STATISTICKÉ METODY V ANALÝZE DAT Z DNA MIKROČIPŮ

STATISTICAL METHODS FOR ANALYSING GENE EXPRESSION MICROARRAY DATA

PAVLÍK T.[1,2],*, JARKOVSKÝ J.[1]

[1] INSTITUT BIOSTATISTIKY A ANALÝZ, PŘÍRODOVĚDECKÁ A LÉKAŘSKÁ FAKULTA MASARYKOVY UNIVERZITYBRNO
[2] KATEDRA APLIKOVANÉ MATEMATIKY, PŘÍRODOVĚDECKÁ FAKULTA MASARYKOVY UNIVERZITY, BRNO

Souhrn
Mohutný rozvoj molekulárně biologických metod v posledních deseti letech je charakterizován produkcí velkého množství dat. Nejinak je tomu i v případě technologie DNA mikročipů, která umožňuje v jednom experimentu sledovat expresi desítek tisíc genů najednou. Kvantum získaných experimentálních dat je však pro relevantní medicínské závěry nutné vhodně analyzovat a interpretovat.
Tento článek je věnován statistickým metodám, které lze pro hodnocení dat získaných z DNA mikročipů použít. Tyto metody lze rozdělit do tří velkých skupin: shlukovací metody, metody pro identifikaci rozdílně exprimovaných genů a klasifikační metody. Shlukovací metody slouží k nalezení homogenních skupin pacientů s podobným expresním profilem nebo skupin genů s podobným chováním, metody pro identifikaci rozdílně exprimovaných genů hledají geny specifické svojí aktivitou pro určitou biologickou tkáň, zatímco klasifikační metody slouží k nalezení diskriminačního pravidla pro přesnou diagnostiku nových pacientů do jedné z definovaných skupin.

Klíčová slova: DNA mikročipy, statistická analýza, shlukování, testování rozdílné exprese, klasifikační metody

Summary
Last decade led to massive progress in the molecular biology methods which was accompanied by the production of large amount of data. This is also the case of the gene expression microarray technology that makes it feasible to study thousands of genes simultaneously. However, for relevant medical inference there is the need for appropriate evaluation and interpretation of this large quantity of experimental data.
This paper is dedicated to statistical methods that can be used for the evaluation of gene expression data. These methods can be split into three main categories: clustering methods, methods for identification of differentially expressed genes and classification techniques. Clustering methods can be used for finding of homogenous groups of patients or genes with similar expression profile, methods for identification of differentially expressed genes find genes specific for activity of certain biological tissue while classification techniques are used for setting up a discrimination rule for precise diagnostics of newly diagnosed patients to one of the previously defined classes.

Keywords: DNA microarrays, statistical analysis, clustering, testing for differential expression, classification techniques

Úvod

Mohutný rozvoj molekulárně biologických metod v posledních deseti letech je charakterizován produkcí velkého množství dat. Nejinak je tomu i v případě technologie DNA mikročipů [1], která umožňuje v jednom experimentu sledovat expresi desítek tisíc genů najednou, což ji řadí mezi nejpoužívanější biologické metody dneška. Kvantum získaných experimentálních dat je však pro relevantní medicínské závěry nutné vhodně analyzovat a interpretovat [2]. K tomu slouží biostatistické metody a postupy, mnohdy neobvyklé a zcela přizpůsobené charakteru dat z DNA mikročipů. Pomocí těchto metod lze dosáhnout takových cílů jako je například identifikace genů specifických svojí aktivitou pro určitou biologickou tkáň nebo genů s rozdílnou expresí mezi dvěma skupinami biologických vzorků, nalezení homogenních skupin pacientů s podobným expresním profilem nebo skupin genů s podobným chováním či nalezení diskriminačního pravidla pro přesnou diagnostiku nových pacientů do jedné z definovaných skupin. Schematické znázornění základních cílů experimentů s použitím DNA mikročipů a statistické metody k jejich dosažení jsou uvedeny na obrázku 1.

Statistická analýza navazuje na fáze analýzy obrazu a normalizace dat a je tak poslední částí microarray experimentu, její výsledky jsou však na těchto i předchozích biologických fázích experimentu zcela závislé ve smyslu jejich kvalitního provedení a tudíž kvalitních vstupních dat. Proto se statistické principy promítají nejen do fáze analýzy obrazu a normalizace dat, ale i do samotné přípravy experimentu ve formě experimentálního designu.

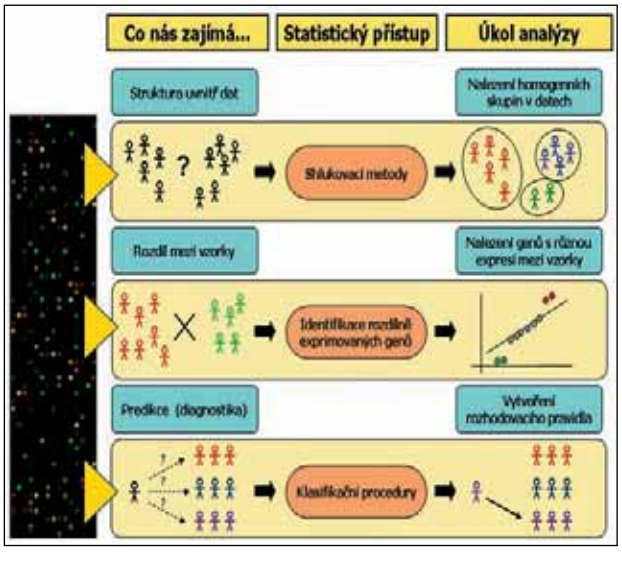
Problémy

Kvůli podstatě DNA mikročipů se statistická analýza tohoto typu dat potýká s nejrůznějšími obtížemi, které někdy nelze pomocí stávajících metod uspokojivě vyřešit, což následně vede k vývoji nových biostatistických postupů specializovaných pro analýzu expresních dat. Zásadní problémy jsou tyto:

1. **Velký počet sledovaných proměnných - genů**. Největší výhoda microarray experimentu je zároveň i jeho největší slabinou, neboť velmi málo statistických metod se umí vypořádat s desítkami tisíc proměnných najednou.
2. **Malá velikost experimentálního vzorku – pacientů**. Studie jsou limitovány jak finanční náročností DNA čipů tak omezeným počtem dostupných pacientů. Většina publikovaných prací je postavena pouze na vzorku v řádu desítek pacientů.
3. **Chybějící hodnoty**. Chybějící hodnoty vznikají na místech podezřelých spotů, které jsou buď manuálně nebo automaticky vyloučeny z další analýzy. Spoty mohou být vyloučeny z více důvodů, např. kvůli prachu a škrábancům na mikročipu, deformovanému obrazu nebo špatnému natištění. Nejjednodušší doplnění chybějících hodnot je metodou řádkového průměru, kdy je chybějící hodnota nahrazena průměrem ostatních expresních hodnot daného genu. Troyanskaya a kol. [3] navrhují použít pro dopočítání chybějící hodnoty metodu *k* nejbližších sousedů, kdy je chybějící hodnota daného genu určena jako průměr odpovídajících hodnot u *k* nejpodobnějších genů. Tento přístup je vhodnější, neboť bere v úvahu funkční vztahy mezi jednotlivými geny.

Statistické metody

Metody používané pro statistickou analýzu dat získaných z DNA mikročipů se dají rozdělit do tří velkých skupin: shlukovací metody, metody pro identifikaci rozdílně exprimovaných genů a klasifikační metody. Cílem shlukovacích metod může být nalezení skupin genů, které vykazují v průběhu experimentu podobné chování, nebo skupin pacientů s podobným expresním profilem. Ke shlukování se používají jak klasické metody (metoda k-průměrů [2], hierarchické metody [4]), tak metody používané zejména v informačních technologiích (self-organizing maps [5]). Do druhé skupiny patří metody, jejichž úkolem je vytipovat podezřelé geny mezi všemi zkoumanými, tedy geny, které by mohly být zodpovědné za různé vlastnosti biologických vzorků. Jedná se zejména o metody násobného testování hypotéz založené na t-statistice a jejích modifikacích. Úkolem klasifikačních metod je na základě výsledku experimentu, tedy získaných intenzit genové exprese, zařadit zkoumaný vzorek do některé z definovaných tříd. Metod používaných k hodnocení dat z DNA mikročipů je velmi mnoho a jsou neustále vylepšovány, ale stejně jako v jiných oblastech matematiky nelze říci, že by nějaká z těchto metod byla nejlepší a její výsledky ty nejsprávnější.



Obrázek 1.: Základní cíle v analýze DNA mikročipů a statistické metody k jejich dosažení.

Shlukovací metody

Shluková analýza je jednou z nejpoužívanějších vícerozměrných statistických metod, a to jak obecně, tak i v případě expresních dat. Jedná se o explorativní techniku, která se používá zejména v případech, kdy nemáme žádné *a priori* znalosti o struktuře uvnitř dat. Úkolem shlukovacích metod je tedy najít v datech skupiny prvků (shluky) tak, že prvky jednotlivých skupin budou v jistém smyslu více podobné než prvky z různých skupin, tzn. nalezené skupiny prvků budou co nejvíce homogenní. V případě DNA mikročipů jsou shlukovány zejména geny a experimentální vzorky (pacienti). Jednoduše řečeno, snažíme se nalézt mezi zkoumanými geny (resp. biologickými vzorky) skupinky genů (resp. biologických vzorků), které vykazují v průběhu experimentu, tedy za působení specifických podmínek, podobné chování.

Mírou podobnosti dvou prvků (ať již genů nebo pacientů) je v tomto případě matematická metrika neboli vzdálenost, k nejvhodnějším patří euklidovská vzdálenost, která odpovídá námi vnímané vzdálenosti ve trojrozměrném prostoru, a korelační vzdálenost, která odráží závislost expresních hodnot dvou prvků. Volba použité vzdálenosti vždy závisí na okolnostech daného experimentu. Při použití shlukovacích metod je však nutné mít na paměti, že tyto metody jsou určeny k hledání a nalezení shluků v datech, a to i v případě, že se v datech žádná vnitřní struktura nenachází. Z toho plyne, že výsledky shlukovací analýzy nelze brát jako dogma a je nutné je vždy konzultovat s expertem.

Některé metody (např. níže zmíněná metoda *k*-průměrů) požadují jako vstupní informaci počet shluků, které má daná procedura vytvořit, proto je nutné počet skupin v datech nějakým způsobem předem odhadnout. Tento odhad můžeme buď získat od klinického experta nebo použitím matematických algoritmů, jako např. procedury *silhouette width* [6].
Příkladem shlukovacích metod vhodných pro expresní data jsou následující procedury:

- **Metoda *k*-průměrů**. Podstatou této jednoduché, a proto oblíbené, procedury je rozdělení prvků do skupin tak, aby byla minimalizována vnitroshluková variabilita a zároveň maximalizována mezishluková variabilita .
- **Hierarchické metody**. Principem hierarchických metod je postupná tvorba shluků, kdy v každém kroku jsou spojeny dva nejpodobnější geny nebo skupiny genů. Tento heuristický přístup zajišťuje minimální vnitroshlukovou variabilitu.
- **Self-organizing maps**. Jedná se o jeden z mnoha algoritmů ze skupiny neuronových sítí. Shluky jsou reprezentovány jako uzly dvourozměrné mřížky, které se iterativně adaptují na data v prostoru. Výhodou tohoto algoritmu je, že sousední uzly v mřížce odpovídají podobným shlukům, výsledné uspořádání tudíž odráží nejen podobnost prvků ve shlucích, ale i podobnost mezi shluky.
- **Metody založené na matematickém modelu**. Tyto metody [7] předpokládají, že expresní hodnoty prvků každého shluku lze popsat matematickou funkcí, která se mezi shluky liší pouze v několika parametrech. Cílem je tedy odhad těchto parametrů pro jednotlivé shluky a následné přiřazení genu ke shluku, do nějž patří s největší pravděpodobností.

Metody pro identifikaci rozdílně exprimovaných genů
Častým a velmi důležitým cílem microarray experimentů je identifikace rozdílně exprimovaných genů, tedy genů, jejichž hladina exprese se liší mezi dvěma (nebo více) biologickými vzorky. Můžeme například zkoumat geny odpovědné za funkci různých druhů tkání, geny odpovědné za přeměnu zdravé tkáně v nádorovou nebo geny jednoznačně determinující jednotlivé nádorové tkáně. Znalost rozdílně exprimovaných genů je zajímavá nejen z hlediska přímé interpretace, ale také z hlediska použití této znalosti ke zpřesnění klasifikačních metod. Vzhledem k tomu, že při hodnocení genové exprese pracujeme někdy i s desítkami tisíc genů zároveň, setkáváme se zde s tzv. *problémem násobného testování hypotéz* [8]. Ten spočívá v tom, že s narůstajícím počtem hodnocených genů nám enormně roste také pravděpodobnost toho, že se při našem testování genů zmýlíme a označíme jako významný gen, který ve skutečnosti žádnou informaci o zkoumaných vzorcích neobsahuje. Každá metoda se s tímto problémem vypořádává po svém, ale v žádném případě nelze říci, že by tento problém byl uspokojivě vyřešen. Zde jsou některé příklady metod pro identifikaci významných genů:

- ***T*-test a jeho modifikace**. Principem této skupiny metod [9] je výpočet číselného skóre pro každý gen, které odráží rozdíl v genové expresi mezi sledovanými podmínkami, a následné určení prahové hodnoty, nad níž lze skóre považovat za statisticky významné. Jako skóre se většinou používá klasická *t*-statistika, někteří autoři však odvodili svá vlastní skóre, přičemž ve většině případů se jedná o modifikace *t*-statistiky.
- **Significance Analysis of Microarrays (SAM)**. Tato procedura [10] je také založena na skóre odvozeném od *t*-statistiky, ale vzhledem k její popularitě je vhodné ji uvést zvlášť. Její výhodou je, že pro určení prahové hodnoty oddělující významné a nevýznamné geny používá SAM permutace, které umožňují vyhnout se některým předpokladům méně propracovaných metod.
- **Optimal Discovery Procedure**. Storey a kol. [11] odvodili proceduru, která při hodnocení exprese jednotlivých genů bere do úvahy také hodnoty exprese ostatních genů, což zlepšuje její srovnávací schopnost v rámci celé skupiny zkoumaných genů.
- **Analýza rozptylu**. Někteří autoři považují normalizaci expresních hodnot, která předchází statistické analýze, za zbytečnou a používají pro identifikaci významných genů analýzu rozptylu [12], v níž pomocí parametrů modelují vliv různých složek experimentu na expresní hodnoty.

Klasifikační metody
Principem klasifikačních metod v analýze dat z DNA mikročipů je vytvoření rozhodovacího pravidla, které by na základě naměřených hodnot genové exprese umožňovalo přiřazení pacienta do jedné z předem definovaných tříd (například zdravý, nemocný). Z toho je zřejmé, že by se „dobré“ rozhodovací pravidlo založené na expresních datech mohlo zařadit po bok stávajících diagnostických metod a výrazně tak přispět ke zpřesnění diagnostiky závažných onemocnění.
Klasifikační pravidlo je vytvořeno v tzv. trénovací fázi, kdy daná metoda studuje dostupná data (tzv. trénovací data) a učí se přiřazovat pacienty do jednotlivých tříd dle jejich expresního profilu. Trénovací data musí být samozřejmě předem validovaná v tom smyslu, že u každého pacienta musí být jednoznačně určeno, do které třídy patří. Měřítkem kvality klasifikačních metod je tzv. *pravděpodobnost misklasifikace*, což je pravděpodobnost přiřazení pacienta do nesprávné kategorie, kterou lze získat použitím klasifikačního pravidla na tzv. validační data.
Pro klasifikaci se většinou používá pouze část z proměnných (genů) tvořících expresní profil pacienta. Je to dáno tím, že mnoho genů nemá vzhledem ke sledovaným třídám dostatečnou diskriminační schopnost, buď proto, že jsou neaktivní, nebo proto, že se u obou skupin pacientů chovají stejně. Vlastnímu použití klasifikační procedury by tedy měl předcházet výběr k tomuto účelu vhodných genů pomocí metod pro identifikaci rozdílně exprimovaných genů.
Z klasifikačních metod používaných pro hodnocení expresních dat lze vybrat následující příklady:

- **Lineární diskriminační analýza a její modifikace**. Diskriminační metody [13] předpokládají, že data v jednotlivých třídách jsou normálně rozdělená. Úkolem těchto metod je pak odhadnout z trénovacích dat parametry pro sledované skupiny, což následně umožňuje zařadit nového pacienta do třídy dle maximální pravděpodobnosti.
- **Metoda *k* nejbližších sousedů (*k*-NN)**. Princip této jednoduché metody je takový, že z trénovacích dat vybereme *k* pacientů, kteří jsou svými expresními profily nejbližší (ve smyslu matematické vzdálenosti) expresnímu profilu nového pacienta, a následně ho přiřadíme do té skupiny, která má mezi těmito *k* pacienty nejpočetnější zastoupení [14].
- **Support Vector Machines (SVM)**. Tato poměrně nová klasifikační metoda se snaží najít mezi uvažovanými třídami rovinu, která by je co nejlépe vzájemně oddělovala. SVM [15] jsou proto zejména vhodné pro klasifikaci, kdy uvažujeme pouze dvě rozdílné třídy.
- **Klasifikační stromy**. Jsou založeny na postupném štěpení datového souboru dle jednotlivých proměnných (genů) [16]. Proměnné, podle nichž se soubor štěpí, jsou vybrány tak, aby měly co nejlepší schopnost oddělit uvažované třídy.

Závěr
V tomto článku jsme se zaměřili na statistické metody a postupy použitelné pro hodnocení výsledků DNA microarray experimentů. Je patrné, že spektrum metod vhodných k dosažení odpovědí na otázky v oblasti výzkumu genové exprese je velmi široké. Otázky, a tedy i statistické metody, lze rozdělit do tří hlavních kategorií. Na první skupinu nejzákladnějších otázek týkajících se vnitřní struktury dat nám dávají odpověď shlukovací procedury, jejichž úkolem je nalezení co nejpodobnějších skupin prvků, ať už genů nebo pacientů. Druhou sadu otázek zabývajících se rozdílnou aktivitou genů ve sledovaných biologických vzorcích nám mohou pomoci rozluštit metody pro identifikaci rozdílně exprimovaných genů založe-

né na statistickém testování hypotéz, kdy pro každý gen testujeme rozdíl mezi dvěma (či více) skupinami naměřených expresních hodnot. Třetí skupina statistických procedur, klasifikační metody, řeší problém predikce spjatý s diagnostikou nových pacientů na základě jejich profilu genové exprese.

I když se na vývoji statistických metod pro analýzu dat z DNA mikročipů pracuje již přibližně 7 let, v žádném případě nelze říci, že by šlo o ukončený proces. Zejména obrovský počet zkoumaných genů a vůči němu poměrně malý počet dostupných pacientů či biologických vzorků představují největší překážky pro biostatistiky na celém světě. Nicméně, DNA mikročipy jsou natolik perspektivní technologií, že se vložené úsilí nepochybně v budoucnu odrazí jak ve vývoji kvalitních matematických metod analýzy, tak následně v prohloubení znalostí o fungování buněčných pochodů i ve zlepšení diagnostiky a léčby těch nejzávažnějších onemocnění.

Poděkování:
Tato práce byla podpořena grantem IGA MZ ČR NR/9076 - 4.

Literatura

1. Schena, M., Shalon, D., Davis, R.W. and Brown, P.O. (1995). Quantitative monitoring of gene expression patterns with complementary DNA microarray. Science, Vol. 270: 467-470.
2. Speed, T. (2003). Statistical analysis of gene expression microarray data. Chapman & Hall/CRC.
3. Troyanskaya, O., Cantor, M., Sherlock, G. et al. (2001). Missing value estimation methods for DNA microarrays. Bioinformatics Vol. 17 No. 6 pp. 520-525.
4. Eisen, M.B., Spellman, P.T, Brown, P.O. and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences USA, 95: 14863-14868.
5. Tamayo, P., Slonim, D., Mesirov, J. et al. (1999). Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. Proceedings of the National Academy of Sciences USA, 96: 2907-2912.
6. Fridlyand, Y.M. (2001). Resampling methods for variable selection and classification: applications to genomics. PhD thesis, University of California at Berkeley, Department of Statistics.
7. De Smet, F., Mathys, J., Marchal, K. et al. (2002). Adaptive quality-based clustering of gene expression profiles. Bioinformatics Vol.18 No. 5 pp.735-746.
8. Ge, Y., Dudoit, S. and Speed, T.P. (2003). Resampling-based multiple testing for microarray data analysis. Technical report, University of California at Berkeley, Department of Statistics.
9. Reiner, A., Yekutieli, D. and Benjamini, Y. (2003). Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinformatics Vol. 19 No. 3: 368-375.
10. Tusher, V.G., Tibshirani, R. and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. Proceedings of the National Academy of Sciences USA, 98: 5116-5121.
11. Storey, J.D., Dai, J.Y. and Leek, J.T. (2005). The Optimal Discovery Procedure for large-scale significance testing, with applications to comparative microarray experiments. UW Biostatistics Working Paper Series, University of Washington.
12. Kerr, M.K. and Churchill, G.A. (2000). Statistical design and the analysis of gene expression microarray data. Technical report, The Jackson Laboratory.
13. Dudoit, S., Fridlyand, J. and Speed, T.P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. Journal of the American Statistical Association Vol.97 No 457: 77-87.
14. Wit, E. and McClure, J. (2004). Statistics for microarrays. Wiley, Chichester, GB.
15. Vapnik, V. (1998). Statistical Learning Theory. Wiley, Chichester, GB.
16. Lee, J.W., Lee, J.B., Park, M. and Song, S.H. (2005). An extensive comparison of recent classification tools applied to microarray data. Computational Statistics & Data Analysis, 48: 869 – 885.