
METODY DETEKCE ALTEROVANÝCH OBLASTÍ DNA Z DAT CGH ARRAYS

STATISTICAL APPROACHES FOR IDENTIFICATION OF AREAS OF DNA ALTERATION ANALYZED BY CGH ARRAYS

BUDINSKÁ E., JARKOVSKÝ J.

INSTITUT BIOSTATISTIKY A ANALÝZ, PŘF A LF, MU BRNO.

Souhrn

CGH arrays se v dnešní době staly významnou technikou v analýze genomu. Využívají se pro detekci změn v počtu kopií genů nebo chromozomů, popřípadě jiných chromosomových přestaveb. Základní metodou je porovnávání DNA vzorku a zdravé kontroly pomocí komparativní genomové hybridizace (CGH). Analýza genetické nestability u nádoru a s ní spojené hledání nádorových markerů se dá využít jak v diagnostice nádorů (klasifikace nádorů do již existujících skupin, hledání nových podskupin), tak i v predikci jejich odpovědi na léčbu. Výhodou technologie CGH arrays je možnost analýzy celého genomu v jediném experimentu. To ovšem znamená, že metoda produkuje velké množství dat, které musí být správně matematicky vyhodnoceny a základním cílem článku je tak představení těchto matematických metod a jejich softwarové implementace.

Klíčová slova: CGH arrays, statistická analýza, body zlomu, analýza segmentů

Summary

Nowadays, array CGH experiments became a powerful technique for analysing changes in DNA by comparing control DNA and DNA of interest. This is widely used for example in genome cancer studies. Analysis of the tumour genome instabilities and the search for tumour markers can be used for the tumour diagnostics (tumour classification, detection of the new clinical groups of tumours) or for the prediction of the response to therapy. The method produce huge amount of data and special statistic techniques for detecting of genomic changes are necessary. The purpose of this paper is to provide brief summary of existing statistical methods used in CGH array analysis and their software implementation.

Keywords: array CGH, statistical analysis, breakpoints, segment analysis

I. Úvod

Analýza microarrays je relativně nová a v dnešní době populární technologie analýzy genomu a jeho exprese. Jde o metodu schopnou analyzovat tisíce genů v jediném experimentu s širokým uplatněním v onkologii. Využívá se jí pro porovnání genomu nádorových buněk s genomem buněk zdravých téhož typu tkáně, což umožňuje vyhledávat nádorové markery a ty dále využít například pro klasifikaci nádorů do již existujících skupin, identifikovat nové podskupiny, určit odpověď nemocného na terapii [1], nebo rozlišit primární nádory od metastáz [2]. Microarray technologie může být využita ve všech třech typech analýzy genomu. Tzv. **CGH arrays** se využívají pro detekci alterovaných úseků v DNA, **expresní microarrays** se zabývají analýzou míry vyjádření (exprese) genu (ta je ekvivalentní množství RNA), a nakonec **proteinové microarrays** studují strukturu již vytvořených proteinů. CGH arrays jsou metoda, kterou lze detekovat alterované úseky v DNA, tedy delece nebo amplifikace. Protože některé CGH čipy pokrývají celý genom, jsou tyto microarrays schopny podat relativně detailní obraz o změnách podél všech chromozomů.

Základním principem CGH microarray technologie je hybridizace nukleových kyselin se sondami umístěnými na podložním skle (tzv. spoty). Zjednodušeně se na každém spotu nachází jeden klon. Zdrojem informace je DNA z jedné nebo dvou tkání (případně buněčných linií). Hybridizované DNA jsou přitom obarveny různými fluorescenčními barvami (nejčastěji červená – tkáň, kterou zkoumáme, zelená – kontrola), po jejichž excitaci laserem o příslušné vlnové délce je získána intenzita světla. Tato intenzita je úměrná množství DNA přichycené k sondě a tedy přenesené i počtu kopií genu ve zdrojové tkáni. Takto získaný obraz čipu je dále analyzován pomocí programů pro analýzu obrazu a jednotlivé intenzity fluorescence (ekvivalentní množství DNA přichycené na odpovídajících sondách) jsou převedeny do číselné podoby, která podstupuje další matematické a statistické zpracování vedoucí k závěrům a interpretacím.

Podobně jako u expresních microarrays se i u CGH arrays využívá nejčastěji takzvaný „referenční“ design, který je založen na porovnávání fluorescenčních intenzit testované (nádorové) DNA a referenční (kontrolní, zdravé) DNA. Jedná se vlastně o komparativní geonomovou hybridizaci (CGH) na microarrays čipech.

Pro následnou statistickou analýzu jsou v případě tohoto designu zajímavé poměry fluorescenční intenzity nádorové a referenční DNA pro každý klon na microarray sklíčku. Tyto poměry se nazývají „CGH ratio“, a obvykle se logaritmuji logaritmem o bázi dva ($\log_2$ ratio):

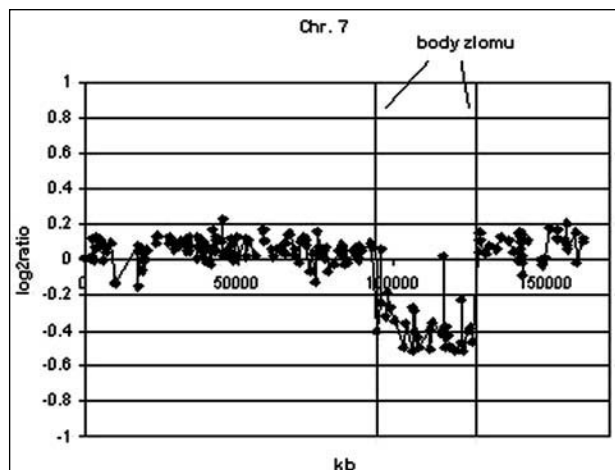
$$CGH\ ratio = \log_2 \frac{\text{intenzita nádorové DNA}}{\text{intenzita kontrolní DNA}}$$

Takhle upravený podíl se spíše řídí přibližně normálním rozložením (vzhledem k schopnosti logaritmické transformace měnit asymetrické rozložení na více či méně symetrické) [3], což je výhodné pro řadu statistických technik. Tyto podíly jsou obvykle vyjadřovány až po úpravě dat (která zahrnuje například odstranění nežádoucích spotů z analýzy) nebo případně normalizaci (pomáhá odstraňovat případné systematické i náhodné chyby), což je fáze společná jak pro analýzu expresních tak pro analýzu CGH arrays. Následujícím krokem v analýze dat CGH arrays je potom průměrování hodnot CGH ratios všech klonů oproti jejich pozici na genomu (Obr. 1).

Tenhle krok poskytuje prvotní informaci o potenciálních změnách a uplatňuje se zde další výhoda logaritmování podílů. Jestliže uvažujeme ve zdravé buňce dvě kopie každého genu

(od otce a od matky), potom v případě že v nádorové buňce nedošlo u daného genu k žádné změně, dostáváme poměr 2/2, což znamená že CGH ratio se rovná nule ($\log_2 2/2 = 0$). CGH ratio menší než -1 indikuje deleci nejméně jedné kopie genu ($\log_2 1/2 = -1$) a větší než 0.58 ($\log_2 3/2 = 0.58$) naopak amplifikaci klonu v testovacím vzorku. Zjednodušeně řečeno, všechny záporné hodnoty znamenají potenciální delece a všechny kladné hodnoty potenciální amplifikace.

Na základě výše uvedených hodnot by bylo možno stanovit dvě dělicí hranice a ty využít pro detekci změn; co se týče delece, označili by jsme tak všechny klony s CGH ratio menší než -1 a za amplifikace by jsme pak považovali všechny klony s CGH ratio větší než 0.58. Z praktického hlediska bohužel není možné tento způsob analýzy použít a to z důvodu existence technické i biologické variability. Vzorek nádoru použitý pro analýzu obvykle nikdy neobsahuje sto procent nádorových buněk a nacházejí se v něm i buňky zdravé, čímž se jednotlivé hodnoty CGH ratio pro delece i amplifikace posouvají směrem k nule. Často je těžké pouhým okem rozlišit tento „šum“, který představuje falešnou odchylku od skutečných změn. Pro řešení tohoto problému byly vyvinuty speciální statistické metody, které jsou schopny s jistotou pravděpodobností tyto dva stavy odlišit. Zdrojem „šumu“ (což je směs náhodných vlivů, které ovlivňují skutečné hodnoty zkoumané proměnné a nelze je od ní uspokojivě oddělit) se kromě přítomnosti zdravých buněk ve vyšetřovaném vzorku mohou potenciálně stát také všechny části zpracovávání microarray sklíčka. Naměřené hodnoty tak mohou být ovlivněny například různými inkorporačními a excitačními vlastnostmi jednotlivých fluorescenčních barviv nebo přirozená fluorescence pozadí apod.



Obrázek 1: Graf reprezentující CGH ratio klonů proti jejich pozici na genomu v kb. Vertikální přímky označují body zlomu ohraničující detekovaný region lidského chromozomu 7.

Je známo, že u nádorů jsou obvykle změnami zasaženy větší regiony DNA [4.], tedy lze očekávat, že tyto změny jsou na grafu pozorovatelné v podobě segmentů amplifikací/delecí. Oblasti, které tyto segmenty oddělují se nazývají „body zlomu“, protože se v nich graf „láme“ buď do záporných nebo do kladných hodnot (Obr. 1). Většina statistických metod používaných pro analýzu CGH arrays se snaží detekovat tyto body zlomu a vymezit tím zmíněné úseky co nejpřesněji. Při hledání segmentů se vychází z předpokladu, že geny jsou na genomu uspořádány v určitém pevném pořadí a v případě, že jsou deletovány například geny s pořadím 1, 2 a 4, je s největší pravděpodobností deletován i gen mezi nimi s pořadovým číslem 3. Tyto předpoklady pomáhají redukovat šum a zvýšit statistickou průkaznost detekovaných změn. Statistické metody

pro hledání alterovaných úseků DNA u CGH arrays se liší od metod pro hledání odlišně exprimovaných genů u expresních microarrays právě těmito předpoklady.

Hledání alterovaných větších úseků je pouze jedním z úkolů analýzy CGH arrays. Dalšími je například detekce bodových mutací, změn o velikosti jednoho nebo několika málo genů a určení počtu amplifikovaných nebo deletovaných kopií genů, neboli míry změny.

II. Metody pro statistickou analýzu CGH arrays

II.1. Metodika

V současnosti existuje řada statistických metod vyvinutých za účelem hledání alterovaných úseků DNA. Jak již bylo zmíněno, většina z nich využívá informaci o pořadí genů na genomu a faktu, že sousední geny mají většinou obdobné CGH ratio (tedy patří do stejné alterované nebo nezměněné oblasti).

Kromě nich ovšem existují i metody, které tyto předpoklady nevyužívají, jako jsou například metody používající dvě pevné dělicí hranice, jednu pro deletované a druhou pro amplifikované regiony, lišící se ve způsobu stanovení těchto hranic. Z nich lze zmínit metodu [5], která navrhuje stanovení hranic na základě rozložení CGH ratio z CGH array experimentů kontrola vůči kontrole, nebo metodu [6] využívající regresní analýzu pořadí CGH ratios. Tyto metody mají obecně nevýhodu, že neberou do úvahy výše uvedené předpoklady a jsou tak mnohem náchylnější k falešně pozitivním nebo falešně negativním výsledkům. Může se například stát, že gen s pořadím číslo 3 se nachází ve velkém alterovaném regionu DNA (dejme tomu o 20 genech), ale díky šumu jeho CGH ratio nemá dostatečně vysokou hodnotu pro přesáhnutí dělicí hranice a je tak označen jako nealterovaný, i když ve skutečnosti alterovaný je. Kromě toho použití kontrolních CGH arrays experimentů pro stanovení dělicí hranice může být zavádějící, protože takovéto hranice neberou do úvahy případné odchylky u reálných experimentů a výsledky tak nemusí úplně odpovídat realitě.

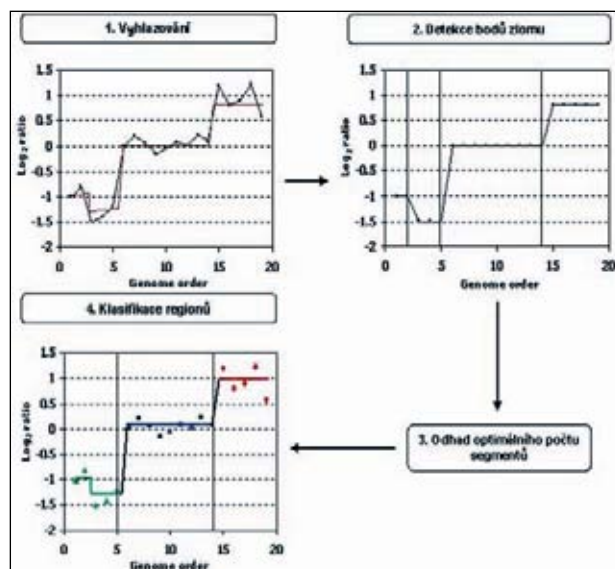
Ostatní metody se snaží CGH data matematicky modelovat a na základě optimalizace různých kritérií pomocí nejrůznějších algoritmů najít nejlepší rozdělení dat do segmentů. Obvykle využívají jeden nebo více z následujících kroků (Obr. 2):

1. vyhlazení dat (používáno pro odstranění šumu)
 2. detekce bodů zlomu (detekce změněných úseků)
 3. odhad optimálního počtu změněných úseků
 4. klasifikace regionů (zařazení detekovaných regionů do tří skupin – deletované, nezměněné, amplifikované)
- a liší se pouze statistickými metodami, které pro tyto kroky používají.

První krok, takzvané **vyhlazování dat**, má za účel redukovat v datech šum způsobený technickou nebo nežádoucí biologickou variabilitou (Obr. 2. - 2). Pro vyhlazování se používají různé metody založené nejčastěji na virtuálním pohyblivém okně o určité velikosti, které se posouvá po hodnotách zleva doprava. V tomto okně, například o velikosti 5 genů, se spočítá jeho statistická sumarizace, například medián (mediánové vyhlazování), kterou se pak nahradí například prostřední hodnota v okně. Jednotlivé metody se liší velikostí okna, hodnotou, která je nahrazována a druhem počítané statistické sumarizace.

Ne všechny vyhlazovací metody jsou založeny na pohyblivém okně. Za všechny zmíníme metodu tzv. „kvantilové regrese“ [7], která byla speciálně navržena pro analýzu CGH arrays. Její výhodou je, že na rozdíl od ostatních technik netvoří při vyhlazování dat obloukovité a málo výrazné přechody mezi odlišnými segmenty. Problémem vyhlazování je, že kromě odstranění šumu dochází někdy i ke ztrátě skutečné informace. V případě dat z CGH arrays tak můžeme „vyhladit“ například bodové změny, nebo změny o několika málo genech a proto ne všechny metody tento krok využívají.

Dalším krokem analýzy je **detekce bodů zlomu**, která je vlastně ekvivalentní úloze najít změněné úseky (segmenty), proto se tato část analýzy nazývá také segmentací. Pro řešení úlohy je využívána řada odlišných metod. Jednu podskupinu tvoří segmentační metody, které předpokládají, že data pocházejí z rozložení, jež je směsí tří normálních rozložení (pro delece, pro geny beze změny, pro amplifikace), které se liší pouze střední hodnotou (např. -1, 0, 0.58).



Obrázek 2.: Schéma metodiky vyhledávání alterovaných segmentů.

Tyto metody se pak snaží správně odhadnout právě parametry tohoto rozložení (průměr, směrodatná odchylka) a tím určit správnou segmentaci. Metody se liší v optimalizačním kritériu (míra kontroly správnosti modelu) a algoritmem, kterým tuto optimalizaci provádějí. Z optimalizačních kritérií zmíníme věrohodnostní (likelihood) kritérium [8, 9, 10, 16], kritérium nejmenších čtverců [11], parciální sumy [12], edge detection filter [13]. Jako optimalizační algoritmy používají genetické algoritmy [8], dynamické programování [10, 11], cirkulární binární segmentaci [12], adaptive weights smoothing [9] nebo EM algoritmus [13, 16]. Další metody využívají skryté Markovovy řetězce [14] nebo modifikovanou metodu hierarchického shlukování - CLAC [15].

Kroky tři a čtyři, odhad optimálního počtu segmentů a klasifikace regionů, představují nezbytné zakončení analýzy, protože v kroku druhém nacházíme pouze nejlepší rozdělení dat do určitého počtu segmentů. V těchto segmentech však mohou být buď některé zbytečně navíc (falešně pozitivní), a je třeba je sjednotit, nebo, jako v případě metody CLAC, musíme zpětně určit počet segmentů, které jsou ze všech možností vybrány (pokud to není již součástí kroku 2). Pro nalezení optimálního počtu segmentů je používáno shlukování, jako například metoda k průměrů [11], nebo nejčastěji různá penalizační kritéria [10, 13, 15]. Je třeba zdůraznit, že některé metody tyto kroky nepotřebují, a kroky 2-4 se tak mohou prolínat.

Vysvětlování principu jednotlivých metod by přesáhlo rozsah tohoto článku, proto zájemce o podrobnější informace odkážeme na literaturu.

II.2. Existující software pro analýzu dat CGH arrays

Softwarové implementace metod pro analýzu dat CGH arrays jsou obvykle dobře dostupné a volně stažitelné, pro některé metody existuje i více implementací. Kromě samostatných programů nebo webových aplikací se metody ve většině případů vyskytují v podobě funkcí známých matematických programů, jako jsou R [16] nebo Matlab [17] a pro analýzu je tak nut-

ná znalost práce v těchto programech. Níže jsou shrnuty příklady implementací:

- **R - balíky:**
 - GLAD [9]
 - cgh_smooth_nakao [7]
 - CLAC [14]
 - snapCGH [23]
- **Matlab:**
 - GLAD [9]
 - M-CGH [18]
 - CGH plotter [11]
 - CGH_segmentation [10]
 - ChARM [13]
- **Excel:**
 - CGH miner for Excel [14]
- **Samostatné programy nebo webové aplikace**
 - ArrayCyGht – webová aplikace pro analýzu a vizualizaci dat CGH arrays [19]
 - CGH Explorer – komplexní java nástroj [20]
 - VAMP – komplexní webová aplikace zahrnující vizualizaci, analýzu i porovnávání a analýzu profilů více pacientů najednou [21]
 - aCGHSmooth [22]

II. 3. Výhody a nevýhody jednotlivých metod

Každá z výše uvedených metod má v porovnání s ostatními své výhody a nevýhody. Pro některé je nezbytné znát a mít k dispozici speciální matematický software (R, Matlab), některé implementace zase vyžadují specifický typ formátu dat. To může být problémem, pokud máme data z jiných zdrojů a nejsme schopni je do daného programu načíst (např. M-CGH vyžaduje Gene-pix nebo Quant array formát). Některé z metod nejsou „komplexní“, a to ve smyslu chybějící analýzy segmentů (po vyhlazování využívají pouze dělicí hranici, kterou určuje sám uživatel – např. [7]). Na druhé straně, komplexní metody mohou vyžadovat nastavení množství parametrů, což vyžaduje dobrou znalost problematiky a metody samotné (např. GLAD, MCGH). Metoda CLAC je komplexní metoda, která se relativně jednoduše používá, avšak vyžaduje několik normal-normal arrays, což u některých experimentů nemusí být z finančních důvodů možné. Některé implementace poskytují předzpracování dat zahrnující kontrolu kvality a normalizaci, spolu s grafickými vizualizacemi a případnými dalšími analýzami. V současnosti je z tohoto pohledu nejkomplexnějším webovým nástrojem nástroj VAMP [21], využívající metodu GLAD, který poskytuje kromě komplexní metody analýzy změněných regionů i všechny ostatní stupně analýzy dat, včetně normalizace a následné analýzy dat více pacientů. Snad zatím jediným nástrojem, který spojuje více různých metod pro analýzu a umožňuje srovnání jejich výsledků je balík snapCGH [23], implementovaný v programu R. Tabulka č. 1 sumarizuje výhody a nevýhody vybraných metod a jejich implementací tak jak je vidí autoři. Matematické porovnání většiny metod najdeme také v referenci č. 24.

III. Závěr

CGH arrays se v dnešní době staly významnou technikou pro analýzu genomu a v jednom experimentu umožňují analýzu tisíců genů. Protože poruchy na úrovni genomu obvykle zasahují širší chromozomální regiony (což je informace kterou u expresních microarrays tolik nevyužijeme), je užitečné při jejich hledání tuto informaci použít a význam-

ně tak redukovat šum v datech. V článku jsou představeny pouze některé, z našeho pohledu nejlepší dostupné metody analýzy CGH arrays, a to prostřednictvím svých základních principů.

Tabulka 1: Porovnání implementací vybraných metod pro analýzu dat CGH arrays. „Komplexnost algoritmu“ znamená, že v programu je použita matematická metoda pro vyhledávání změněných segmentů (tedy ne pouze stanovení dělicí hranice).

Program	Implementace	Předzpracování dat	Grafy	Je algoritmus komplexní	Hlavní nevýhody	Hlavní výhody
MCGH	Matlab	ANO	ANO	ANO	Speciální vstupní formát dat (GenePix and QuantArray)	Komplexnost
CGH plotter	Matlab	NE	ANO	ANO	-	Lze použít i pro expresní microarrays
arrayCyGht	web	ANO	ANO	NE	Nekomplexní	Jednoduchý, propojení s databázemi
GLAD	R, Matlab	NE	ANO	ANO	Není „user friendly“	Komplexnost
VAMP	web	ANO	ANO	ANO	-	Komplexní zpracování CGH arrays dat, analýza více pacientů
cgh_smooth_nakao	R	NE	ANO	NE	nekomplexní, není „user friendly“	-
CGH-miner	R, Excel	NE	ANO	ANO	Vyžaduje normal – normal arrays	Implementace v Excelu

Tyto metody, na rozdíl od metod pro expresní arrays, nejsou doposud natolik unifikovány. Každá z nich má své výhody i nevýhody, mnohdy se musí uživatel rozhodovat mezi kvalitou výsledku a jednoduchostí používání. Metody se obvykle volí podle preferencí uživatele, v praxi se však doporučuje zanalyzovat data vždy více způsoby a výsledky pak porovnat. V každém případě je nutná alespoň základní znalost matematického pozadí metody využitě pro analýzu. Vzhledem k faktu, že mnohé metody pro analýzu používají pouze informaci o pořadí jednotlivých klonů na genomu a ne jejich skutečnou polohu, je důležité nalezené regiony podrobit následné kontrolní analýze uživatelem s ohledem na skutečnou pozici. Zjištěné aberace je dále vhodné ověřit pomocí jiných přesnějších molekulárně biologických technik (např. PCR). Jako zatím nejkomplexnější nástroj pro analýzu CGH arrays se jeví být webová aplikace VAMP, která ale bohužel implementuje pouze jednu z metod a neumožňuje tak porovnání. Z tohoto hlediska jsme zatím zjistili pouze jedinou aplikaci, která poskytuje porovnání tří metod najednou a to balík snapCGH, implementovaný v programu R. V této oblasti doposud chybí nástroj, který by propojoval metody všech částí analýzy CGH arrays a poskytoval širší porovnání výsledků. Z tohoto důvodu zůstáváme při jejich používání doposud odkázáni na znalost práce s různými software a detailní znalost matematického pozadí využívaných metod. Zároveň jde o oblast s dosud neukončeným matematickým vývojem, kde je možno očekávat rozvoj jak dosavadních, tak vývoj zcela nových metod analýzy dat.

Poděkování

Tato práce byla podpořena grantem IGA MZ ČR NR/9076 – 4.

Literatura:

- Natrajan R, Williams RD, Hing SN et al. (2006). Array CGH profiling of favourable histology Wilms tumours reveals novel gains and losses associated with relapse. *J Pathol*, DOI: 10.1002/path.2021.
- Wa CV, DeVries S, Chen YY et al. (2005). Clinical application of array-based comparative genomic hybridization to define the relationship between multiple synchronous tumors. *Modern Pathology* 18:591-597.
- Zar H. Jerrold (1999). *Biostatistical Analysis*. Prentice Hall, Upper Saddle River NJ, USA, 4. ed.: 275.
- Loeb LA, Christians FC (1996). Multiple mutations in human cancers. *Mutat Res*;350:279-86.
- Pollack, J., Sorlie, T., Perou, C. et al. (2002). Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumours. *Proceedings of the National Academy of Sciences, USA* 99, 12963–12968.
- Cheng, C., Kimmel, R., Neiman, P. and Zhao, L. P. (2003). Array rank order detection of gene copy-number changes in human cancer. *Genomics* 82, 122–129.
- Eilers, P. H. C. and de Menezes, R. X. (2005): Quantile smoothing of array CGH data. *Bioinformatics*, 21: 1146–1153.
- Jong, K., Marchiori, E., van der Vaart, A., Ylstra, B., Meijer, G. And Weiss, M. (2003) Chromosomal breakpoint detection in human cancer. In *LNCS*, vol. 2611, Springer.
- Hupé, P., Stransky, N., Thierry, J.-P. et al. (2004). Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics* 20, 3413–3422.
- Picard, F., Robin, S., Lavielle, M. et al. (2005). A statistical approach for CGH microarray data analysis. *BMC Bioinformatics*; 6:27.
- Autio, R., Hautaniemi, S., Kauraniemi, P. et al. (2003). Cgh-plotter, Matlab toolbox for CGH data analysis. *Bioinformatics* 19, 1714–1715.
- Olshen, A. B., Venkatraman, E. S., Lucito, R. and Wigler, M. (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 5, 557–572.
- Myers, C.L., Dunham, M.J., Kung, S.Y., Troyanskaya, O.G. (2004) Accurate detection of aneuploidies in array CGH and gene expression microarray data. *Bioinformatics* 20(18):3533-3543.
- Fridlyand, J., Snijders, A.M., Pinkel, D., Albertson et al. (2004). Hidden Markov models approach to the analysis of array CGH data. *J. Multivariate Anal.* 90, 132–153.
- Wang P, Kim Y, Pollack J et al. (2005). A method for calling gains and losses in array CGH data. *Biostatistics* 6(1):45-58.
- R Development Core Team (2006). *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>.
- Mathworks (2006). *MATLAB: The Language of Technical Computing*, Massachusetts: The MathWorks, Inc.
- Wang, J., Meza-Zepeda, L.A., Kresse, S.H., Myklebost, O. (2004). MCGH: analysing microarray based CGH experiments. *BMC Bioinformatics* 5, 74.
- Kim, S.Y., Nam, S.W., Lee, S.H. et al. (2005). ArrayCyGHt: a web application for analysis and visualization of array-CGH data. *Bioinformatics* 21(10): 2554-5.
- Lingjarde, O.Ch., Baumbusch, L.O., Liestol, K. et al. (2005). CGH-Explorer: a program for analysis of array-CGH data. *Bioinformatics* 21(6):821-822.
- La Rosa P, Viara E, Hupe P, Pierron G (2006). VAMP: visualization and analysis of array-CGH, transcriptome and other molecular profiles. *Bioinformatics*. July 4.
- Jong, K., Marchiori, E., Meijer, G. et al. (2004). Breakpoint identification and smoothing of array comparative genomic hybridization data. *Bioinformatics* 1, 1–2.
- Smith, ML. (2006). Sample of aCGH analysis. www.bioconductor.org/packages/1.8/bioc/vignettes/snapCGH/inst/doc/snapCGHguide.pdf.
- W.R. Lai, M.D. Johnson, R. Kucherlapati, and Peter J. Park (2005). Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data, *Bioinformatics* 21(19):3763-3770.