Premošćivanje jaza između tehnologije mikropostroja i rutinske kliničke dijagnostike: pristup smanjenju dimenzionalnosti profila genske ekspresije zasnovan na slučajnim šumama

Debeljak, Željko

Premošćivanje jaza između tehnologije mikropostroja i rutinske kliničke dijagnostike: pristup smanjenju dimenzionalnosti profila genske ekspresije zasnovan na slučajnim šumama

Bridging the gap between microarray technology and routine clinical diagnostics: a Random Forest approach to the gene expression profile dimensionality reduction

Željko Debeljak

Odjel za medicinsku biokemiju. Klinička bolnica Osijek, Osijek
Department of medical biochemistry, Osijek University Hospital, Osijek, Croatia

Sažetak

Uvod: Analiza genske ekspresije zasnovana na mikropostrojima je tijekom proteklog desetljeća prepoznata kao koristan alat od strane znanstvene zajednice, ali nije ušla u rutinsku dijagnostičku primjenu. Kako je skupa i podložna značajnim eksperimentalnim varijacijama, na trenutnom tehnološkom stupnju razvoja ta tehnologija nije prikladna za rutinske kliničko-dijagnostičke primjene. U svrhu premošćivanja jaza između mogućnosti navedene tehnologije i potreba kliničke dijagnostike razvijeni su različiti računalni alati za smanjenje dimenzionalnosti. Njihova osnovna svrha je odabir malog skupa kandidata za biomarkere iz ogromnog skupa sadržanog u profilima genske ekspresije prikladnog za rutinsko postavljanje dijagnoze.

Cilj: Slučajna šuma (engl. *Random Forest,* RF) se nametnula kao pouzdan pretkazatelj. Ipak, njene su mogućnosti u odabiru relevantnih gena privukle manje pažnje. Cilj ove studije je evaluacija prikladnosti na RF-u zasnovanoga odabira biomarkera iz skupova genskih profila. Tri takva skupa, preuzeta iz literature, prikupljena tijekom manjih kliničkih pokusa izabrana su u navedenu svrhu.

Rezultati: Dobiveni rezultati ukazuju da RF može lako identificirati dobre univarijatne klasifikatore, tj. pojedinačne biomarkere kada je složenost skupa mala. Za nešto složenije probleme pouzdani dvodimenzionalni klasifikator može se također pronaći. Ipak, ako je odnos između dijagnoze/prognoze i profila genske ekspresije vrlo složen ili ako je skup premalen, na RF-u zasnovano smanjenje dimenzionalnosti ne omogućava odabir pouzdanog skupa kandidata za biomarkere.

Zaključci: Unutar ograničenja zadanih složenošću skupa RF predstavlja prikladan alat za izbor kandidata za biomarkere.

Ključne riječi: genska ekspresija; mikropostroj; probiranje biomarkera; slučajne šume; izbor svojstava

Abstract

Introduction: Although recognized as a valuable tool by scientific community, microarray based gene expression profiling has not accessed routine diagnostic application during the last decade. Since this approach is expensive and prone to substantial experimental variation, it is not suited for routine clinical diagnostic purposes at the current state of technology. In order to bridge that gap, different computational dimensionality reduction tools have been developed. The principle of their application is selection of a limited set of biomarker candidates from huge gene expression profiles appropriate for routine diagnostic assessment.

Aim: Random forest (RF) has been established as a reliable predictor. However, its relevant gene selection capabilities gained less attention. The aim of this study was to evaluate suitability of RF for biomarker selection from gene expression profile datasets. Three datasets taken from literature, obtained during small-scale clinical experiments, were chosen for that purpose.

Results: The results obtained show that RF could easily identify good univariate classifiers, i.e. single biomarkers when the problem at hand is of low complexity. For more complex problem a reliable two-dimensional classifier candidate could be also found by this approach. However, when the relationship between diagnosis/prognosis and gene expression profiling results are highly complex or the dataset is too small, RF-based dimensionality reduction fails to select a reliable set of biomarker candidates.

Conclusions: Within dataset complexity limitations, RF represents an appropriate tool for biomarker candidate selection.

Keywords: gene expression; microarray; biomarker screening; random forests; feature selection.

Pristiglo: 24. srpnja 2006. *Received: July 24, 2006*

Prihvaćeno: 7. rujna 2006. *Accepted: September 7, 2006*

Uvod

Trenutni status tehnologije mikropostroja čini mogućom usporednu ekspresijsku analizu desetaka tisuća ljudskih gena s jednog mikropostroja (1). Unatoč tome, u odnosu na rutinske laboratorijske dijagnostičke tehnike koje mogu proizvesti točne i klinički vrijedne rezultate za nekoliko minuta ili sati uz nisku cijenu, ova tehnologija predstavlja vrlo skup, spor i neučinkovit dijagnostički alat. Za pouzdanu dijagnostičku primjenu, osim mikropostroja, reagensa i čitača mikropostroja ona zahtijeva sofisticiranu računalnu podršku i mjerenje najmanje u triplikatu. Ova svojstva čine tehnologiju neprikladnom za rutinski dijagnostički rad.

Ipak se dijagnostička primjena tehnologije mikropostroja namijenjene analizi genske ekspresije može barem razmotriti. Korisnik može primijeniti analizu genske ekspresije u svrhu diferencijacije dvaju usko povezanih kliničkih stanja, odnosno u svrhu postavljanja dijagnoze u iznimno složenim slučajevima. Osim toga, skup genskih profila prikupljen tijekom kliničkog pokusa koji je zasnovan na kontrolnoj i testnoj skupini sudionika može se primijeniti za izbor genskih podskupova relevantnih za identifikaciju i/ili diferencijaciju analiziranoga kliničkog stanja. Kvantitativna analiza RNA ili proteinskih produkata ovih gena u tjelesnim tekućinama ili tkivima je znatno jednostavnija, jeftinija, brža i pouzdanija alternativa u odnosu na mikropostroje. Izbor relevantnih gena i evaluacija kliničke korisnosti određivanja odgovarajućih RNA i proteinskih produkata predstavlja okvir za primjenu mikropostroja u svrhu probiranja novih biomarkera.

Rezultati analize genske ekspresije prikupljeni tijekom dobro organiziranih kliničkih pokusa predstavljaju bogat izvor podataka o ispitivanom kliničkom stanju. U stvari, matrice podataka prikupljene tijekom takvih pokusa su prebogate informacijom. One sadrže stotine tisuća brojčanih podataka koji ih čine presloženima za jednostavnu vizualnu provjeru i analizu. Uz takve postavke izbor nekolicine relevantnih gena iz skupa od više desetaka tisuća gena predstavlja izazovan zadatak za smanjenje dimenzionalnosti. Stoga je potrebna računalna podrška.

Računalne provedbe metoda strojnoga i statističkog učenja poznate pod nazivom filtri mogu se primijeniti u navedene svrhe. U slučaju monogenskih stanja/bolesti konvencionalni statistički alati poput ANOVA-e, t-testa i njihovih neparametarskih pandana mogu se primijeniti sa ili bez modifikacija (1,2). Ti se alati mogu pouzdano primijeniti čak i u slučajevima koje karakterizira nezavisna promjena nekolicine gena. Ipak, u većini situacija analizirana klinička stanja su posljedica visoko međuzavisnih, multigenskih promjena. U takvim uvjetima multivarijatne računalne i statističke metode predstavljaju prikladan alat za izbor relevatnih gena ili postavljanje dijagnoze (3). Ponekad složenost međuovisnosti određenoga kliničkog stanja i odgovarajućih profila genske ekspresije onemogućava smanjenje njene dimenzionalnosti (4). U nekim slučaje-

Introduction

Current state of microarray technology makes parallel expression analysis of tens of thousands human genes possible on a single slide (1). However, in comparison to routine laboratory diagnostic procedures which could produce accurate and valuable diagnostic results in a matter of minutes or hours at low cost, it represents quite an expensive, slow and inefficient diagnostic tool. For reliable diagnostic application, besides slides, reagents and microarray reader, it requires sophisticated computational assistance and at least three replicates of a single experiment. These properties make the technology unsuitable for routine diagnostic work.

Nevertheless, diagnostic application of gene expression microarray technology could be considered. One can use gene expression analysis for differentiation between closely related clinical states, i.e. as a tool for establishment of diagnosis in some extremely complex cases. In addition, a gene expression dataset obtained during a clinical experiment based on a control and test group of participants can be applied for selection of a subset of genes relevant for identification and/or differentiation of examined clinical conditions. Quantitative analysis of RNA or protein products of these genes in body fluids or tissues is a much easier, cheaper, faster and more reliable alternative to gene expression profiling. Selection of relevant genes and evaluation of clinical usefulness of the determination of corresponding RNA and protein products provide the framework for application of gene expression profiling for new biomarker screening purposes.

Gene expression results obtained during well-designed clinical experiments represent a rich source of data related to examined clinical condition. In fact, data matrices obtained in such experiments are too rich with information. They contain hundreds of thousands of numerical results that make them too complex for simple visual inspection and analysis. In such settings, the selection of a few relevant genes from a set of tens of thousands of genes represents challenging dimensionality reduction task. Therefore, computational support is needed.

Computational implementations of machine and statistical learning methods called filters could be used for such tasks. In case of monogenetic conditions/diseases, conventional statistical tools like ANOVA, t-test and their nonparametric counterparts with or without modifications could be applied (1, 2). Even in cases characterized by independent change in a few genes, these tools can be reliably used. However, in most instances analyzed clinical conditions are consequences of highly interdependent multigenetic changes. In such instances multivariate computational and/or statistical methods represent the only appropriate tool for selection of relevant genes or diagnostic assessment (3). Sometimes the complexity of relationship between a certain clinical condition and

vima čak niti uspostavljanje pouzdanoga kvantitativnog, prediktivnog modela zasnovanoga na cjelokupnim profilima genske ekspresije nije moguće. To je osobito slučaj u situaciji kada se rezultati mjerenja genske ekspresije na mikropostroju koriste u svrhu prognoze (5) ili u slučajevima kada je analizirani skup sudionika premalen i heterogen. Ipak, u mnogim situacijama multivarijatni filtri mogu odabrati skup gena koji predstavljaju dobre kandidate za biomarkere za određena klinička stanja.

Nažalost, broj multivarijatnih filtra nije baš velik. Osim na uzajamnoj informaciji zasnovanih pristupa (6,7,8), određeni alati za statističko i strojno učenje/predviđanje mogu se iskoristiti za multivarijatni odabir relevantnih gena, tj. za smanjenje dimenzionalnosti. Računalni hibridi koji uključuju artificijelne genetičke algoritme i neke alate za učenje poput strojeva potpornih vektora (engl. *support vector machines*) predstavljaju najčešće korištene alate za multivarijatni odabir svojstava i postavljanje dijagnoze (9).

Breiman je relativno nedavno razvio jedan drugi multivarijatni filtar/prediktivni alat, poznat pod nazivom slučajna šuma (10,11,12). Ta je metoda ukorijenjena u starijoj metodi poznatoj pod nazivom stabla klasifikacije i regresije (engl. *Classification And Regression Tress,* CART) (13). U odnosu na CART, RF uvodi randomizaciju (14). Uzorci i geni se randomiziranim postupkom dijele vrlo mnogo puta. Prilikom svake podjele skupa objekata na podskup za učenje i podskup za provjeru znanja razvija se pojedinačno stablo odluke iz slučajno izabranog podskupa gena koje predstavlja najprikladnije rješenje u pogledu točnosti predviđanja kliničkog stanja. Kvaliteta svake pojedine točke grananja (gena) se analizira rerandomizacijom, a dobiveni rezultati u obliku Ginijeve mjere koriste se za rangiranje gena (10,11). Osim toga, ansambl sastavljen od pojedinačnih stabala odluke može se koristiti za postavljanje dijagnoze zasnovane na cjelokupnom profilu genske ekspresije. Detaljan opis RF metodologije može se naći u navedenim publikacijama.

Dok je prediktivna kvaliteta RF-a opsežno evaluirana te je prikladnost navedene metode za prediktivne svrhe dokazana (10,11,12), prva primjena RF-a u svrhu rangiranja gena tek je nedavno objavljena (15). Iz toga se može zaključiti da primjena RF-a u svrhu odabira kandidata za biomarkere još uvijek nije podrobno ispitana. Cilj ove studije je evaluacija prikladnosti RF za izbor biomarkera na osnovi skupova profila genske ekspresije.

Materijali i metode

Tri dobro opisana skupa profila genske ekspresije preuzeta iz literature izabrani su za analizu odabira relevantnih gena zasnovanog na RF. Poimenično, skupovi "AML/ALL", "Meduloblastom" i "Karcinom kolona" su izabrani na početku ove studije. Ti su se skupovi uvriježili kao testni skupovi za prediktivne alate, kao i testni skupovi za odabir relevantnih gena (5,16,17 i u njima citirane publikacije). Svi

corresponding gene expressions disables dimensionality reduction (4). In some cases even the establishment of a reliable quantitative, predictive model based on acomplete set of gene expressions is impossible. This is especially the case when microarray results are used for prognostic purposes (5) or in cases when analyzed dataset of participants is small and heterogeneous. However, application of multivariate filters could in many instances produce a set of genes that represent good candidate biomarkers of a certain clinical condition.

Unfortunately, the number of multivariate filters is not very large. Besides mutual information based approaches (6,7,8), certain statistical and machine learning/prediction tools could be used for multivariate selection of relevant genes, i.e. dimensionality reduction. Computational hybrids that incorporate artificial genetic algorithms and some learning tools, like support vector machines, represent the most frequently used multivariate feature selection and diagnostic assessment tool (9).

Another multivariate filter/prediction tool known as random forests has been relatively recently introduced by Breiman (10,11,12). This method is rooted in an older method known as classification and regression trees (CART) (13). In comparison to CART, RF introduces randomization (14). Samples and genes are randomly partitioned many times. On each partition of a set of objects into a learning and test set, a single decision tree is grown based on a randomly selected subset of genes that represents the most suitable solution in terms of predictive accuracy for a clinical condition. Quality of each branching point (gene) of a tree is analyzed by rerandomization and obtained results in terms of Ginis' measure are used for gene ranking (10,11). Besides, an ensemble composed of all decision trees could be used for diagnostic assessment based on complete gene expression profiles. Detailed description of RF methodology could be found in the list of publications.

While predictive quality of RF has been extensively validated and RF suitability for predictive purposes has been proven (10,11,12), the first application of this method in gene ranking has been only recently published (15). Therefore biomarker candidate selection based on application of RF has not been thoroughly analyzed. The aim of this study was to evaluate suitability of RF for biomarker selection from gene expression profile datasets.

Materials and methods

Three well-described gene expression profile datasets taken from literature were chosen for analysis of RF-based relevant gene selection. Namely, AML/ALL, medulloblastoma and colon cancer datasets were initially selected. These datasets have been established as benchmarks for prediction tools as well as benchmarks for relevant gene selection (5,16,17 and publications cited therein). All datasets were collected during the clinical experiments that

su skupovi prikupljeni tijekom kliničkih pokusa koji su uključivali dvije skupine sudionika s dva, u određenom smislu suprotna klinička stanja. U danim okolnostima postavljanje dijagnoze/prognoze predstavlja dvoklasni prediktivni problem za svaki od izabranih skupova.

Skup "AML/ALL" je sastavljen od 72 profila genske ekspresije koji sadrže 7129 genskih ekspresija bolesnika koji su bolovali od akutne mijeloične (25 uzoraka) ili akutne limfatične leukemije (47 uzoraka) i taj skup predstavlja problem klasifikacije kliničkog stanja. Skup je izabran jer su molekularna osnova razvoja bolesti i rutinska dijagnostička diferencijacija između odabranih bolesti poznati vrlo detaljno (18). Takvo znanje omogućava evaluaciju postupka za odabir relevantnih gena. Preostala dva skupa su izabrani jer predstavljaju različite tipove dijagnostičkih problema i različite stupnjeve prediktivne složenosti. Skup "Karcinom kolona" se sastoji od 62 profila genske ekspresije koji sadrže 2000 genskih ekspresija po bolesniku. Sudionici su ili bolesnici koji boluju od karcinoma kolona (22 uzorka) ili zdravi pojedinci (40 uzoraka). Skup "Meduloblastom" predstavlja najsloženiji slučaj u kojem se traži prognoza bolesti. Taj je skup sastavljen od 60 uzoraka uzetih od bolesnika koji boluju od meduloblastoma i čije se preživljavanje prati kroz određeni period. Skup sadrži 7129 genskih ekspresija po uzorku. Za podroban opis odabranih skupova pogledati članak koji su publicirali Mukherjee i suradnici te publikacije citirane u tom članku (5).

Kroz cijeli odlomak navodi se i složenost uz ostala svojstva skupova, iako je predmet članka smanjenje dimenzionalnosti, a ne predviđanje. Treba naglasiti da ako se ispravan prediktivni model zasnovan na cjelokupnom profilu genskih ekspresija ne može pronaći, tada smanjenje dimenzionalnosti nema smisla. S druge strane, u određenim situacijama kada se razumno smanjenje dimenzionalnosti ne može postići ispravni prediktivni modeli zasnovani na cjelokupnom profilu genskih ekspresija mogu biti korisni. Iz tih su razloga svojstva vezana uz generalizaciju/predviđanje navedena u prethodnom tekstu.

U ovoj je studiji primijenjena provedba RF u statističkom jeziku R 2.2.0. (11,19). Osim broja stabala koji je postavljen na vrijednost 30 000, sve ostale, od korisnika prilagodljive varijable su postavljene na njihove predefinirane vrijednosti tijekom cijele studije. Povećani broj stabala omogućava bolje rangiranje genskih ekspresija bez ugrožavanja generalizacijskih svojstava (10). Svi su proračuni provedeni na osobnom računalu s operativnim sustavom Windows™. Dijagram toka prikazan na slici 1 daje slijed postupaka od elementarnog značenja za ovu studiju:

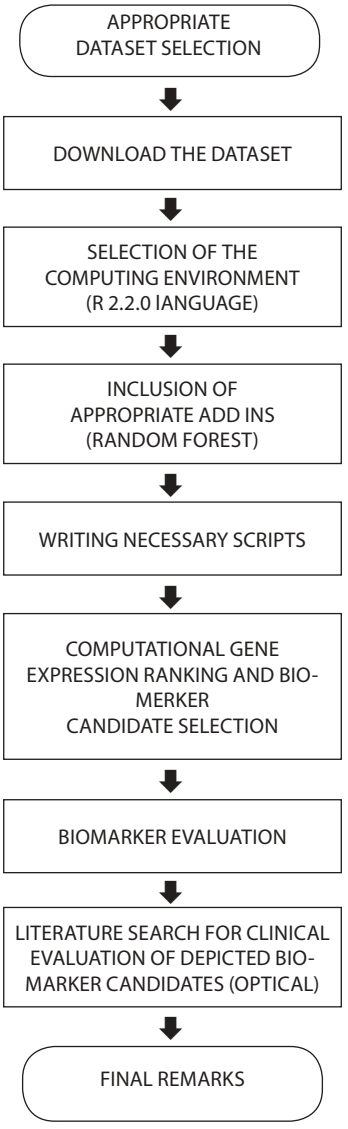
Na kraju ovog odlomka potrebno je razjasniti neke terminološke razlike. Kroz cijeli tekst koji slijedi izrazi "točnost" i "dijagnostička učinkovitost" se koriste kao sinonimi. Prvi izraz koristi računalna, a drugi biomedicinska zajednica za istu veličinu. U slučajevima kada se koristi izraz "prediktivna točnost" on označava točnost izračunatu na odvoje-

involved two groups of participants characterized by, in a certain sense, opposite clinical conditions. Under given circumstances diagnostic/prognostic assessment represents two-class prediction problem for all datasets.

AML/ALL dataset was composed of 72 gene expression profiles consisting of 7,129 gene expressions of patients suffering from acute myeloid (25 samples) or acute lymphatic leukemia (47 samples) and it represented a problem of clinical condition classification. This set was chosen since the molecular basis of disease development and routine diagnostic differentiation between selected diseases have been captured in great detail (18). Such knowledge enables evaluation of relevant gene selection procedure. The other two datasets were selected since they represent different types of diagnostic problems and different levels of prediction complexity. Colon cancer dataset consisted of 62 gene expression profiles that contained 2,000 gene expressions per participant. Participants were either colon cancer patients (22 samples) or apparently healthy subjects (40 samples). Medulloblastoma dataset represented the most complex case where disease prognosis was sought. This set was composed of 60 samples taken from patients suffering from medulloblastoma whose survival was monitored for a certain period of time. 39 patients survived, while 21 patients died during the monitoring. The dataset contained 7,129 gene expressions per sample. For detailed description of selected datasets, see a paper published by Mukherjee et al. and publications cited therein (5).

Throughout this section, prediction complexity is stated along with other dataset properties, although dimensionality reduction rather than prediction represents the subject of the article. It should be stated that if a valid predictive model based on a complete gene expression profile could not be established, dimensionality reduction based on an invalid model would not make much sense. On the other hand, in certain cases when reasonable dimensionality reduction could not be made, valid predictive models based on a complete gene expression profile are still useful. For these reasons dataset properties related to generalization/prediction are also included in the text. R 2.2.0. statistical language (19) implementation of RF was used in this study (11). Besides the number of trees which was set to
30 000, all other user-defined variables were set to their default values throughout the study. Increased number of trees enabled better gene expression ranking without compromising generalization properties (10). All calculations were executed on a Windows™ driven personal computer. Flowchart in Figure 1 provides a sequence of steps essential for this study:

At the end of this section some terminology discrepancies should be cleared out. Throughout the text hereinafter the terms "accuracy" and "diagnostic efficacy" are used interchangeably. The first term is used by computa-



Slika 1. Organizacija studije

Figure 1. Study organization

nom, tj. nezavisnom skupu. Prediktivna točnost je stoga parametar validacije matematičkog modela koji opisuje kvalitetu predviđanja umjesto kvalitete opisa koja se kvantificira točnošću izračunatom na istom onom skupu na kojem je provedeno učenje.

Rezultati

RF-analiza izabranih skupova profila genske ekspresije je provedena prema opisanim eksperimentalnim postavkama. 30 najviše rangiranih gena dobivenih za skup AML/ALL navedeno je u Tablici 1.

Nasuprot genetičkim algoritmima ili nekim filtrima zasnovanim na uzajamnoj informaciji koji daju podskupove relevantnih gena s ograničenim brojem članova, genski

tional community while biomedical community uses the second term for the same quality. In cases when the term "predictive accuracy" is used, separate, i.e. independent test set accuracy is calculated. Predictive accuracy is therefore a mathematical model validation parameter that describes predictive quality instead of descriptive quality that is quantified by accuracy obtained on the same set which was used for learning.

Results

RF analysis of chosen gene expression profile datasets was conducted according to described experimental settings. 30 highest-ranking genes obtained for AML/ALL dataset are listed in Table 1.

Tablica 1. 30 najviše rangiranih gena dobivenih primjenom RF na skup "AML/ALL".

Table 1. 30 highest-ranking genes obtained by application of RF on AML/ALL dataset.

Gene	Gene product function (20)
CD33	Myeloid cell line specific antigen
Microsomal glutathione S-transferase	Leukotriene C4 synthesis; characteristic of polymorfonuclear cells
Zyxin	Cell adhesion mediator
Cystatin C	Inhibitor of cystein proteases
Transcription factor 3	Protein synthesis
Amyloid beta precursor-like protein 2	Promotes transcriptional activation; makes amyloid plaques
Cyclin D3	Cell cycle control
Cystatin A	Inhibitor of cystein proteases
D component of complement	Component of alternative complement pathway
Terminal transferase (TdT)	DNA polymerase; early lymphocyte marker
Non-erythrocytic spectrin	Cytoskeletal protein
CD63	Cell adhesion mediator
Properdin P complement factor	Component of alternative complement pathway
Proteasome iota chain	Protein degradation
Rho G	Ras homolog gene family
Azurocidin	Antibiotic protein from granulocyte azurophilic granules
Granulin	Cell growth
Lambda immunoglobin light chain	Acquired immune response
Clusterin	Protein whose role has been associated with apoptosis
MB-1 (CD 79a)	Part of B lymphocyte antigen receptor
Cathepsin D	Lysosomal aspartyl protease involved in proliferation and mitosis
Macmarcks	Actin filament crosslinking protein involved in cell motility, mitosis and phagocytosis
Myeloperoxidase	Oxydative burst in myeloid cells
Myosin light chain	Part of cytoskeleton
Fumarylacetoacetase	Tyrosine catabolism pathway
Topoisomerase II beta	DNA replication
Proteoglycan 1	Hematopoietic cell granule proteoglycan
Protective protein for beta-galactosidase	Glycoprotein which associates with lysosomal enzymes beta-galactosidase and neuraminidase
Immunoglobulin-associated beta (CD79b)	Part of B lymphocyte antigen receptor
Leukotriene C4 synthase	Leukotriene C4 synthesis; characteristic of polymorfonuclear cells

odabir zasnovan na RF rangira sve gene polazišnog skupa. Ipak, samo najviše rangirani geni predstavljaju one gene čija ekspresija čini najznačajniju razliku između analiziranih kliničkih stanja. U studiji AML/ALL izabrano je 30 najviše rangiranih gena. Ovaj broj predstavlja manje od 0,5% broja gena polazišnog skupa, što je značajno smanjenje dimenzionalnosti.

In contrast to genetic algorithm or some of mutual information based filters that provide relevant gene subsets with restricted number of elements, RF based gene selection ranks all genes contained in the starting set. However, only the highest-ranking genes represent those genes whose expression is the most relevant for the distinction between analyzed clinical conditions. In AML/ALL study,

Među ostalim genima prikazanima u Tablici 1 mogu se naći CD33, TdT i mijeloperoksidaza. Proteinski produkti ovih gena su dobro poznati imunokemijski ili citokemijski biomarkeri za diferencijaciju AML/ALL (18). Osim prethodno navedenih citokemijskih biomarkera specifična i nespecifična esteraza se rutinski koriste za diferencijaciju akutnih mijeloičnih i limfatičnih leukemija. Nažalost, polazišni skup gena ne sadrži gene koji odgovaraju ovim enzimima. Među serumskim enzimima lizozim se često koristi za diferencijaciju AML/ALL (21). Odgovarajući gen je od strane RF rangiran na 250. poziciju. Osim CD33, CD10 i CD13 se ponekad koriste za imunokemijsku diferencijaciju AML i ALL. Odgovarajući geni su rangirani na 542. i 392. poziciju. Geni koji kodiraju lizozim, CD10 i CD13 su postavljeni među 10% najviše rangiranih gena. Kako tijekom izbora relevantnih gena zasnovanog na RF iz AML/ALL skupa nije bilo uplitanja ili manipulacije, može se reći da odabrani filtar prepoznaje najznačajnije detalje koji razlikuju AML i ALL, tj. primjena odabranog filtra je uz navedene eksperimentalne postavke prikladna za odabranu namjenu.

Osim poznatih biomarkera generiran je i velik broj novih kandidata. Među ostalima je izabran i cistatin C. Ovaj biomarker za diferencijaciju AML/ALL je nedavno prepoznat i od strane drugih autora (22). Razvijena kvantitativna metoda PCR za određivanje cistatina C je dala ohrabrujuće rezultate u odnosu na diferencijaciju odabranih tipova leukemije. Ovi rezultati donose daljnju potvrdu prikladnosti primjene RF za probiranje biomarkera na osnovi danog skupa mikropostroja. Ipak, ovi preliminarni rezultati nove dijagnostičke indikacije za mjerenje genske ekspresije cistatina C zahtijevaju daljnju evaluaciju. Od osobitog bi značenja bilo ispitivanje prikladnosti određivanja serumske koncentracije cistatina C u svrhu razlikovanja AML i ALL.

Konačno, većina gena nabrojenih u Tablici 1 su također prepoznati od strane drugih autora (7,8,23) i većina njih je koristila univarijatne filtre. Ova činjenica ukazuje da se problem razlikovanja AML/ALL može svesti na univarijatni ili linearno razdvojivi problem (Slike 2a i 2b).

Prije svega, Slike 2a i 2b jasno prikazuju prikladnost odabranih dvodimenzionalnih klasifikatora leukemija. Slika 2a predstavlja slučaj u kojem se uzorci mogu linearno odvojiti. Čak se može postaviti i univarijatna klasifikacija AML i ALL na osnovi pojedinačne koordinate uz razumnu točnost (96%). Slika 2b je još dojmljivija. Bilo koja koordinata (TdT, glutation S-transferaza) se uz visoku pouzdanost može iskoristiti kao univarijatni klasifikator AML i ALL (točnost za univarijatnu, na glutation S-transferazi zasnovanu klasifikaciju iznosi 94%). Ovi pronalasci daju daljnju potvrdu prikladnosti izbora gena zasnovanog na RF za problem diferencijacije AML i ALL. Štoviše, takvi univarijatni klasifikatori su gotovo savršeni kandidati za rutinsku dijagnostičku diferencijaciju AML i ALL jer omogućavaju diferencijaciju zasnovanu na pojedinačnom kvantitativnom mjerenju s PCR. Ipak, ove slike naznačuju da se problem

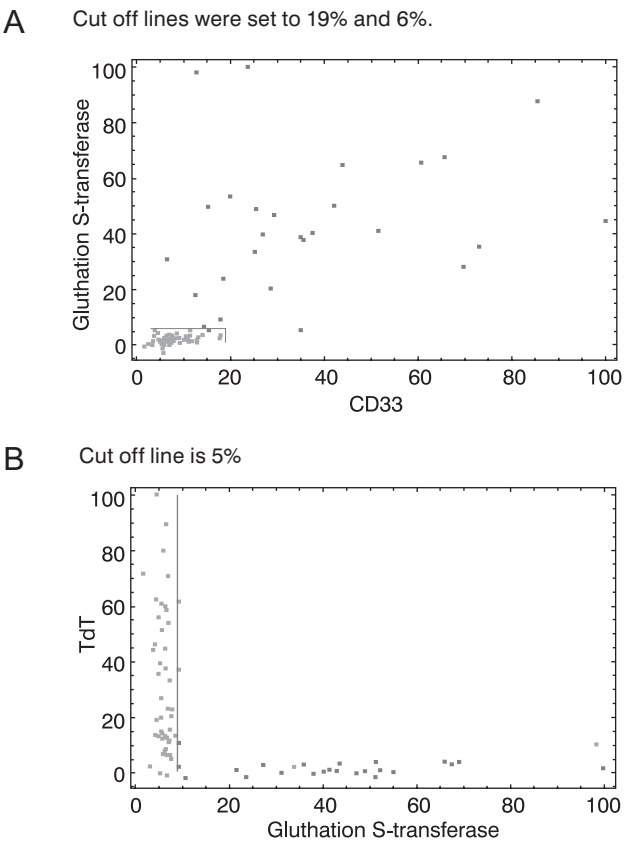
30 highest-ranking genes were selected. This number represents less than 0.5% of the starting number of genes, which is a considerable dimensionality reduction.

Among other genes depicted in Table 1, CD33, TdT and myeloperoxidase could be found. Protein products of these genes are well-known immunochemical or cytochemical biomarkers for AML/ALL differentiation (18). Besides aforementioned cytochemical biomarkers, specific and nonspecific esterase have been routinely used for differentiation of acute myeloid and lymphatic leukemias. Unfortunately, the starting set of genes did not contain genes corresponding to these two enzymes. Among serum enzymes, lysozyme has been frequently used for AML/ALL differentiation (21). A corresponding gene was ranked by RF at the 250th position. Besides, CD33 CD10 and CD13 are sometimes used for immunochemical differentiation of AML and ALL. Corresponding genes have been ranked at the 542nd position and 392nd position, respectively. The genes that code for lysozyme, CD10 and CD13, have been placed among top 10% genes. Since there was no involvement or manipulation during the process of RF-based relevant gene selection from AML/ALL dataset, it could be stated that the selected filter captured the most important details that discriminate between AML and ALL, i.e. application of the selected filter is suitable for a given purpose under provided experimental conditions.

Besides established biomarkers, RF-based selection generated a large number of new candidates. Among others, cystatin C was selected. This biomarker for AML/ALL differentiation was recently recognized by other authors (22). The quantitative PCR method developed for determination of cystatin C provided encouraging results regarding differentiation of selected types of leukemia. These results bring further evidence of the suitability of biomarker screening based on application of RF on a given microarray dataset. However, these preliminary results of new diagnostic indication of cystatin C expression measurement need further evaluation. It would be especially important to examine the suitability of serum cystatin C concentration determination for AML/ALL discrimination.

Finally, most of the genes listed in Table 1 were found by other authors (7,8, 23), and most of them used univariate filters. This fact suggests that AML/ALL problem could be reduced to a univariate or linearly separable problem (Figures 2a and 2b).

First of all, Figures 2a and 2b clearly show suitability of the selected two-dimensional differentiation between leukemias. Figure 2a presents the case where samples could be linearly separated. Even the univariate classification of AML and ALL patients could be established based on a single coordinate with reasonable accuracy (96%). Figure 2b is even more striking. Any of the coordinates (TdT, glutathione S-transferase) could be used as a univariate AML/ALL classifier with considerable certainty (accuracy



Slika 2. Razdvajanje uzoraka iz skupa "AML/ALL" u prostoru razapetom s ekspresijama CD33 i glutation S-transferaze (a) ili glutation S-transferaze i TdT (b). Uzorci AML su predstavljeni tamnijim točkama, dok su uzorci ALL predstavljeni svjetlijim točkama. Koordinate - genske ekspresije su podijeljene na ljestvici. Najveće mjerene genske ekspresije su postavljene na 100%. CD33, glutation S-transferaza i TdT su preuzete iz liste najviše rangiranih gena (Tablica 1).

Figure 2. Separation of samples from AML/ALL dataset in the space spanned by CD33 and glutathione S-transferase (a) or glutathione S-transferase and TdT (b) expressions. AML samples are represented by dark dots, while ALL samples are represented by light dots. Gene expression coordinates are scaled. The highest measured gene expression values were set to 100%. CD33, glutathione S-transferase and TdT were taken from the highest-ranking gene list (Table 1.).

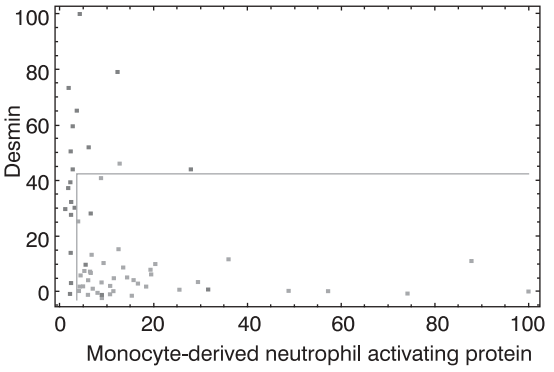
smanjenja dimenzionalnosti skupa AML/ALL može jednostavno riješiti univarijatnim filtrima.

Na ovom mjestu treba istaknuti da su tijekom validacije prediktivnih modela za klasifikaciju AML/ALL zasnovanu na cjelokupnom profilu genskih ekspresija mnogi autori postigli gotovo savršene rezultate. U svim značajnim slučajevima prediktivna je točnost bila 90% ili više. Ova činjenica ukazuje da bi se prikladnost bilo kojeg filtra, uključujući i multivarijatni filtar zasnovan na RF, trebala evaluirati na zahtjevnijim klasifikacijskim problemima. Skupovi karcinom kolona (Slika 3) i meduloblastom (Slika 4) predstavljaju takve probleme (5).

Već je na prvi pogled očigledno da visokokvalitetno razdvajanje analiziranih klasa u prostorima razapetim primjenom dvaju najviše rangiranih gena nije lako postići, osobito u slučaju prognoze meduloblastoma. Ovo je u suglasju s rezultatima koje je objavio Mukherjee sa suradnicima (5).

for univariate, glutathione S-transferase based classification is 94%). These findings provide further evidence for correctness of the RF-based gene selection approach to AML/ALL differentiation problem. Moreover, such univariate classifiers are almost perfect candidates for routine diagnostic AML/ALL differentiation since they enable the differentiation based on a single quantitative PCR measurement. However, these figures imply that the AML/ALL dimensionality reduction problem could be easily solved by univariate filters.

It should be stated at this point that many authors obtained almost perfect results during the validation of predictive models for AML/ALL classification based on a complete gene expression profile. In all relevant cases, predictive accuracy was 90% or higher. This fact suggests that suitability of any filter including RF-based multivariate filter should be evaluated on more demanding classification



Slika 3. Skup "Karcinom kolona". Tamnije točke predstavljaju bolesnike koji boluju od karcinoma kolona, a svjetlije točke predstavljaju zdrave pojedince. Obje koordinate (ekspresije monocitnog proteina koji aktivira neutrofile i dezminskog gena) su podijeljene na na interval 0 - 100%. Linije graničnih vrijednosti postavljene su na 3 i 45%.

Figure 3. Colon cancer dataset. Dark dots represent colon cancer patients and light dots stand for apparently healthy individuals. Both coordinate axes (monocyte-derived neutrophyl activating protein and desmin gene expressions) are scaled to 0 - 100% interval. Cut off lines were set to 3 and 45%, respectively.

Slika 3 predstavlja tipično multivarijatno rješenje za dani dijagnostički problem. Pažljivim izborom graničnih vrijednosti na obje koordinate bolesnici s karcinomom kolona mogu se razlikovati od zdravih pojedinaca uz prihvatljivu točnost (92%). Ovaj rezultat podrazumijeva da su najmanje dva najviše rangirana gena potrebna za rutinsku dijagnostičku primjenu. U svrhu poboljšanja dijagnostičke učinkovitosti mogla bi se razmotriti primjena tri ili više visoko rangiranih genskih ekspresija. Ipak, povećanjem broja mjerenja genskih ekspresija potrebnih za postavljanje dijagnoze ovaj pristup postaje neprikladan za rutinsku dijagnostičku primjenu.

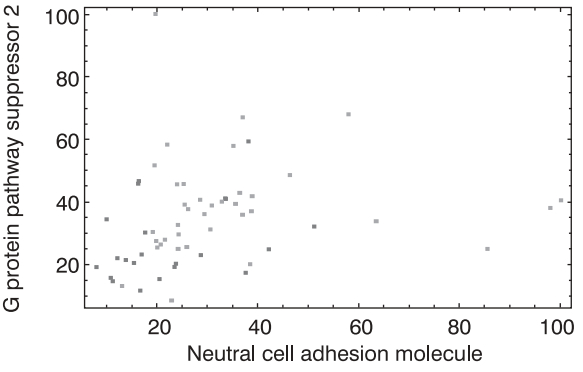
Konačno, Slika 4 prikazuje da korisno smanjenje dimenzionalnosti ne može biti provedeno u svim slučajevima. Praktično je nemoguće razlikovati bolesnike koji su preživjeli tijekom terapije meduloblastoma od bolesnika koji su umrli (Slika 4). To podrazumijeva da je uključivanje tri ili više genskih ekspresija i/ili uspostavljanje nelinearnih veza između genske ekspresije i preživljavanja potrebno za točno postavljanje dijagnoze. Taj problem ukazuje da rangiranje gena zasnovano na RF u svrhu izbora biomarkera preživljanja tijekom terapije meduloblastoma nije prikladan korak. U danim eksperimentalnim uvjetima, koji uključuju malu skupinu heterogenih bolesnika/tretmana, jedini je pristup primjena tehnologije mikropostroja uz adekvatni prediktivni alat poput RF. Ova činjenica predstavlja prednost alata poput RF koji se mogu koristiti kao multivarijatni filtar, a u isto vrijeme su prediktivni alati koji se mogu koristiti za postavljanje dijagnoze/prognoze na temelju cjelokupnog profila genske ekspresije. Ovim je

problems. Colon cancer (Figure 3) and medulloblastoma (Figure 4) datasets represent such problems (5).

It is obvious at first glance that high quality separation of analyzed classes in the spaces spanned by two highest-ranking genes is not easy to establish, especially in medulloblastoma prognosis case. This is in agreement with results published by Mukherjee et al. (5).

Figure 3 presents typical multivariate solution for a given diagnostic problem. By careful selection of cut off values on both coordinates colon cancer patients could be differentiated from apparently healthy individuals with acceptable accuracy (92%). This result implies that expressions of at least two highest-ranking genes is needed for routine diagnostic utility. In order to improve diagnostic efficacy, application of three or more highly ranked gene expressions could be evaluated. However, by raising the number of gene expression measurements necessary for diagnostic assessment, this approach becomes inappropriate for routine diagnostic application.

Finally, Figure 4 shows that useful dimensionality reduction cannot be made in all cases. Medulloblastoma survivors and non-survivors are almost indistinguishable based on selected two-dimensional coordinate system (Figure 4). This implies that inclusion of three or more gene expressions and/or establishment of nonlinear relationship between gene expressions and survival is a necessary prerequisite for accurate prognostic assessment. This problem shows that RF-based gene expression ranking for a biomarker of medulloblastoma survival selection is not an appropriate step. Under given experimental con-



Slika 4. Skup "Meduloblastom". Tamnije točke predstavljaju bolesnike koji su umrli tijekom perioda praćenja, a svjetlije točke predstavljaju one bolesnike koji su preživjeli. Obje koordinate (ekspresija gena neuralne adhezijske molekule i supresora 2 G-proteinskog puta) su podijeljene na interval 0 - 100%.

Figure 4. Medulloblastoma dataset. Dark dots represent patients who died during monitoring period and light dots stand for survivors. Both coordinate axes (neural adhesion molecule and G protein pathway suppressor 2 gene expressions) are scaled to 0 - 100% interval.

alatima svojstveno još nešto: validacija modela. Povećavanjem složenosti odnosa profila genske ekspresije i biološke varijable koju ispitujemo korisnost odabranih biomarkera opada. Stoga adekvatna validacija modela, koja predstavlja neizravnu mjeru složenosti analiziranog problema, daje uvid u korisnost smanjenja dimenzionalnosti. U skladu s ovom činjenicom, prihvatljivi rezultati validacije modela zasnovanog na cjelokupnom profilu genske ekspresije koji je dobiven primjenom RF-a ili sličnog alata tvore osnovu za daljnje smanjenje dimenzionalnosti i probiranje biomarkera. U većini slučajeva to je dostižan cilj. Ipak, ukoliko korisno smanjenje dimenzionalnosti nije moguće, validirani model dobiven primjenom RF zasnovan na cjelokupnom profilu genske ekspresije preostaje kao alternativa.

Zaključci

Prije bilo kakvog pokušaja izbora kandidata za biomarkere iz cjelokupnog profila genske ekspresije, prikladnost prediktora, u ovom slučaju modela dobivenog primjenom RF, i/ili stupanj složenosti odnosa između profila genske ekspresije i dijagnoze/prognoze treba biti poznat. Ako je moguće pronaći pouzdan model zasnovan na cjelokupnom profilu genske ekspresije može se pristupiti izboru kandidata za biomarkere.

U slučaju jednostavnih problema RF se pokazao kao koristan filtar. Na osnovi ovog pristupa identificirani su dobro poznati univarijatni klasifikatori dvaju tipova leukemije. Štoviše, otkrivene su mnoge nove i obećavajuće alternative. U nešto složenijem slučaju identifikacije karcinoma koditions, which include a small group of heterogeneous patients/treatments, the only approach is application of microarray technology equipped with an adequate prediction tool like RF. This fact represents the advantage of tools like RF that could be used as a multivariate filter and, at the same time, they are prediction tools that could be used for prognostic/diagnostic assessment based on a complete gene expression profile. These tools are characterized by one more important property: model validation. By increasing the complexity of relationship between gene expression profile and biological variable under examination, the utility of selected biomarkers declines. Therefore adequate model validation, which provides indirect measure of analyzed relationship complexity, gives insight in usefulness of dimensionality reduction. According to this fact, acceptable validation results obtained by RF or a similar tool based on a complete gene expression profile provide the basis for further dimensionality reduction and biomarker screening. In most cases, this is an achievable task. Still, if useful dimensionality reduction is not possible, validated RF model based on complete gene profile remains as an alternative.

Conclusions

Before any attempt to select biomarker candidates from a complete gene profile is made, the suitability of a predictor, in this case RF-derived model, and/or degree of relationship complexity between gene expression profile and diagnosis/prognosis should be established. If a reliable model can be obtained based on a complete gene set,

lona pojavio se dvodimenzionalni klasifikator zasnovan na genskoj ekspresiji dva najviše rangirana gena dobivena primjenom RF. Predloženi klasifikator i pripadajuće granične vrijednosti pokazali su obećavajuće klasifikacijske mogućnosti. Konačno, primjena RF u najsloženijem slučaju prognoze tijeka liječenja meduloblastoma nije dala korisne kandidate za biomarkere. Zbog izrazito složenog odnosa profila genske ekspresije i prognoze (5) i malog, heterogenog skupa razumno smanjenje dimenzionalnosti primjenom RF nije moguća.

Cjelokupne rang-liste gena iz ove studije su dostupne na zahtjev. Pokazano je da su neki novi biomarkeri, poput ekspresije cistatina C primijenjene za klasifikaciju AML/ALL, ušli u proces dijagnostičke evaluacije. Ipak, enzim-imunotestovi, imunonefelometrija, imunoturbidimetrija i protočna citometrija, nasuprot kvantitativnom PCR-u su prisutni u većini rutinskih kliničkih laboratorija. Bilo bi zanimljivo provjeriti mogu li neki od proteinskih produkata relevantnih gena biti korisni biomarkeri ispitivanih kliničkih stanja.

Zahvale

Autor se želi zahvaliti Ana-Mariji Šimundić na kritičnom osvrtu i korisnim savjetima.

Adresa za dopisivanje:

Željko Debeljak
Odjel za medicinsku biokemiju
Klinička bolnica Osijek
J. Huttlera 4
31 000 Osijek
Tel: +385 (0) 31 511 660
e-pošta: *debeljak.zeljko@kbo.hr*

the selection of biomarker candidates can be initiated. In case of low complexity problems, RF proved to be a useful filter. Based on this approach, well-established univariate classifiers of two leukemia types were identified. Moreover, many new promising alternatives were revealed. In a bit more complex case of colon carcinoma, identification was possible of a two-dimensional classifier based on top two gene expressions obtained by RF. The proposed classifier, along with proposed corresponding cut off values, showed promising classification capabilities. Finally, application of RF on the most complex case of establishing medulloblastoma prognosis did not yield useful biomarker candidates. Due to the highly complex relationship between gene expression profiles and the prognosis (5) and the small and heterogeneous set, reasonable dimensionality reduction by application of RF was not possible.

Complete gene ranking lists from this study for all examined datasets are available on request. It has been shown that some of the new biomarkers, like cystatin C gene expression applied for AML/ALL classification, have already entered diagnostic evaluation. However, enzyme immunoassay, immunonephelometry and immunoturbidimetry and flow citometry are, unlike-quantitative PCR, already present in most of the routine clinical laboratories. It would be interesting to evaluate whether some of relevant gene protein products could be useful as biomarkers of examined clinical conditions.

Acknowledgements

Author would like to thanks Ana-Mariji Šimundić for critical reading of manuscript and usefull suggestions.

Corresponding author:

Željko Debeljak
Department of medical biochemistry
Osijek University Hospital
J. Huttlera 4
31 000 Osijek
Croatia
Phone: +385 (0) 31 511 660
e-mail: *debeljak.zeljko@kbo.hr*

Rječnik računalnih izraza korištenih u ovoj studiji

Izraz	Značenje
"Nezavisna" varijabla/svojstvo	U ovom kontekstu - gen
"Zavisna" varijabla	U ovom kontekstu – kliničko stanje
Objekt	U ovom kontekstu – profil genske ekspresije pojedinog uzorka
Strojno i statističko učenje	Skupina računalnih tehnika razvijenih u svrhu rješavanja općih klasifikacijskih i regresijskih problema
Smanjenje dimenzionalnosti	Smanjenje broja varijabli potrebnih za učenje (smanjenje broja gena potrebnih za postavljanje dijagnoze/prognoze)
Stabla klasifikacije i regresije (engl. *classification and regression trees,* CART)	Jedna od tehnika strojnog učenja zasnovana na generiranju specifičnog tipa stabala odluke
Univarijatni klasifikatori	Pristup klasifikaciji objekata, zasnovan na pojedinačnim varijablama (postavljanje dijagnoze/prognoze zasnovano na ekspresiji samo jednog gena)
Filteri	Metode strojnoga i statističkog učenja razvijene u svrhu odabira relevantnih varijabli
Multivarijatni odabir svojstava	Izbor većeg broja relevantnih varijabli koji pridaje značenje i njihovoj međuzavisnosti
Uzajamna informacija	Veličina izvedena iz teorije informacije koja se može koristiti kao kriterij filtriranja
Genetički algoritam	Računalni pristup odabiru svojstava (u ovom kontekstu) zasnovan na oponašanju procesa prirodnog odabira i mutacije
Prediktivni alat	Alat za strojno ili statističko učenje koji omogućava predviđanje vrijednosti zavisne varijable zasnovan na uspostavaljanju kvantitativne veze između zavisne i nezavisnih varijabli
RF	Tehnika strojnog učenja koja kao građevni blok koristi stabla klasifikacije i regresije i koja se može primjeniti kao filtar, ali i kao prediktivni alat
Strojevi potpornih vektora (engl. *support vector machines*)	Jedan od prediktivnih alata
Prediktivna točnost	Udio točnih predviđanja klase na skupu koji nije korišten tijekom učenja
Opisna točnost	Udio točnih predviđanja klase na skupu koji je korišten tijekom učenja
Ispravan model	Prediktivni model čija je prediktivna točnost evaluirana i dokazano prihvatljiva

Dictionary of computational terms used in the study

Term	Meaning
"Independent" variable/feature	In this context, gene
"Dependent" variable	In this context, clinical condition
Object	In this context, gene expression profile of a single sample
Machine and statistical learning	A group of computational techniques developed for solving general classification and regression problems
Dimensionality reduction	Reduction in the number of variables needed for learning (reduction in gene expressions needed for establishment of diagnosis/prognosis)
Classification and regression trees	One of the machine learning techniques based on generation of a specific type of decision trees
Univariate classifiers	Approach to classification of objects, based on a single variable (establishment of diagnosis/prognosis based on the expression of a single gene)
Filters	Machine and statistical learning methods developed for relevant variable selection
Multivariate feature selection	Selection of multiple relevant variables that delineates possible variable interdependence
Mutual information	Quantity derived within information theory that can be used as a filter criterion
Genetic algorithm	Computational approach to feature selection (in this context) based on natural selection and mutation
Prediction tool	Machine or statistical learning tool that enables dependent variable value prediction based on the development of quantitative relationship between a dependent and independent variable
Random Forest	A machine learning technique that uses classification and regression trees as a building block which can be used as a filter, and as a prediction tool at the same time
Support vector machines	One of the machine learning prediction tools
Predictive accuracy	A fraction of correct class predictions obtained on a dataset that was not used during learning
Descriptive accuracy	A fraction of correct class predictions obtained on the same dataset that was used during learning
Valid model	A prediction model whose predictive accuracy has been evaluated and proven to be acceptable

Literatura

1. Brazma A, Vilo J. Gene Expression Data Analysis. FEBS Lett 2000; 480:17-24.
2. Dudoit S, Fridlyand J, Speed T. Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. J Am Stat Assoc 2002;97:77-87.
3. Guyon I, Weston J, Barnhill S, Vapnik V. Gene Selection for Cancer Classification Using Support Vector Machines. Mach Learn 2002;46(1-3):389-422.
4. Guyon I, Elisseef A. An Introduction to Variable and Feature Selection. J Mach Learn Res 2003;3:1157-82.
5. Mukherjee S, Tamayo P, Rogers S, Rifkin R, Engle A, Campbell C, Golub TR, Mesirov JP. Estimating Dataset Size Requirements for Classifying DNA Microarray Data. J Comput Biol 2003;10(2):119-42.
6. Kohavi R, John GH. Wrappers for Feature Selection. Artif Intell 1997;97(1-2):273-324.
7. Bogunović N, Marohnić V, Debeljak Ž. Efficient Gene Expression Analysis by Linking Multiple Data Mining Algorithms. In Proceedings of the 27th Annual International Conference of the IEEE-EMBS; 2005.
8. Su Y, Murali TM, Pavlovic V, Schaffer M, Kasif S. RankGene: Identification of Diagnostic Genes Based on Expression Data. Bioinformatics 2003;19(12)1578–9.
9. Peng S, Xu Q, Ling XB, Peng X, Du W, Chen L, Molecular Classification of Cancer Types from Microarray Data Using the Combination of Genetic Algorithms and Support Vector Machines. FEBS Lett 2003;555:358-62.
10. Breiman L. Random Forests. Mach Learn 2001;45:5-32.
11. Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. J Chem Inf Comput Sci 2003;43:1947-58.
12. Zhang H, Yu CY, Singer B. Cell and tumor classification using gene expression data: Construction of forests. Proc Natl Acad Sci USA 2003;100(7):4168-72.
13. Breiman L, Friedman J, Olshen R, Stone C Classification and Regression Trees. Belmont, USA: Wadswarth; 1984.
14. Lunneborg CE. Data Analysis by Resampling: Concepts and Applications. Pacific Grove, USA: Duxbury; 2000.
15. Díaz-Uriarte R, Alvarez de Andrés S. Gene Selection and Classification of Microarray Data Using Random Forest. BMC Bioinformatics 2006;7:3-16.
16. Wang Y, Tetko IV, Hall MA, Frank E, Facius A, Mayer KFX, Mewes HW, Gene Selection from Microarray Data for Cancer Classification – a Machine Learning Approach. Comput Biol Chem 2005;29:37-46.
17. Lee JW, Lee JB, Park M, Song SH. An Extensive Comparison of Recent Classification Tools Applied To Microarray Data. Comput Stat Data Anal 2005;48:869-85.
18. McKenzie SB. Clinical Laboratory Hematology. Upper Saddle River, USA: Pearson Education, Inc.; 2004.
19. R Development Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2005.
20. http://www.ncbi.nlm.nih.gov/ accessed July 17th 2006.
21. Thomas, L. Clinical Laboratory Diagnostics. Frankfurt, Germany: TH-Books; 1998.
22. Sakhinia E, Faranghpour M, Yin JAL, Brady G, Hoyland JA, Byers RJ. Routine Expression Profiling of Microarray Gene Signatures in Acute Leukaemia by Real-time PCR of Human Bone Marrow. Br J Haematol 2005;130(2):1365-2141.
23. Ben-Dor A, Bruhn L, Friedman N, Nachman I, Schummer M, Yakhini Z, Tissue Classification with Gene Expression Profiles. J Comput Biol 2000;7(3-4):559-83.