Pacific Symposium on Biocomputing 13:315-326(2008)

PREDICTING DNA METHYLATION SUSCEPTIBILITY USING CpG FLANKING SEQUENCES

S. KIM[1,2], M. LI[3], H. PAIK[3], AND K. NEPHEW[3]
[1]*School of Informatics,* [2]*Center for Genomics and Bioinformatics,* [3]*Medical Sciences, Indiana University, Bloomington, IN 47408, USA*
E-mail: {sunkim2,menli,hyupaikknephew}@indiana.edu

H. SHI[4], R. KRAMER[5] AND D. XU[5,6]
[4]*Department of Pathology and Anatomical Sciences, Ellis Fischel Cancer Center,* [5]*Department of Computer Sciences and* [6]*Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO 65212, USA*
E-mail: {ShiHu@health.,kramer@,xudong@}missouri.edu

T-H. HUANG[7]
[7]*Human Cancer Genetics, The Ohio State University, Columbus, OH 43210, USA; E-mail: Tim.Huang@osumc.edu*

DNA methylation is a type of chemical modification of DNA that adds a methyl group to DNA at the fifth carbon of the cytosine pyrimidine ring. In normal cells, methylation of CpG dinucleotides is extensively found across the genome. However, specific DNA regions known as the CpG islands, short CpG dinucleotide-rich stretches (500bp - 2000bp), are commonly unmethylated. During tumorigenesis, on the other hand, global de-methylation and CpG island hypermethylation are widely observed. *De novo* hypermethylation at CpG dinucleotides is typically associated with loss of expression of flanking genes, thus it is believed to be an alternative to mutation and deletion in the inactivation of tumor suppressor genes. In this paper, we report that sequences flanking CpG sites can be used for predicting DNA methylation levels. DNA methylation levels were measured by utilizing a new high throughput sequencing technology (454) to sequence bisulfite treated DNA from four types of primary leukemia and lymphoma cells and normal peripheral blood lymphocytes. After measuring methylation levels at each CpG site, we used 30 bp flanking sequences to characterize methylation susceptibility in terms of character compositions and built predictive models for DNA methylation susceptibility, achieving up to 75% prediction accuracy in 10-fold cross validation tests. Our study is first of its kind to build predictive models for methylation susceptibility by utilizing CpG site specific methylation levels.

Pacific Symposium on Biocomputing 13:315-326(2008)

1. Introduction

DNA methylation, the addition of a methyl group to the fifth carbon of a cytosine residue within the context of a CpG dinucleotide, is the only known epigenetic modification of DNA that can be inherited without changing the DNA sequence [1]. In normal cells, CpG methylation is extensively found across the genome and widely believed to act to silence gene expression and/or retrotransposition of parasitic repeat sequences [2]. However, significantly less CpG methylation is observed within specific regions known as the CpG islands, short CpG dinucleotide-rich stretches (500bp - 2000bp), commonly found within the promoter and first exon of active genes [2].

Patterns of epigenetic modifications that arise during tumorigenesis are quite different from normal cells [3]. These alterations include global hypomethylation of CpG dinucleotides as well as localized hypermethylation at CpG islands [4]. It is now firmly established that CpG island hypermethylation is a powerful mechanism of transcriptional repression in cancer genomes, including silencing of tumor suppressor genes [5]. However, while DNA hypermethylation of several promoter CpG islands is frequently observed in cancers, other CpG island-containing genes remain unaffected by this epigenetic modification [6]. This observation indicates that some CpG island sequences are more susceptible to aberrant methylation, while others remain resistant to alteration by DNA methylation. While the reason for this differential susceptibility to DNA methylation is unknown, recent reports suggest that DNA pattern information may play a key role in distinguishing between methylation-sensitive and -resistant CpG islands. [7,8,9,10,27,26]

A widely used experimental method to measure DNA methylation is Methylation Specific PCR (MSP), a bisulfite conversion based PCR technique [11]. The target DNA is first modified with sodium bisulfite which converts un-methylated cytosine to uracil while methylated cytosine remains 5-methyl cytosine. The technique is accurate, however limited to detecting only highly specific regions of individual genes. Several genome-wide methylation detection methods have been developed in the past few years, such as differential methylation hybridization (DMH) [12] and methylation DNA immunoprecipitation (MeDIP) [13]. Both methods are microarray based experiments. DMH uses methylation specific restriction enzymes and MeDIP uses 5-methyl cytosine antibody to distinguish methylated and unmethylated DNA. These methods can be used for genome-wide methylation detection, allowing researchers to quickly profile methylation pattern

Pacific Symposium on Biocomputing 13:315-326(2008)

alteration. However, technical issues, such as hybridization efficiency for microarrays and antibody efficiency, affect their accuracy. Usually, MSP or bisulfite sequencing is performed to further validate the methylation levels of the genes of interest.

In the past, due to the limitation of sequencing technology, bisulfite sequencing has been performed only for determining specific regions of individual genes. With the development of high throughput sequencing technology, we are now able to perform methylation profiling on genome-wide scale. In particular, the 454 sequencing technology (`454.com`) combines emulsion PCR and pyrosequencing technique and it can determine up to 100Mbp in a single biological experiment with an average read length of 250 bp. The approach produces sequences of a very high quality with an accuracy over 99% [30] and resolution of single 5-methylcytosine, thus highly reliable for profiling methylation patterns. In this paper, we utilized the methylation data measured using the new high throughput sequencing technique to build predictive models for methylation susceptibility.

2. Related work and Motivation

There has recently been a significant research development in predicting DNA methylation susceptibility based on DNA patterns. Feltus et al [24] showed that a classification function based on the frequency of seven sequence patterns was able to discriminate methylation-prone from methylation-resistant CpG island sequences with over 80% accuracy in an experiment designed using over-expressed DNMT1. Feng et al [27] developed a support vector machine classifier for predicting methylation status of CpG islands and showed relationship between nucleotide sequence contents and transcription factor binding sites. Bock et al [26] showed that CpG island methylation in human lymphocytes was highly correlated with DNA sequence, repeats, and predicted DNA structure.

All previous studies used rather coarse-grained methylation information in long DNA regions, rather than CpG site specific information. With such coarse grained information, patterns as short as 3bp were used to build predictive models. Use of such short patterns utilizes frequencies of patterns, not specific patterns, for the predictive models. For example, a pattern of length 4bp, CCGC, is over-represented in unmethylated CpG Islands with a p-value of 5.18×10^{10} in Bock et al [26]. This means that CCGC occurs both in methylation susceptible and resistant sequences, but their occurrence frequencies in susceptible and resistant sequences are significantly

Pacific Symposium on Biocomputing 13:315-326(2008)

different. On the other hand, Handa and Jeltsch's analysis [20] reported that flanking sequences of up to +/-four base-pairs surrounding the central CG site that are characteristic of high (5'-CTTGCGCAAG-3') and low (5'-TGTTCGGTGG-3') levels of methylation in human genomic DNA.

In this paper, we investigated on whether specific DNA sequences, not just their frequencies, can be used for predicting methylation susceptibility. We used CpG flanking sequences with CpG site specific methylation information measured by sequencing bisulfite treated DNA from four types of primary leukemia and lymphoma cells and normal peripheral blood lymphocytes with a new high-throughput sequencing technology (454). This study is first of its kind that uses CpG site specific methylation information to build predictive models.

3. Research Goal

Once we measured methylation level at each CpG site (explained in "Method" Section), we investigated two research questions:

(1) Is there any significant difference in DNA character composition of CpG flanking sequences of methylation susceptible sites and methylation resistant sites?
(2) Is it possible to use the CpG flanking sequence composition to build predictive models for methylation susceptibility?

The first research question is directly motivated by using the high throughput sequencing technique. Indeed, the answer to the first question is positive as shown in Section 6.2. In addition to the significant difference in DNA character composition of CpG flanking sequences, we also observe that CpG sites in the same region of a gene are quite different in terms of methylation level as shown in Figure 5. If methylation levels of multiple CpG sites in the same genomic region are different, we may be able to distinguish methylation susceptible CpG sites from resistant sites using sequence-specific features, especially CpG flanking sequences. Thus the second question of modeling methylation susceptibility using machine learning techniques was investigated in this paper.

4. Data

A massively parallel sequencing (454-sequencing) experiment was designed on 25 gene-related CpG islands in four different tumor types, such as acute

Pacific Symposium on Biocomputing 13:315-326(2008)

lymphoblastic leukemia (ALL), chronic lymphocytic leukemia (CLL), follicular lymphoma (FL), and mantle cell lymphoma (MCL), and normal peripheral blood lymphocytes (PBL) [31]. These 25 genes, listed below, were previously reported to be highly methylated in leukemia and lymphoma.

CYP27B1 (Chr12: 56446851-56447152), *PON3* (Chr7: 94863816-4864092), *KCNK2* (Chr1: 213322021-213322250), *PCDHGA12* (Chr5: 140790654-140790834), *DDX51* (Chr12: 131193754-131194006), *PTPN6* (Chr12: 6930334-6930791), *DAPK* (Chr9: 89302286-89302749), *CDKN2B* (Chr9: 21998564-21998941), *TP53* (Chr17: 7531401-7531767), *CDKN1C* (Chr11: 2863438-2863596), *TRIM36* (Chr5: 114543737-114544155), *ZNF677* (Chr19: 58449649-58450000), *LRP1B* (Chr2: 142604668-142604910), *LHX4* (Chr1: 178469412-178469574), *NKX2-3* (Chr10: 101282587-101282884), *ALDH1L1* (Chr3: 127381582-127381801), *EFNA5* (Chr5: 107035339-107035619), *CCND1* (Chr11: 69165258-69165515), *DLC-1* (Chr8: 13034845-13035136), *TGFB2* (Chr1: 216586512-216586822), *ZNF566* (Chr19: 41671827-41672234), *ADAM12* (Chr10: 128066859-128067044), *MYOD1* (Chr11: 17697405-17697613), *MME* (Chr3: 156280330-156280527), and *MGMT* (Chr10: 131155100-131155259).

Prior to sequencing, bisulfite treatment was performed. Bisulfite treatment converts all unmethylated cytosines to uracil while methylated cytosines remain unaltered after the treatment. Thus, by aligning the sequences of the bisulfite treated DNA and comparing altered/unaltered cytosines, we can measure DNA methylation level. 454 pyro-sequencing on the bisulfite treated DNA is the most accurate method that can be used to measure DNA methylation. As a result, a total of 294,631 sequences was generated with an average read length of 131 bp (range 35-300bp).

From the sequences of bisulfite treated cells, we collected sequences of 30 bases centered around each CpG site and grouped the sequences into two classes: *methylation susceptible site sequences (MS)* and *methylation resistant site sequences (RS)*. See "Method" Section for detail. There were 41 CpG sites, thus 41 sequences in MS. We randomly selected 41 sequences (41 CpG sites) out of 84 sequences (81 CpG sites) in RS and used the sequence sets to build predictive models.

Pacific Symposium on Biocomputing 13:315-326(2008)

5. Method

5.1. *Estimating methylation level of a CpG site*

The methylation level of a CpG site was estimated by counting the number of `C`'s and `T`'s in two columns that are predominantly either `CG` or `TG`. For example, such columns are highlighted in the alignment in Figure 1.

Figure 1. An alignment of bisulfite treated sequences and identification of methylated sites.

Intuitively, counting characters in an alignment of sequences will give us a good estimation of methylation level of a CpG site. There are two problems with this simple approach. First, aligning 2,000 to 3,000 sequences for each sequenced region of 25 genes is very time consuming. Second, even though we use a high performance machine to align the sequences, it is only an estimation of methylation level of a CpG site. In particular, there are sequencing errors which makes the estimation of methylation level of a CpG site more complicated. We used a sequence sampling technique to estimate methylation level as follows:

(1) Sample 20% of sequences in each DNA region. This results in a set of 400 to 500 sequences, which is large enough to estimate methylation level and also can be aligned using ClustalW [22] in a reasonable amount of time.
(2) Then look for two columns where predominantly either CG (methylated) or TG (unmethylated). Estimation of methylation level of a CpG site is computed by counting the number of CG's.

We repeated the sampling task 25 times, so there were 25 estimated methylation level per each CpG site. Then we defined a CpG site as a *methylation susceptible site* when the estimated methylation level $X \geq$

Pacific Symposium on Biocomputing 13:315-326(2008)

$T_{susceptible}$ with a p-value less than 0.01, assuming X following a Gaussian distribution. We define a CpG site as a *methylation resistant site* when the estimated methylation level $X \leq T_{resistant}$ with a p-value less than 0.01, assuming X following a Gaussian distribution. The two threshold values are set to $T_{susceptible} = 0.5$ and $T_{resistant} = 0.01$.

5.2. *Preparing input data to prediction algorithms*

We collected sequences of 30 bases around each CpG site and grouped the sequences into two classes: methylation susceptible site sequences (MS) and methylation resistant site sequences (RS). An alignment of sequences in MS has 30 columns, each of which becomes an attribute. Each sequence in MS is designated as "UP" class label. We used the WEKA package [23], thus each attribute has four possible values:

```
@attribute C1 {A,T,G,C}
```

and each sequence of 30 bases in MS is represented as

```
T,T,T,A,T,T,T,A,T,T,G,T,A,A,C,G,G,T,T,A,A,G,G,T,T,G,G,T,T,T,UP
```

Sequences in RS were represented in the same way as those in MS, except the class label being designated as "DOWN."

For classification tests, we used four machine learning algorithms in WEKA: SMO that implements John C. Platt's sequential minimal optimization algorithm [21] for training a support vector classifier using polynomial or RBF kernels; IBk-type classifier [17] that is a simple distance measure to find the training instance closest to the given test instance, and predicts the same class as this training instance; Multilayer Perceptron that uses backpropagation to classify instances (all nodes are sigmoid); and a Naive Bayes classifier [18].

5.3. *Character composition analysis*

To analyze character composition of CpG flanking sequences, we compared sequences in MS and RS using Weblogo [19] (`http://weblogo.berkeley.edu`) and the two sample logo [14] (`http://www.twosamplelogo.org/`). Weblogo uses only one sequence input file and compare its character frequencies to a random model, thus we used MS and RS separately and generated two Weblogos. The two sample logo uses two input sequence sets simultaneously, thus the name "two sample", and effectively highlighted character composition difference in MS and RS.

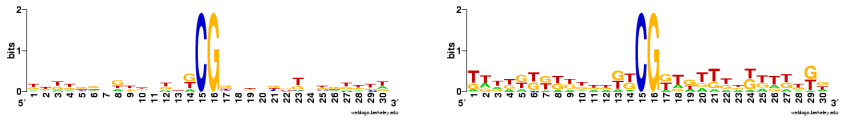
Pacific Symposium on Biocomputing 13:315-326(2008)

6. Result

6.1. *CpG methylation is not uniform*

Our data analysis of methylation level showed that methylation level was not uniform even in the upstream region of a single gene (see Figure 5 for *CYP27B* gene). Thus we conjectured that there should be biological mechanisms, possibly sequence specific ones, for DNA methylation, which inspired us to investigate into building predictive models for methylation susceptibility.

6.2. *Analysis of character composition*



(A) methylation susceptible (B) resistant

Figure 2. Relative entropy of CG flanking sequences using WebLogo. Logos of methylation susceptible sequences and Logos of methylation resistant sequences.

We computed logos (relative entropy with respect to a random model) of CG flanking sequences using WebLogo. As shown in Figure 2, flanking sequences of susceptible CpG sites (MS) are more like random characters while flanking sequences of methylation resistant CpG sites (RS) consistently lack cytosine (C) with adenine (A), guanine (G), and thymine (T) over-represented.

To further investigate how different character composition of sequences in MS and RS, we used the two sample logo technique [14] (`http://www.twosamplelogo.org/`).

The two sample logo in Figure 3 highlights character composition difference clearly: characters in the upper panel for methylation susceptible sequences and characters in the lower panel for methylation resistant sequences. The over-represented characters in the two sample logo analysis agree well with the methylation susceptibility experiments using DNMT1 in Handa and Jeltsch's analysis [20], which reported flanking sequences of up to +/-four base-pairs surrounding the central CG site that were characteristic of high (5'-CTTGCGCAAG-3') and low (5'-TGTTCGGTGG-3') levels of methylation in human genomic DNA. Five positions of the two sample logo (11, 12, 14, 17, and 18) agreed with Handa and Jeltsch's analysis. Only two

(13 and 20 in the two sample logo) out of eight flanking positions disagree with Handa and Jeltsch analysis [20]. However, in Handa's analysis, G is prominent in both methylation susceptible and resistant sequences. In our analysis, T is prominent in methylation susceptible sequences, which may be worth further investigation. In summary, among eight CpG flanking characters, only one position (13 in the two sample logo) disagrees with a published methylation susceptibility analysis.

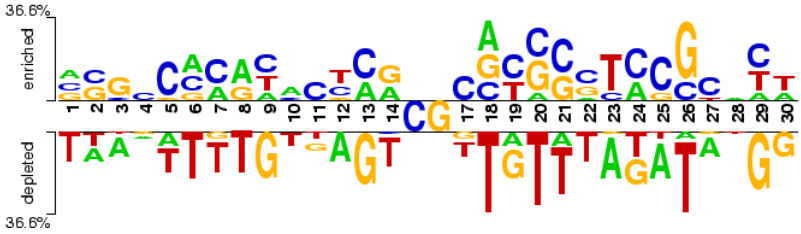


Figure 3. Two sample logo plot showing the DNA character composition difference.

6.3. *Predicitive model for methylation susceptibility*

Since this is to determine either of two classes, UP and DOWN, we used four classification methods in WEKA: SMO, Multilayer Perceptron (MP), Naïve Bayes (NB), Instance Based Classifier (IBk). The prediction accuracy was measured with the standard 10 fold cross validation. As shown in Figure 4, all four algorithms achieved over 70% accuracy with 30bp flanking sequences. We expect that the prediction accuracy drops as the length of the flanking sequences decreases. To measure the effect of flanking sequence length, we used flanking sequences of 30 bp to 2bp with a 2bp decrease in length (one from the left end and the other from the right end). The prediction accuracy of SMO, NB, and MP did not drop much until the flanking sequence length was reduced to 10bp, which agrees with the experimental result in Handa and Jeltsch's analysis.

6.4. *Is this cancer specific?*

Given that methylation levels are measured for four types of primary lymphoma and leukemia cells and normal peripheral blood lymphocytes, it is natural to ask whether there is difference in methylation level between cancer types and normal cells. Our quick, initial analysis was not able to find

Pacific Symposium on Biocomputing 13:315-326(2008)

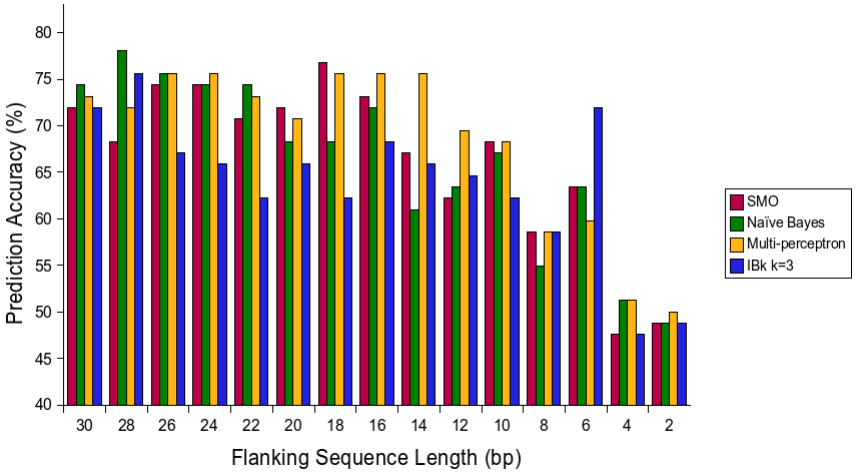


Figure 4. Prediction accuracy of four classification models using 10 fold cross validation. X-axis is the length of CpG flanking sequences.

distinctive methylation patterns between four leukemia cells and one normal cells, as shown in Figure 5 for *CYP27B* gene. In general, we observed that methylation levels in leukemia cell lines were higher than in the normal cell lines, as expected. However, this was not true at all CpG sites. For example, some CpG sites in the upstream region of the *CYP27B* gene showed higher methylation in the normal cell line compared to the cancer cell lines. We plan to investigate this question further with new data sets.

7. Discussion

In this paper, we utilized CpG site specific methylation information to characterize CpG site methylation susceptibility. First, we showed that there was significant difference in DNA character composition between methylation susceptible and resistant sequences. In particular, comparison of methylation susceptible and resistant sequences using the two sample logo technique showed that over-represented characters in methylation susceptible sequences are in agreement with the analysis by Handa and Jeltsch showing CpG flanking sequence specificity for methylation susceptibility. Secondly, we used the CpG flanking sequences to build predictive models for methylation susceptibility and achieved over 75% prediction accuracy in 10 fold cross validation tests. This study is first of its kind that uses CpG site specific methylation information to build predictive models.

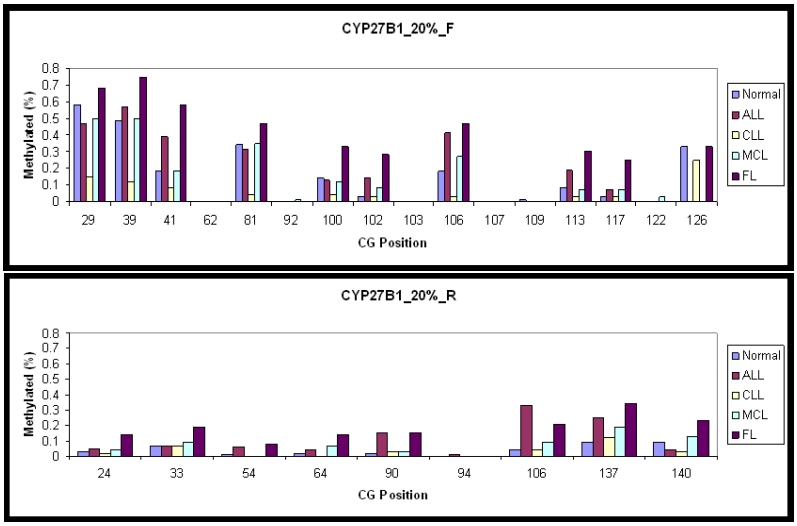


Figure 5. Methylation level of each CpG sites in the upstream region of *CYP27B* gene. Two fragments, Forward (upper panel) and Reverse (lower panel), were sequenced after bisulfite treatment.

Further study includes characterization of leukemia specific methylation pattern signatures and related sequence and machine learning analysis.

Acknowledgement

This work is supported by the National Cancer Institute grants U54 CA11300 and R01 CA85289.

References

1. Jaenisch, R. and Bird, A. *Nat Genet*, 33 Suppl: 245-254, 2003.
2. Feinberg AP, Tycko B *Nature Reviews Cancer* 2004 4: 143-153
3. Herman JG, Baylin SB *New England Journal of Medicine* 2003 349:2042-54
4. Toyota M, Issa JP *Semin Oncol* 2005 32:521-30
5. Nephew KP, Huang TH *Cancer Letters* 2003 190:125-33
6. Jones P A *Semin Hematol*, 2005 42: S3-8, 2005.
7. Keshet I, Schlesinger Y, Farkash S, Rand E, Hecht M, et al. *Nat Genet* 2006 38: 149-153.
8. Goh L, Murphy SK, Muhkerjee S, Furey TS *Bioinformatics* 2007 23: 281-288.
9. Fang F, Fan S, Zhang X, Zhang MQ *Bioinformatics* 2006 22: 2204-2209.
10. Bock C, Walter J, Paulsen M, Lengauer T *PLoS Comput Biol* 2007 3: e110.
11. Herman JG, Graff JR, Myohanen S, Nelkin BD, Baylin S *Proc Natl Acad Sci U S A* 1996 Sep 3;93(18):9821-6

Pacific Symposium on Biocomputing 13:315-326(2008)

12. Yan PS, Chen CM, Shi H, Rahmatpanah F, Wei SH, Caldwell CW, Huang TH. *Cancer Res* 2001; 61:8375-80.
13. Weber M, Davies JJ, Wittig D, Oakeley EJ, Haase M, Lam WL, Schubeler D. *Nat Genet* 2005; 37:853-62.
14. Vacic V., Iakoucheva L.M., and Radivojac P. *Bioinformatics*, 22(12): 1536-1537. 2006
15. Taylor K.H., Kramer R.S., Davis J.W., Xu D., Caldwell1 C.W., and Shi H., *Cancer Research*, in press. 2007
16. Zheng Z. and Webb G., *Machine Learning*, 41(1): 53-84. 2000
17. Aha D. and Kibler D., *Machine Learning*, 6:37-66. 1991
18. John G.H. and Langley P. *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*. pp. 338-345. 1995
19. Crooks G.E., Hon G., Chandonia J.M., Brenner S.E. *Genome Res.*, 14:1188-1190, 2004
20. Handa V. and Jeltsch A. *J Mol Biol.* 348(5):1103-12. 2005
21. Platt J. *Advances in Kernel Methods - Support Vector Learning*. 1998
22. Chenna R., Sugawara H., Koike T., Lopez R., Gibson T.J., Higgins D.G., Thompson J.D. *Nucleic Acids Res.* 31(13):3497-500. 2003
23. Witten I. H. and Frank E. *Data Mining: Practical Machine Learning Tools and Techniques* Morgan Kaufmann, San Francisco, 2 edition, 2005.
24. Feltus F.A., Lee E.K., Costello J.F., Plass C. and Vertino P.M., *Proc Natl Acad Sci U S A* 100(21):12253-8. 2003.
25. Feltus F.A., Lee E.K., Costello J.F., Plass C., Vertino P.M., *Genomics* 87(5):572-9. 2006.
26. Bock C., Paulsen M., Tierling S., Mikeska T., Lengauer T., Walter J., *PLoS Genet.* **2(3)** e26, 2006.
27. Fang F., Fan S., Zhang X., Zhang M.Q., *Bioinformatics.* 22(18):2204-9. 2006.
28. Das R., Dimitrova N., Xuan Z., Rollins R.A., Haghighi F., Edwards J.R., Ju J., Bestor T.H., Zhang M.Q., *Proc Natl Acad Sci U S A.* 103(28):10713-6. 2006.
29. Goh L., Murphy S.K., Muhkerjee S., Furey T.S., *Bioinformatics.* 23(3):281-8, 2006.
30. Margulies, M. Eghold, M. et al. *Nature* 2005 Sep 15; 437(7057):326-7
31. Taylor KH, Kramer RS, Davis JW, Guo J, Duff DJ, Xu D, Caldwell CW, Shi H. *Cancer Res* 2007 67: 8511-8518