Logic is introduced. A polynomial monitoring algorithm for piecewise linear signals is introduced and implemented in MATLAB with the Multi-Parametric Toolbox.

Key words

Temporal logic, signals, monitoring, formal specification, oscillations.

1. Introduction

Recent development of computers made it possible not only to acquire, store and process gigantic amount of data, but also to make models of natural processes and study them in silico. The rapid development of fields like systems biology, computational biology or synthetic biology creates a demand on tools for automatic performing and evaluation of such experiments.

One of these tools is temporal logic, formalism for description and reasoning about phenomena taking place in time. It originated in philosophy and found its use in computer science to formal description and verification of technical processes. Nowadays, these techniques are increasingly used for formal analysis of natural processes (Calzone et al., 2006; Rizk et al., 2008). The usage in this field places new demands on temporal logics. The thesis deals with a class of temporal logics suited to reasoning about nontrivial real-time processes. As a model of these processes, various types of oscillations (understood in broader sense) are considered, because oscillatory behaviour plays a crucial role in nature (Hess, 2000).

The main contribution of the thesis is an extension of the Signal Temporal Logic (STL) (Maler and Nickovic, 2004). This new temporal logic (denoted STL*) enables expressing more complex biological behaviour like oscillations. In the thesis, there is also introduced a monitoring algorithm for STL* formulae working in polynomial time. The theoretical description of the algorithm is supplemented by a prototype implementation evaluated on a biological case study.

2. Methods

Formulae of the logic STL* are interpreted over real-valued continuous signals (functions from dense time domain to real values representing quantities of the investigated system, i.e., concentrations of species, abundances or other measurable attributes). These signals are

defined similarly as in (Maler and Nickovic, 2004), but are restricted to a linear form to enable efficient monitoring (i.e., determinig the satisfaction of a formula over a continuous signal). Basic properties of continuous signals are expressed via atomic predicates (Boolean functions form signal variables to the set of truth values {true, false}). These atomic predicates constitute atomic propositions in STL* formulae.

The logic STL* extends STL by adding a unary temporal operator * (called freeze operator) which freezes the values of the signal in some time instant. These values are then accessible later in other parts of the formula. This operator enables expressing of properties like "the value of the variable x is nondecreasing on some interval", "the variable x copies the values of the variable y with a delay of 4 seconds" or "the variable x is bigger than y was 5 time units ago". These properties cannot be expressed in the logic STL without using some extra information about the signal.

The monitoring algorithm for STL* formula was inspired by (Calzone et al., 2006). The idea is to construct a parse tree of the formula and check the satisfaction in a bottom-up manner, starting with construction of satisfaction sets (i.e., sets of the time instants and frozen time instants in which the formula is satisfied) for atomic predicates and continuing with construction of satisfaction sets for compound formulae on higher levels of the parse tree.

The monitored signals are supposed to be piecewise linear, which does not limit the expressivity of the logic (every measurement or output of a numerical simulation can be seen as a piecewise linear signal). Such an assumption in combination with linearity of atomic predicates enables us to express the satisfaction sets for atomic predicates as polygons in 2D space. Construction of the satisfaction sets for higher formulae is then performed as polygonal operations, for which efficient algorithms exist (de Berg et al., 2008).

3. Results

A prototype of the introduced monitoring algorithm was implemented in software MATLAB (Matlab, 2011). For the polygonal operations, the Multi-Parametric Toolbox (MPT) (Kvasnica et al., 2004) was used. Performance and functionality of the algorithm were tested by a series of tests and demonstrated on a biological case study.

A system of three transcriptional repressors built into bacteria Escherichia coli (Elowitz and Leibler, 2000) was analysed by the means of STL* logic. I was able to successfully prove the desired properties of the modeled system e.g. sustained oscillations. The parameters of the model for which this behaviour occurs could be automatically estimated. It would be also possible to verify the oscillatory behaviour in the real world system, i.e., bacteria, if I had the access to the measured data.

4. Conclusion

In my thesis, I have introduced an extension STL* of the logic STL and proposed a monitoring algorithm working in polynomial time for this new temporal logic. This algorithm takes a STL* formula and a signal and checks if the formula is satisfied over the signal. It works by transforming the problem to series of polygonal operations in plane.

The theoretical time complexity of the algorithm ($O(kn^4)$, where $\underline{k}$ is the length of the monitored formula and $\underline{n}$ is the number of points of the signal) was confirmed by tests and a case study. However, due to inefficiency of the Multi-Parametric Toolbox (MPT) used for

polygonal operations, the computations took a long time (hours for signal of 100 points). It was probably due to the fact that the MTP is not specialised for this operations in plane.

As a straightforward direction of future development, specialised algorithms for polygonal operations could be implemented. This would significantly fasten the monitoring algorithm. Next step could be the introduction of a measure of robustness (Donzé and Maler, 2010) to quantify the degree of satisfaction of a formula over a signal. This would improve the utilization of STL* in tasks like parameter estimation.

Another contribution of this thesis can be seen in the field of knowledge representation. The logic STL* can be used for succinct representation of knowledge e.g. in signals or systems and their behaviour. Representation in a form of logic formulae can be utilised in different fields of artificial intelligence (Baral, 2003).

5. References

Baral C. Knowledge Representation, Reasoning and Declarative Problem Solving. Cambridge, Cambridge University Press, 2003. 530 p. ISBN 0-511-03065-7.

de Berg M, Cheong O, van Kreveld M, Overmars M. 2008. Computational Geometry: Algorithms and Applications. Springer, Berlin, 3rd edition.

Calzone L, Chabrier-rivier N, Fages F, Soliman S. 2006. Machine learning biochemical networks from temporal logic properties. Trans Comput Syst Biol, 4220: 68–94.

Donzé A, Maler O. 2010. Robust satisfaction of temporal logic over real-valued signals. FORMATS'10, Berlin, Heidelberg, Springer-Verlag. 92–106.

Elowitz MB, Leibler S. 2000. A synthetic oscillatory network of transcriptional regulators. Nature, 403(6767): 335–338.

Hess B. 2000. Periodic patterns in biology. Naturwissenschaften, 87: 199–211.

Kvasnica M, Grieder P, Baotić M. 2004. Multi-Parametric Toolbox (MPT). URL: http://control.ee.ethz.ch/~mpt/about.php#geometry.

Maler O, Nickovic D. 2004. Monitoring temporal properties of continuous signals. Proc. of FORMATS-FTRTFT. Springer. 152–166.

Matlab. 2011. Mathworks: software MATLAB, (version 2011b). Available at http://www.mathworks.com/products/matlab/.

Rizk A, Batt G, Fages F, Soliman S. 2008. On a continuous degree of satisfaction of temporal logic formulae with applications to systems biology. In: Heiner M, Uhrmacher AM (ed.) Proc. of the 6th conf. on CMSB'08. Springer. 251–268.

Study of expression of genes specific for BRAF mutated colon tumours in early phases of tumour development

Barbora Hanáková[1], supervisor: Mgr. Eva Budinská, Ph.D.[2]

[1] *Faculty of Science, Masaryk University, Brno*
[2] *Bioinformatics Core Facility, Swiss Institute of Bioinformatics, Lausanne, Switzerland*

Abstract

The aim of this study is to demonstrate the dynamics of changes in the expression of genes that are specific for BRAF mutation in colorectal carcinomas. The study of changes in gene expression in the early phases of tumour development can help us to understand the causes of molecular changes leading to aggressive tumours.

I will focus on the available datasets in the different phases of tumour development. After that, I will identify specific genes for the particular phases of tumour development.

Key words

BRAF V600E mutation, colorectal cancer, multiple top-scoring pairs, gene signature

1. Introduction

Colorectal carcinoma, a big problem in recent years, is the 2nd most occurred tumour and Czech Republic is holding sad primacy in the incidence per population. This carcinoma can be developing covertly for a couple of months and can be detected in advanced stage.

As we know, the mutation of protooncogene BRAF appears in tumours who are not yet aggressive, which means that the cancer cells do not metastatize at that time. The protooncogene is an important part of the MAPK/ERK pathway, which regulates the growth and proliferation of cells. The mutation V600E in particular, where valin is in the position of 600 amino acid substituted by glutamic acid, leads to the permanent activation of the BRAF protein in the pathway. The cell does not respond to physiological regulation and in most cases neither to the treatment by inhibitors of EGFR. Patients with this mutation have the worst overall survival rate and early detection of tumours with this mutation may increase their chances.

2. Methods

Datasets containing notes from microchips were gained from the website of National Center for Biotechnology Information. For my analysis, it was important to choose datasets that contain gene expression from the cells of normal tissue, adenoma and primary stages of carcinoma.

Datasets were normalized by RMA (Robust MultiArray Average) (Irizarry et al., 2003). This method is generally used for the normalization of genomic data. After that, I used mTSP (multiple Top-Scoring Pairs) (Popovici et al., 2012). This classifier is based on classification of 32 pairs of genes. All analyses were done in R software, version 2.13.2.

2.1. mTSP

I had to find out suitable procedure how to compare and classify genes specific for BRAF mutated tumours into two classes. The first one was tumour with BRAF mutation and second one was tumour without this mutation. Classifier mTSP is able to recognise tumour with BRAF mutation with 96% sensitivity and 86% specificity.

This classifier is based on TSP (Geman et al., 2004) who was created earlier for pairwise comparisons of data from microarrays. This method studies the differences between two classes by finding the pairs of genes which level of expression is passing from one class to another class.

Selection of pairs of genes (G1, G2) is simply described as a selection of two elements from the set of 64 genes with repetition and depends on the order of elements (if G1<G2, then the sample is a BRAFm, otherwise it is predicted to be BRAFwt). This way all possible pairs are compared and a score as a function of the proportion of samples is computed to each pair who was correctly classified. Score was gained from the training data. Pairs were sorted according their score and then were selected the pairs whose score was above 0.6. This value says that 60% of samples were classified correctly. This threshold is the lowest possible limit to ensure robustness and specificity. Then some pairs of genes were eliminated in order to create only unique pairs. This means that the gene may occure only at one position in one pair. In this way, 32 pairs of genes were created.

The decision rule uses characteristics of average as the most representative measure. The pairs were selected according to certain rules then all of them have to have their share in final value. If the expression of one or more of the genes has extreme values, then these values should influence the final value. If average of expressions of all genes G1 is lower than the average of expressions of all genes G2, then the sample is BRAF mutated, otherwise the sample is without BRAF mutation.

Unfortunately, all 64 genes are not available in the datasets. . This method was also applied to datasets containing less then these 32 pairs of genes. Based on this results, it was found out that the classifier works reliably with only 8 pairs of genes.

2.2. Hypothesis testing of score

Score is calculated from values that you got from mTSP – score is equal to subtraction average of expressions of all genes G1 from average of expressions of all genes G2. The aim of this testing was to find out if there are statistical significant differences among score of normal tissue, adenoma and carcinoma. I could not use gene expression in this case because of following relation: if expression of gene 1 is bigger or lower than expression of gene 2, they do not show us absolute differences between two genes.

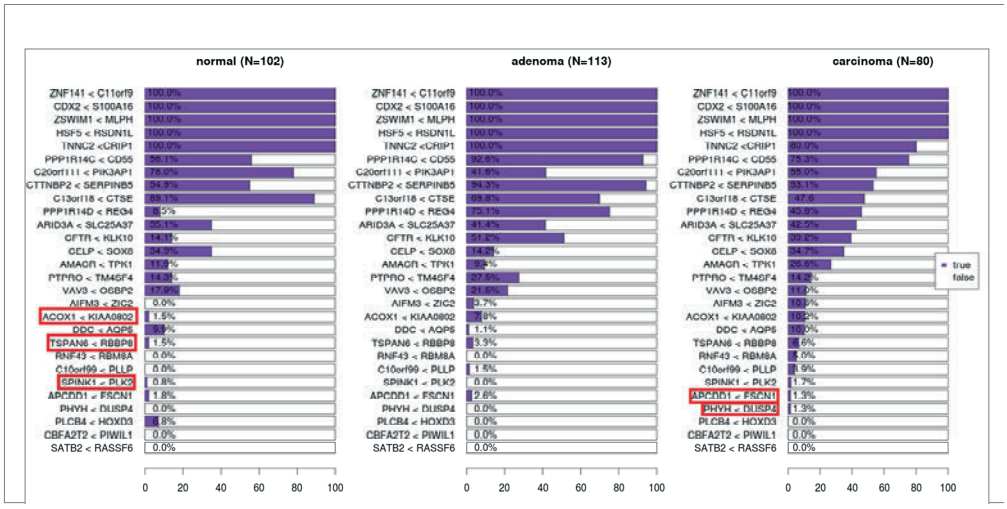
I used Shapiro-Wilk test to verify the normality of data. According the results, I could not reject the null hypothesis that the score of carcinoma has normal distribution. In the case of normal tissue and adenoma, I could reject the null hypothesis about normality of data. Due to this fact, the selection of appropriate statistical test was constricted on the non-parametric tests that can be used when data has asymetric distribution. The fact that these three groups of patients are completely independent, played another important role in selection. Based on these criteria, I decided to choose Mann-Whitney test with significance level $\alpha=0.05$. I used two-tailed test because I wanted to know if the score is bigger or lower.

After that, Benjamini-Hochberg correction for multiple comparison was used.

3. Results

3.1. Specific genes for BRAF mutation in early phases of tumour development

Figure 1. Barplot for pairs of genes in normal tissue, adenoma and carcinoma



According to mTSP I predicted 7 patients with carcinoma (8.75%), 14 patients with adenoma (11.57%) and 3 patients with normal tissue (2.94%) as a BRAF mutated.

The rule G1<G2 was held for some of the pairs of genes only in some patients who were by mTSP classified as BRAF mutated. According to this fact, I defined these genes as specific for BRAF mutation.

Specific genes for normal tissue: SPINK1<PLK2, TSPAN6<RBBP8 and ACOX1<KIAA0802.

Specific genes for carcinoma: PHYH<DUSP4 and APCDD1<FSCN1. Unfortunately I did not find out any of these connexion in adenomas.

3.1. Testing of score

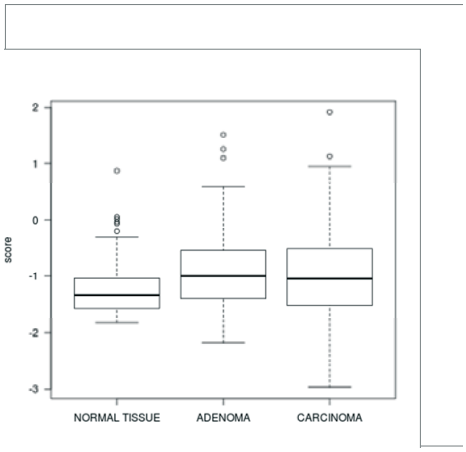
By comparing group with normal tissue and group with adenoma I rejected the null hypothesis, score of population with normal tissue is lower than the score of population with adenoma. In the case of groups with normal tissue and carcinoma I rejected the null hypothesis again. The score of population with normal tissue is lower than the score of population with carcinoma. But, I did not rejected the null hypothesis by comparing score of adenoma and carcinoma.

4. Conclusion

In this work, I presented the studies of expression of genes who were found out as a specific for V600E BRAF mutated colorectal cancers. I applied mTSP to 7 available datasets, which contain notes from microchips. Based on the results, I defined specific pairs of genes for

normal tissue, adenoma and primary stages of carcinoma. In the last step of my analysis I found out that the score of normal tissue is lower than the score of adenoma and carcinoma.

Figure 2. Boxplot of score



5. References

Geman D, d'Avignon C, Naiman DQ, Winslow RL. 2004. Classifying Gene Expression Profiles from Pairwise mRNA Comparisons, Statistical Applications in Genetics and Molecular Biology 3: article 19.

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay D, Antonellis K J, Scherf U, Speed T P. 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4: 249-264.

Popovici V, Budinska E, Tejpar S, Weinrich S, Estrella H, Hodgson G, Xie T, Bosman F, Roth AD, Delorenzim M. 2012. Identification of Poor-Prognosis BRAF-Mutant-Like Population of Patients With Colon Cancer, Journal of Clinical Oncology 30:1288-1295.

Spatial modelling of vegetation based on bioclimatic data

Lenka Krupková, supervisor: Mgr. Klára Komprdová, Ph. D.

Faculty of Science, Masaryk University, Brno

Abstract

The main goal of this study is to create a model of probability distribution of vegetation of the seven most important tree species in the area of the southern Siberia. The work is built on the theory of the analogy of current vegetation of southern Siberia and glacial vegetation of the Central and Eastern Europe.

Key words

Bioclimatic modelling, ice age, last glaciation, Siberian vegetation

1. Introduction

This work pursues bioclimatic modelling of vegetation in the area of the southern Siberia. On the basis of recent researches (Kuneš et al., 2008) is possible to say that current southern Siberian vegetation is similar to the last-glacial vegetation of Central and Eastern Europe. According to the newest findings (Barron and Pollard, 2002), it is presumed that in the coldest eras of the last glaciation, which occurred before 35 000 – 13 000 years, there grew in the Central and Eastern Europe forests. The aim of this study is to create a model of the vegetation of the southern Siberia on the basis of data sampled in this area.

1.1. Current vegetation of the southern Siberia

The present days vegetation of the southern Siberia is characterised by the three main biomes: taiga – mainly forests of pine tree and spruce with sporadic presence of larch, birch and some other deciduous trees; tundra – mainly grassland with sporadic presence of stunted coniferous and broadleaf wood and common presence of lichens and mosses; extensive grasslands. Similar could be vegetation of glacial central-east Europe.

2. Materials and methods

2.1. Data

The dataset contains records of species composition in phytosociological plots on 633 southern-Siberian localities in two separated areas (mountains Altai and Sajan). It also contains information about occurrence of approximately 1 300 species and information about environmental variables such as elevation, average temperature (June, January, annual), average precipitation (annual, winter, summer) and computed values of radiation. The areas of the Altai and Sajan were divided into cca 140000 squares, using the ArcGIS geographical information system. Each square was assigned to environmental variable values.

2.2. Modelling method

There were modelled the presence of the seven most important tree species: *Abies sibirica, Betula pendula, Pinus sibirica, Pinus sylvestris, Picea obovata, Larix sibirica* and *Populus tremula.*

Logistic regression is used as the modelling method. This regression technique is also applicable on classification problems. There is modelled probability of presence of some event. Dependent variable is binomial and is estimated by the set of continuous or discrete predictors.

One of the main assumptions of regression is absence of correlation among predictors. Therefore it was necessary to check if used predictors are uncorrelated.. Strong correlations are present among different types of temperatures and also precipitation. Therefore the model with only one from those predictors was chosen in each case to prevent correlation problem. Radiation and heat were not statistically significant in any model. The probabilities of occurrence were calculated separately for tree species from simple logistic equations. These probabilities were assigned to all squares by means of the geographical information system (ArcGIS) and the potential habitat distribution map for each tree for the entire area of the southern Siberia was created. Accuracy of the model was calculated for both Altai and Sajan. In every testing model, presences and absences were known. Models were compared for each tree species separately. The best is considered the model who has the highest accuracy for presence and absence of species.

3. Results and discussion

Three different models were made for each tree species. First model is computed on whole dataset (Sajan + Altai). The second one is computed on Sajan dataset and tested by samples from Altai and finally the third model is computed on Altai dataset and tested on Sajan.

These models were compared in the next step by the Nagelkerke R^2 coefficient (coefficient of determination designed especially for the logistic regression similar to the classical R^2 used in the linear regression) and by the accuracy of classification presence and absence of species. For computing the accuracy, it is necessary to convert probabilities to values 0 and 1. This is done by the following rule: value 0 is assigned to localities with probability of presence less then middle value of the interval reached probabilities for the model, value 1 is assigned to all others.

Comparison of Nagelkerke R^2 and accuracies of classification for each model is presented in Table 1.

Table 1. Accuracy of classification in percentage, P – presence, A – absence and Nagelkerke R^2 of each model

Species	Model (percentage of correct classification)										Model (R^2)		
	Total		Altai/ Altai		Altai/ Sajan		Sajan/ Altai		Sajan/ Sajan		Total	Altai	Sajan
	P	A	P	A	P	A	P	A	P	A			
Abies sibirica	99	56	53	43	14	38	71	53	97	77	0.48	0.58	0.54
Betula pendula	99	38	33	51	84	58	61	38	84	80	0.40	0.40	0.43
Pinus sibirica	87	62	20	67	61	60	27	53	83	67	0.36	0.34	0.36
Pinus sylvestris	100	29	42	38	76	62	58	47	70	72	0.37	0.58	0.46
Larix sibirica	96	42	51	47	49	62	75	26	82	69	0.25	0.13	0.34
Picea obovata	90	46	29	46	86	52	33	40	69	74	0.21	0.17	0.40
Populus tremula	100	56	50	56	86	42	63	40	90	71	0.32	0.26	0.34

From the comparison it is seen that the best model is the one computed on Sajan dataset. Model computed for the Altai is the worst one. This can be caused by the smaller range of gradient of conditions on Altai area, therefore it is not possible to compute accurate model.

Potential habitat distribution map for each tree was created using probability. It is shown map of occurrence probability of *Abies sibirica* in the Altai (Fig. 1) area and in in the Sajan area (Fig. 2).

Figure 1. Map of probability of presence of Abies sibirica in the area of Altai

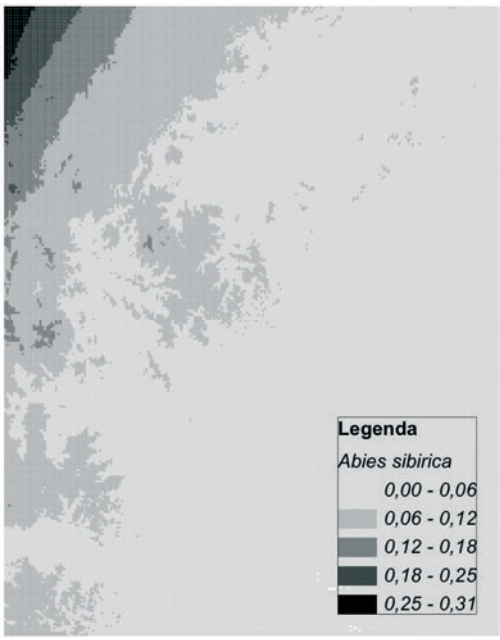
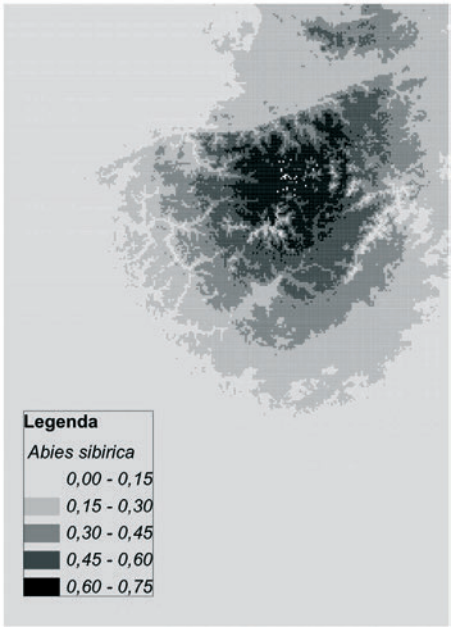


Figure 2. Map of probability of presence of Abies sibirica in the area of Sajan



4. Conclusion

Well functioning models of southern Siberian vegetation were made, nevertheless, in the consequence of competition of species and similar ecological requirements of modelled species, it is not appropriate to nest these models on the glacial Central and Eastern Europe. For this purpose will be made models of vegetation types.

3. References

Barron E, Pollard D. 2002. High-Resolution Climate Simulations of Oxygen Isotope Stage 3 in Europe. Quaternary research, 58: 296-309.

Kuneš P, Pelánková B, Chytrý M, Jankovská V, Pokorný P, Petr L. 2008. Interpretation of last-glacial vegetation of Eastern Central Europe using modern analogues from southern Siberia. Journal of Biogeography, 35: 2223-2236.

Modelling of acidification of forest soils with the inclusion of uncertainties

Petra Malcová

Faculty of Science, Masaryk University, Institute of Biostatistics and Analyses MU, Research Centre for Toxic Compounds in the Environment, Brno; e-mail: 356608@mail.muni.cz

Abstract

The aim of this work is to make a summary of the current state of knowledge in the modelling of acidification of forest soils in the catchment Lysina, in this issue to select the appropriate mathematical model that will be implemented in the software Maple and then to make the analysis of the uncertainty on the selected model. In the first part, there is introduced the basic overview of processes that cause acidification of forest soils. It contains a brief description of selected mathematical models addressing acidification of forests and a more detailed description of the model MAGIC, which is subsequently used in the work. The last part solves the uncertainty in the selected model. This chapter contains a theoretical summary of the uncertainties and sensitivities in mathematical models and also application of knowledge to the model that solves the acidification of forest soils on catchment level. The result of modelling is that in the last twenty years attended to improving soil condition. In the future it is expected further improvement in soil condition, but not as pronounced as in previous years.

Key words

Mathematical model, acidification, atmospheric deposition, uncertainties, sensitivity.

1. Introduction

Nowadays the topic of modelling of acidification of forest soils on river basin level is becomes very current environmentally. Acidification is one of the most serious problems of soil caused by human activity. This issue is addressed to the Ministry of Environment, along with other scientific institutions.

2. Acidification of forest soils

Acidification of forest soils is a long and gradual process that has both of natural and anthropogenic causes. The air pollutants sulfur and nitrogen oxides are important source of hydrogen cations of anthropogenic origin. This source has become one of the most important in the 20th century and therefore it is most associated acidification of soils. As a result of changes of soil acidity, it is affecting not only the soil chemistry and nutrient cycles, but also the entire forest ecosystem (Hruška and Cienciala, 2002).

3. Lysina basin

Lysina Basin is located in the peak part of the Slavkov Forest. About 10 km away in the Sokolov brown coal basin is placed Tisová power. The geological bedrock consists of granite with a low content of basic cations and is covered with a layer of brown podsolic soil with a

thickness of about 1 meter. About 70% of the basin is covered with a spruce monoculture and the rest clearings planted with young spruces. Basin appears to be all the typical signs of chronic acidity (Hruška et al., 1996). The study was verified of data from this basin.

3.1. Modelling of the most development of major ions in the soil

For modelling of the most important ions in the soil, equations of the model MAGIC were used. At first, the past of of individual inos has been displayed and then calculated their rate of changes.

3.1.1. Results of modelling

The evolution of base cations and sulfur dioxide over the past 20 years, is represented in figures 1 and 2 and graphs showing the expected future development in the soil are depicted in figures 3 and 4. It is evident at the graph of basic cations that their amount in the soil over the last twenty years is still declining. The next graph shows that quantity of carbon dioxide which contrarily helps to the acidification of soils was reduced too. Compared to 1990, the amount of carbon dioxide was minimized and a number of other acids was reduced. Even though the amount of basic cations was reduced, one can say that there was a slight improvement of the soil over the last twenty years. When modeling predictions of ions, there was found out that number of major ions, who mostly contribute to acidification, will decrease in the future. Amount of basic cations in the soil will not change radically, but it is positive fact that they are not expected to decline. Based on these results, it is still expected an improvement of soil conditions, but this improve will not as fast as in the previous years.

4. Models with the inclusion of uncertainty

In modelling, it is investigated system replaced by model. Because of it, there always occurs a simplification of the system, so we get an incomplete model of the original system. The model contains elements called uncertainty.

4.1. Uncertainty Analysis

4.1.1. Interval arithmetic

The interval arithmetic is used to describe of a date uncertainty made by inaccuracy of measurement or due to the existence of several alternative methods for estimating the parameters. In the Maple system, it was divided into more classical functions (minimum, average, maximum) and plotted in one graph. Figures 5 and 6 show the dispersion of equation for the solution of ammonium cation and carbon dioxide. The ammonium cation graph shows that the default parameter settings are correct because negative values can´t occur.

4.1.2. Probabilistic analysis

Probabilistic analysis is the most widely used method of description of uncertainty. Individual parameters are treated like a random variables with the probability distribution of values that may acquire. The aim of this analysis is to determine the probability distribution of output values of the model initially at time 1 and then in time 10. The result is that the probability distribution doesn´t modify and it remains triangular.

4.2. Analysis of sensitivity

While uncertainties occur in the analysis, we want to determine the overall uncertainty in model output using the input characterized by uncertainty, at the sensitivity analysis we

examine how outcomes are influenced by various parameters of the model (Hřebíček et al., 2010; Hřebíček et al., 2011).

4.2.1. Local sensitivity

Local sensitivity determines how the model output is sensitive to changes of input parameters in the "close" some of their representative value. So it examines the local effect of changes of only one of parameters in the solution of the model. Calculated listed indexes of local sensitivity says, that in all the model equations are locally the most sensitive parameters which represent the input flow of ions in the equations. These parameters can have the greatest effect on the change in model solutions.

4.2.2. The global sensitivity

Global sensitivity analysis method is used in models where one of the parameters has no default value and can have a wider range of values. The Sobols method was used for the calculation (Hřebíček et al., 2011; Urbánek, 2009).

The analysis shows that the one of the most sensitive parameters is the initial quantity of the ions. It is obvious that the resulting rate of change of all ions is the most depended to this parameter. Contrary to the other ions, in the case of the equation for sulfur dioxide was found out that all parameters are very significantly involving in the distortion of solution. Withholding atmospheric deposition is very sensitive parameter because it is the largest flow of ions into the soil. The other parameters can be considered less sensitive.

Figure 1. Development of basic cations in the soil over the past 20 years.

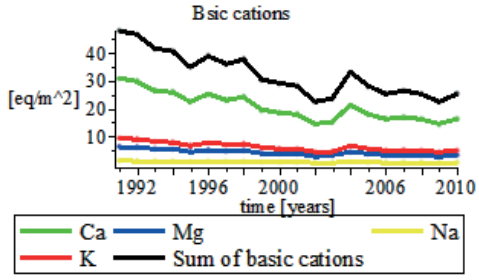


Figure 2. Development of sulphur dioxide in the soil over the past 20 years.

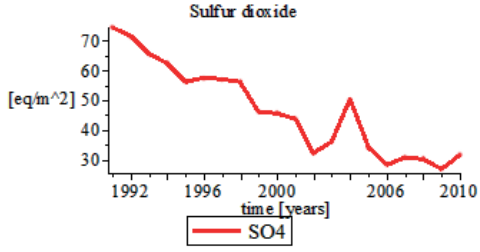


Figure 3. Expected development of basic cations in the soil by 2040.

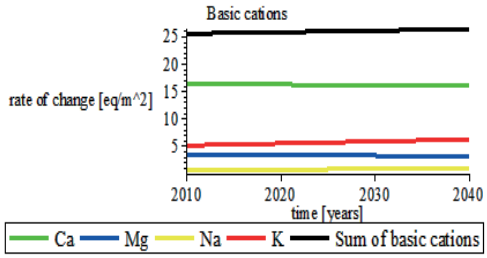


Figure 4. Expected development of sulphur dioxide in the soil by 2040.

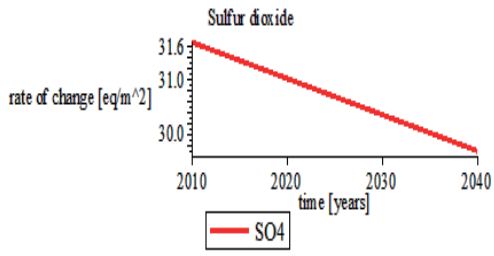


Figure 5. Dispersion equation for the solution of ammonium cation.

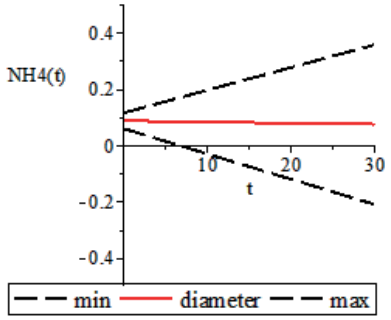
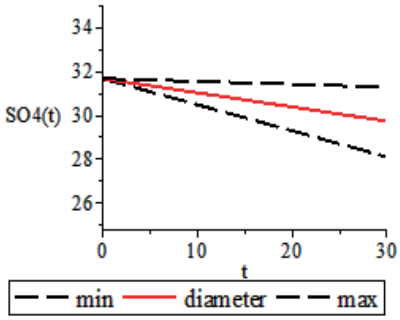


Figure 6. Dispersion equation for the solution of sulphur dioxide.



5. Conclusion

The acidification of forest is very dynamic and complicated process where is important to deal with the past. It is necessary to have large amounts of data and parameters to the exact modelling. The modelling is done using Maple software on data from the basin Lysina, the Czech Geological Survey. There is apparent on the basis of the modelling result that soil condition will improve in the future only slightly. The presented results are only approximate, because I did not obtain sufficient data, especially data relevant to soils. Because the modelling means a simplification of reality, the last part contains an analysis of model uncertainty. At first, there was detected the uncertainty at the output of model with the help of characterized uncertainty at the input and then, how the outputs are influenced by various parameters of the model - a sensitivity analysis.

6. References

Hruška J, Moldan F, Krám P. 1996. Effect of acid rain on surface water - modelling of past and future basin Lysina in the Slavkovský forest. Universe, 75: 373 – 376.

Hruška J, Cienciala E. 2002. Long-term acidification and nutrient degradation of forest soils - limiting factor of the current forestry. Ministry of Environment, 159 pp.

Hruška J, Oulehle F, Krám P, Skořepová I. 2009. Effects of acid rain on forest and aquatic ecosystems – 2. Effect of sulfur and nitrogen deposition on soil and forests. Živa, 3/2009, 141 – 144.

Hřebíček J, Pospíšil Z, Urbánek J. 2010. Introduction to the mathematical modelling with using Maple. Academic publishing CREM, s r.o., Brno, 118 pp.

Hřebíček J, Kubásek M, Kohút L, Matyska L, Tokárová L, Urbánek J. 2011. Scientific computing in mathematical biology. Masaryk University, Brno, 117 pp.

Urbánek J. 2009. Numerical stability in the environmental modelling with uncertainties. Masaryk University. Brno, Thesis, 68 pp.

Statistical models for ecological assessment of reservoir using phytobenthos

Lucie Panáčková

Faculty of Science, Masaryk University, Brno

Abstract

This essay analyses the relationship between species composition of phytobenthos and trophic state of selected reservoirs in the basin of the Morava River in the years 2008, 2009 and 2011. For evaluation of the relationships of variables describing individual reservoirs the methods were used which evaluate the reservoir trophic state of stagnant water by means of diatom index, diversity indices, coefficients of similarity and multivariate statistical methods. By using of Spearman correlations there have been shown the relationships between indexes of biological water quality assessment, the index number of taxa of diatoms, the diversity index and the indexes of evaluation of saprobity and trophy. Thanks to the cluster analysis it has been verified that the contents of chemical substances is reflected in the occurrence of diatoms in reservoirs, their number and diversity. The number of diatoms and their diversity increases with the water quality. But also geo-morphological characteristics of the reservoirs have a significant influence on chemical composition of the reservoirs.

Key words

phytobenthos, biodiversity, diatom index, multivariate statistical methods

1. Introduction

Water reservoirs are significant not only in terms of components of the landscape, but also they fulfil many important requirements of a mankind. There are some important functions of water reservoirs. They are resources of drinking water and they serve as flood control, source of water energy for hydroelectric power, for recreation, fish farming and as a means of balancing of the flow of rivers. However, bigger consumption of water means also bigger pollution. Biological status of the reservoir can be determined by using fish, benthic invertebrates, phytoplankton, macrophytes and phytobenthos. This essay deals with phytobenthos assessment. Evaluating phytobenthos, a diatom part of phytobentos is used because changes in the aquatic environment evocate an immediate reaction of diatoms by changing their taxonomic and quantitative composition. Because of short lives diatoms are able to build a new biota in a few weeks (Schaumburg et al., 2004).

2. Methods

The data for the assessment of the relationship between species composition and trophic state of phytobenthos were obtained from the biomonitoring of the Morava River. They include first and second sampling in 21 reservoirs. The data contain the records of the chemical mixed sample of the reservoirs and tributaries phytobenthos species composition. Reservoirs are described by using of diatom indexes (indexes evaluating the biological water quality: TDI, IBD, IPS, IDG, indexes evaluating saprobity and trofic: Rott's (SI) index, Rott's (Ti) index and Sládeček's saprobic index, index of the number of taxa of diatoms and the index

diversity). These indexes are calculated by using the programme Omnidia, chemical components and geo-morphological characteristics of the reservoirs.

Prior to the analyzing of the relationship between phytobenthos species composition and trophic conditions, it was required to make basic descriptive statistic of data and some other extra analysis for better orientation in the data to select appropriate evaluation analysis. For comparison of the first and second samplings, Wilcoxon paired test was used, who determines whether it will be necessary to assess particularly the first and second samplings. For the evaluation of relationships between variables, Spearman correlation was used.

There was also performed a cluster analysis on the data. The aim was to create 2 groups of clusters according to chemistry and the occurrence of diatoms in the reservoirs for subsequent comparison of dams. Before application of methods for chemical parameters, the data was standardized. The data on the occurrence of diatoms were adjusted, too, and consequently they were converted into binary data. It was used several clustering methods, but the best results were obtained using Jaccard coefficient and Ward method for the occurrence of diatoms, Euclidean distances and the furthest neighbour methods for chemical parameters.
It was chosen the appropriate number of clusters on the demographers according to knowledge of the dams so that the group gave the most meaningful results.

3. Results and discussion

For comparison of the first and second samplings Wilcoxon paired test was used, who did not show at the indexes (the level of statistical significance $p < 0.05$) statistically significant difference between the first and second samplings, therefore it was not necessary to assess the first and second samplings extra.

The calculated Spearman correlations between indexes and chemicals showed a statistically significant correlation at significance level 0.05. Index of IBD correlates with BOD5, TOC, N-NO2 and total P, index TDI BOD5, total Mn and P, Rott's (TI) index BOD5, TOC, Mn, N-NO2, total P and chlorophyll, Rott's (Si) index BOD5, TOC, Mn, N-NO2,
N-NO3, total P and chlorophyll, Sládeček's saprobic index BOD5, Mn, total P and total N.

On the demographer according to chemical parameters there were chosen three clusters marked with Roman numerals I, II and III. According to the occurrence of diatoms there were selected six clusters identified by Arabic numerals from one to six.

Using cluster analysis according to the chemical parameters there was achieved three groups of reservoirs, which are reservoir with a similar chemical composition. The diversity of groups is also geographic question and that is mainly a matter of altitude, the shape of the reservoir and its subsoil. The first cluster is located at South Moravia, region characterized by lowlands. Reservoirs of cluster II are largely situated in the Highlands and the cluster III is typical East Moravia. The cluster analysis using the chemical parameters showed that the chemistry of reservoirs is associated with water quality and quantity of diatoms. The reservoirs with lower quality of water have fewer diatoms than the cleaner reservoirs. On the contrary, the second cluster analysis that was applied to the occurrence of diatoms showed six groups of clusters of reservoirs. The cluster 1 is reservoir Nové Mlýny in 2008 and the lower reservoir in 2011 (first sampling). In addition to reservoir Nové Mlýny is also Vranov reservoir (its first sampling), which also lies on the River Dyje. The cluster 2 comprises reservoirs, which belong to the larger mesotrophic, eutrophic to hypertrophic

reservoirs. The cluster 3 belongs to higher put reservoirs, but there is a surprising presence of the second sampling mesotrophic reservoir Bystřička. The explanation may perhaps be a remnant of diatoms similar to other reservoirs of the cluster at an earlier period when the first sampling from the reservoir Bystřička at this time indicates a eutrophic reservoir. The group 4 is the largest cluster group, which includes reservoirs with similar geographical characteristics of the reservoir as it was in the group of cluster 3, but cluster 4 has better water. The cluster 5 contain only two reservoirs which includes the first sampling from reservoir Landštejn and the second sampling from reservoir Nová Říše. Reservoirs are very close both the geographic distance and the higher altitude and small size of reservoirs. There is a good quality of water which is showed by the results of diatom indexes. So these reservoirs have a very low saprobity. The cluster 6 contains reservoirs Bojkovice, Horní Bečva and Ludkovice that are smaller and do not reach great depth.

4. Conclusion

Between variables of Spearman correlations groups of correlations were tested, who are between indexes which evaluate trophic and saprobity (TDI Rott's (TI) index, saprobic Rott's (SI) index and SLA saprobic index) as well as water quality assessment index (IBD, IPS and IDG) and indexes of diversity and number of taxa. Spearman correlation was used for relations between the indexes and chemicals. Index of IBD correlates with BOD5, TOC, N-NO2 and total P, index TDI with BOD5, total Mn and P, Rott's (TI) index with BOD5, TOC, Mn, N-NO2, total P and chlorophyll, Rott's (Si) index with BOD5, TOC, Mn, N-NO2, N-NO3, total P and chlorophyll, Sládeček's saprobic index with BOD5, Mn, total P and total N.

In addition, reservoirs were divided into groups of clusters by using cluster analysis. Two groups were formed as clusters for subsequent comparison of reservoirs. The cluster analysis divided the dataset into three clusters according to chemical parameters and six clusters according to the occurrence of diatoms. The analysis showed that the distribution of reservoirs into clusters by using cluster analysis closely related with a trophy of reservoirs and the geo-morphological characteristics (especially altitude) and ground of reservoirs, who are certainly related to the occurrence of diatoms.

By using cluster analysis according to the chemical parameters, it was possible to verify that the contents of chemical substances are reflected in the occurrence of diatoms in the reservoirs, their number and diversity. The number and diversity of diatom increase with water quality. But, a geo-morphological characteristic of the reservoirs also has the significant influence over chemical composition of the reservoir. . For unambiguous identification, if diatoms have some specifics on the trophic status of reservoirs or their presence affects other properties of the reservoir is a need of further analysis.

5. References

Schaumburg J, Schranz Ch, Hofmann G, Stelzer D, Schneider S, Schmedtje U. 2004. Macrophytes and phytobenthos as indicators of ecological status in German lakes - a contribution to the implementation of the Water Framework Directive. Limnologica 34: 302-314.

Regression diagnostic tools in survival analysis

Ivana Svobodová

Faculty of Science, Masaryk University, Brno

Abstract

Regression models are very useful tool for analyzing and summarizing survival data; however, they have their own statistical assumptions. Another problematic aspect of survival data is a phenomenon called censoring. Thus, special statistical techniques are necessary for survival data analysis and modelling. The aim of this contribution is to introduce various regression diagnostic tools for survival models. These tools can be used to assess overall model goodness-of-fit, to check proportional hazards assumption and to look for potential outliers and influential observations. The use of the presented methods is documented on real data.

Key words

Regression diagnostic tools, survival analysis, overall goodness of fit, analysis of residuals, proportional hazards assumption, outliers, influential observations.

1. Introduction

In survival analysis, regression models are used to describe relationship among an outcome variable and one or more independent variables. However, models have own assumptions and their validity must be satisfied. If assumptions do not hold, models can be wrong and results are then incorrect. Thus models assumptions must be verified. More models can be applied to one type of data; important step is also selection of the model, that best fits the dataset.

Regression diagnostic tools help us with checking of model's assumptions; with these tools we can also select the model that is the best for our dataset.

2. Survival analysis

In survival analysis, the main outcome variable is the time to an event of interest; the generic name for the time is survival time (Clark et al., 2003). If we cannot observe survival time for some patients, we say that their survival time is censored. Censoring is specific for survival data. Survival analysis deals with following problems: (1) estimation of survival time distribution; (2) comparison of survival of different groups of patients; (3) prognostic evaluation of different variables (Marubini and Valsecchi, 1995).

We have two main groups of regression models in survival analysis: proportional hazard models (PH models) and accelerated failure time models (AFT models).

3. Regression diagnostic tools

Regression diagnostic tools are used to assess model adequacy, i.e. whether explanatory variables are correctly selected and whether outliers or influential observation are not

presented; we can choose from a group of models the model that best fits to the dataset. Regression diagnostic tools are used to check validation of model's assumptions, e.g. proportional hazard assumption, which is assumed in PH models.

Residuals are a large group of regression diagnostic tools. Residuals and related diagnostics can be used for examining different aspects of model adequacy. Essentially, they are defined as a difference between observed and model-predicted quantity (Bradburn etl al., 2003). There are four main types of these residuals. Martingale residuals are defined as a difference between observed and predicted number of events. By transformation of these residuals, we obtain new type of residuals, deviance residual. Deviance residuals are more symmetrically distributed. With martingale and deviance residuals, we can assess model goodness of fit. If residuals are symmetrically distributed around zero, model is appropriate. Subjects that are away from whole dataset are identified as outliers, i.e. poorly predicted subjects. Next type of residuals, Schoenfeld's, are used for examining of proportional hazard assumption. This assumption means that failure rates of any two individuals are proportional (Marubini and Valsecchi, 1995). If this assumption holds, Schoenfeld's residuals are randomly distributed around zero and do not show any obvious trend. Last type, score residuals, measures the leverage exerted by each subject on parameter estimates (Marubini and Valsecchi, 1995). Subjects that are away from whole dataset are identified as influential observations.

There are specific regression diagnostic tools for assessing overall goodness of fit. The Grønnesby and Borgan test is based on the martingale residuals. The basis of the test is a grouping of subjects by their risk score and summing residuals in each group (May and Hosmer, 1998). If the model is appropriate, sum in each group should be close to zero. Akaike's information criterion (AIC) is a statistic that trades off a model's likelihood against its complexity (Bradburn etl al., 2003). AIC can be used for comparing parametric models and selects the model who is the most appropriate.

Besides Schoenfeld's residual, proportional hazard assumption can be assessed by other methods. First we can assess this assumption graphically. Data are divided into groups. If a PH model is valid, a plot of the logarithm of the cumulative hazard function in each group against the time give rise to lines are parallel (Bradburn etl al., 2003). This plot is known as log(-log(survival)). For assessing with test, there are several ways. First, we can assess proportional hazard assumption by test, which tests whether the effects of covariate changes with time. The test is known as time-dependent covariate test. There are also tests, who test association between residuals and time, namely weighted residuals score test and linear correlation test (Ng'andu, 1997).

4. Results

Above-mentioned regression diagnostic tools were applied to a real dataset. Data were created by file of patients with chronic myelogenous leukemia that were diagnosed in chronic phase from year 2000 to 2008. Time to event was time to complete cytogenetic response. Cox's proportional hazard model was created.

Overall goodness of fit was tested by Grønnesby and Borgan test. Observed and expected numbers of events in each group were approximately same and the fitted model was correct.

Proportional hazard assumption was tested by graphical method and by weighted residuals score test. Test denoted one variable as a non-proportional.

Outliers were detected by martingale and deviance residuals and influential observations by score residuals. There was one outlier and two influential observations.

5. Conclusion

Regression diagnostic tools helped us in finding outliers and influential observations. A few patients were denoted as a problematic. Also, the non-proportionality was found in one case. . For these reasons, regression diagnostic tools are very useful and their application is very important for each model development.

6. References

Bradburn MJ, Clark TG, Love SB, Altman DG. 2003. Survival Analysis Part III: Multivariate data analysis choosing a model and assessing its adequacy and fit. British Journal of Cancer 89(4): 605-611.

Clark TG, Bradburn MJ, Love SB, Altman DG. 2003. Survival Analysis Part I: Basic concepts and first analyses. British Journal of Cancer 89(2): 232-238.

Marubini E, Valsecchi MG. 1995. Analysing survival data from clinical trials and observational studies. New York: J. Wiley, 414 p. ISBN 04-719-3987-0.

Maz S, Hosmer DW. 1998. A Simplified Method of Calculating an Overall Goodness-of-Fit Test for the Cox Proportional Hazards Model. Lifetime Data Analysis 4(2): 109-120.

Ng'andu NH. 1997. An empirical comparison of statistical tests for assessing the proportional hazards assumption of Cox's model. Statistics in medicine 16(6): 611-626.