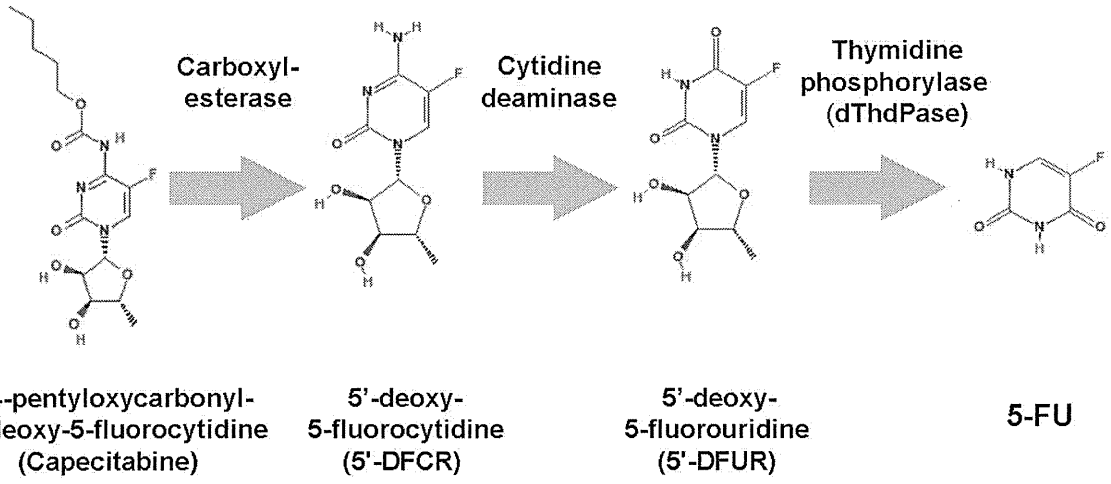
with a C4-C6 alkyl chain were the most susceptible to human carboxylesterase, which led to the development of capecitabine. In 1998, Ishikawa *et al.* at the Nippon Roche Research Center investigated the efficacy of capecitabine and 5-FU in xenograft models implanted with human colon cancer cells [46]. Their results supported the notion that the inefficient conversion of 5'-DFUR to 5-FU by dThdPase in tumors would represent a mechanism of resistance. In contrast, even in tumors with sufficient levels of dThdPase, capecitabine was not effective if DPD levels were very high, and its efficacy was consequently found to be well-correlated with and dependent on the ratio of these two enzymes – dThdPase and DPD – in tumors [46]. The efficacy of capecitabine can be optimized by selecting patients who have tumors with a high ratio of dThdPase to DPD activities.

Figure 4. The metabolism of capecitabine.



HFS is a cutaneous adverse event that occurs in some patients treated with fluoropyrimidines, which can severely disrupt the daily lives of patients. It is also a leading cause of interruption of capecitabine regimens as well [47]. In order to test the hypothesis that the occurrence of HFS could be related to tissue-specific expression of drug-metabolizing enzymes in the skin of the palms and soles, Milano *et al.* measured the expression of dThdPase (activation pathway), DPD (catabolic pathway) and cell proliferation (Ki67) in the skin of the palm (target tissue for HFS) and of the lower back (control area) with punch biopsy specimens [48]. Their study revealed that dThdPase and DPD expression levels were significantly greater in the palm relative to the back, and that dThdPase-facilitated local production of 5-FU in the palm during capecitabine treatment could explain the occurrence of HFS. In addition, the accumulated findings from clinical trials show the benefits of DPD inhibition on decreasing the risk of HFS [47].

The efficacy of co-administration of a series of DPD inhibitors with capecitabine has been investigated. A DPD inhibitor, RO0094889, which is a prodrug of 5-vinyluracil, was designed to generate 5-vinyluracil selectively in tumor tissues by sequential conversion by three enzymes responsible for the metabolism of capecitabine [49]. RO0094889 and various DPD inhibitors have been analyzed for co-administration with capecitabine. Nevertheless, HFS occurs more frequently with 5-FU delivered by continuous infusion [5] or with the 5-FU oral derivative capecitabine, rather than with bolus 5-FU therapy.

4. Conclusions

Recently clinical studies on S-1 and capecitabine, comparing their efficacy and adverse events, have been reported, mainly from Korea [50,51]. The accumulated results will provide benefits that can optimize the treatment of cancer patients. The information obtained from the studies described in this review may give us better direction for the appropriate use of the oral 5-FU drugs. For example, the assessment of the dThdPase and DPD levels may provide evidence of patients who would be good/poor responders to therapy. Patients with low dThdPase activity and inefficient conversion of 5'-DFUR to 5-FU, may present resistance to capecitabine. The activities of carboxylesterase and cytidine deaminase may also affect the efficacy of capecitabine. Among patients with high DPD activity, S-1 may exhibit better efficacy; on the other hand, capecitabine may show more powerful effects along with DPD inhibitors in tumor cells. Although recent studies support the notion that the continuous 5-FU infusion chemotherapies can be replaced with orally-administrable 5-FU drugs in some regimens, it will be necessary for us to remember that the metabolism of orally-administered 5-FU differs from that of infusional 5-FU, because orally-administered 5-FU undergoes more diverse metabolism in the gastrointestinal tract and in the liver, with various enzymes. On the other hand, it is essential to elucidate the pharmacokinetic mechanism of each of the newly-developed drugs, to ensure the selection of the proper drug(s) for each patient in the clinical setting, and to further develop the optimized drug derivatives. This will require the collaboration of clinicians, molecular biologists and preclinical drug researchers.

Acknowledgements

The authors declare no conflicts of interest in connection with the current study.

References

1. Heidelberger, C.; Chaudhuri, N.K.; Danneberg, P.; Mooren, D.; Griesbach, L.; Duschinsky, R.; Schnitzer, R.J.; Pleven, E.; Scheiner, J. Fluorinated pyrimidines, a new class of tumour-inhibitory compounds. *Nature* **1957**, *179*, 663–666.
2. Rutman, R.J.; Cantarow, A.; Paschkis, K.E. The catabolism of uracil *in vivo* and *in vitro*. *J. Biol. Chem.* **1954**, *210*, 321–329.
3. Handschumacher, R.E.; Welch, A.D. Microbial studies of 6-azauracil, an antagonist of uracil. *Cancer Res.* **1956**, *16*, 965–969.
4. Skipper, H.E.; Schabel, F.M. Jr.; Wilcox, W.S. Experimental evaluation of potential anticancer agents. XIII. On the criteria and kinetics associated with "curability" of experimental leukemia. *Cancer Chemother. Rep.* **1964**, *35*, 1–111.
5. Meta-analysis Group In Cancer. Efficacy of intravenous continuous infusion of fluorouracil compared with bolus administration in advanced colorectal cancer. *J. Clin. Oncol.* **1998**, *16*, 301–308.
6. Saif, M.W.; Syrigos, K.N.; Katirtzoglou, N.A. S-1: A promising new oral fluoropyrimidine derivative. *Expert Opin. Investig. Drugs* **2009**, *18*, 335–348.

7. Wohlhueter, R.M.; McIvor, R.S.; Plagemann, P.G. Facilitated transport of uracil and 5-fluorouracil, and permeation of orotic acid into cultured mammalian cells. *J. Cell. Physiol.* **1980**, *104*, 309–319.
8. Longley, D.B.; Harkin, D.P.; Johnston, P.G. 5-fluorouracil: mechanisms of action and clinical strategies. *Nat. Rev. Cancer* **2003**, *3*, 330–338.
9. Houghton, J.A.; Houghton, P.J.; Wooten, R.S. Mechanism of induction of gastrointestinal toxicity in the mouse by 5-fluorouracil, 5-fluorouridine, and 5-fluoro-2'-deoxyuridine. *Cancer Res.* **1979**, *39*, 2406–2413.
10. Schuetz, J.D.; Wallace, H.J.; Diasio, R.B. 5-fluorouracil incorporation into DNA of CF-1 mouse bone marrow cells as a possible mechanism of toxicity. *Cancer Res.* **1984**, *44*, 1358–1363.
11. Meta-Analysis Group In Cancer. Toxicity of fluorouracil in patients with advanced colorectal cancer: effect of administration schedule and prognostic factors. *J. Clin. Oncol.* **1998**, *16*, 3537–3541.
12. Diasio, R.B.; Harris, B.E. Clinical pharmacology of 5-fluorouracil. *Clin. Pharmacokinet.* **1989**, *16*, 215–237.
13. Heggie, G.D.; Sommadossi, J.P.; Cross, D.S.; Huster, W.J.; Diasio, R.B. Clinical pharmacokinetics of 5-fluorouracil and its metabolites in plasma, urine, and bile. *Cancer Res.* **1987**, *47*, 2203–2206.
14. Koenig, H.; Patel, A. Biochemical basis for fluorouracil neurotoxicity. The role of Krebs cycle inhibition by fluoroacetate. *Arch. Neurol.* **1970**, *23*, 155–160.
15. Okeda, R.; Shibutani, M.; Matsuo, T.; Kuroiwa, T.; Shimokawa, R.; Tajima, T. Experimental neurotoxicity of 5-fluorouracil and its derivatives is due to poisoning by the monofluorinated organic metabolites, monofluoroacetic acid and alpha-fluoro-beta-alanine. *Acta Neuropathol.* **1990**, *81*, 66–73.
16. Matsubara, I.; Kamiya, J.; Imai, S. Cardiotoxic effects of 5-fluorouracil in the guinea pig. *Jpn. J. Pharmacol.* **1980**, *30*, 871–879.
17. Santi, D.V.; McHenry, C.S. 5-Fluoro-2'-deoxyuridylate: covalent complex with thymidylate synthetase. *Proc. Natl. Acad. Sci. USA* **1972**, *69*, 1855–1857.
18. Jackson, R.C.; Grindley, G.B. The biochemical basis for methotrexate cytotoxicity. In *Folate Antagonists as Therapeutic Agents*, 2nd edition; Sirotnak, F.M., Burchell, J.J., Ensminger, W.D., Eds.; Academic Press: New York, NY, USA, 1984; Volume 1, pp. 289–315.
19. Yoshioka, A.; Tanaka, S.; Hiraoka, O.; Koyama, Y.; Hirota, Y.; Ayusawa, D.; Seno, T.; Garrett, C.; Wataya, Y. Deoxyribonucleoside triphosphate imbalance. 5-Fluorodeoxyuridine-induced DNA double strand breaks in mouse FM3A cells and the mechanism of cell death. *J. Biol. Chem.* **1987**, *262*, 8235–8241.
20. Mitrovski, B.; Pressacco, J.; Mandelbaum, S.; Erlichman, C. Biochemical effects of folate-based inhibitors of thymidylate synthase in MGH-U1 cells. *Cancer Chemother. Pharmacol.* **1994**, *35*, 109–114.
21. Grem, J.L.; Fischer, P.H. Enhancement of 5-fluorouracil's anticancer activity by dipyridamole. *Pharmacol. Ther.* **1989**, *40*, 349–371.

22. Showalter, S.L.; Showalter, T.N.; Witkiewicz, A.; Havens, R.; Kennedy, E.P.; Hucl, T.; Kern, S.E.; Yeo, C.J.; Brody, J.R. Evaluating the drug-target relationship between thymidylate synthase expression and tumor response to 5-fluorouracil. Is it time to move forward? *Cancer Biol. Ther.* **2008**, *7*, 986–994.
23. Lembersky, B.C.; Wieand, H.S.; Petrelli, N.J.; O'Connell, M.J.; Colangelo, L.H.; Smith, R.E.; Seay, T.E.; Giguere, J.K.; Marshall, M.E.; Jacobs, A.D.; *et al.* Oral uracil and tegafur plus leucovorin compared with intravenous fluorouracil and leucovorin in stage II and III carcinoma of the colon: results from National Surgical Adjuvant Breast and Bowel Project Protocol C-06. *J. Clin. Oncol.* **2006**, *24*, 2059–2064.
24. Boku, N.; Yamamoto, S.; Fukuda, H.; Shirao, K.; Doi, T.; Sawaki, A.; Koizumi, W.; Saito, H.; Yamaguchi, K.; Takiuchi, H.; *et al.* Fluorouracil *versus* combination of irinotecan plus cisplatin *versus* S-1 in metastatic gastric cancer: A randomised phase 3 study. *Lancet Oncol.* **2009**, *10*, 1063–1069.
25. Mansfield, P.F.; Hohn, D.C.; Fornage, B.D.; Gregurich, M.A.; Ota, D.M. Complications and failures of subclavian-vein catheterization. *N. Engl. J. Med.* **1994**, *331*, 1735–1738.
26. Agnelli, G.; Verso, M. Therapy Insight: venous-catheter-related thrombosis in cancer patients. *Nat. Clin. Pract. Oncol.* **2006**, *3*, 214–222.
27. Lokich, J.J.; Moore, C.L.; Anderson, N.R. Comparison of costs for infusion *versus* bolus chemotherapy administration–Part two. Use of charges *versus* reimbursement for cost basis. *Cancer* **1996**, *78*, 300–303.
28. Giller, S.A.; Zhuk, R.A.; Lidak, M.Iu. Analogs of pyrimidine nucleosides. I. N1-(alpha-furanidyl) derivatives of natural pyrimidine bases and their antimetabolities. *Dokl. Akad. Nauk. SSSR.* **1967**, *176*, 332–335 (article in Russian).
29. Toide, H.; Akiyoshi, H.; Minato, Y.; Okuda, H.; Fujii, S. Comparative studies on the metabolism of 2-(tetrahydrofuryl)-5-fluorouracil and 5-fluorouracil. *Gann* **1977**, *68*, 553–560.
30. Fujii, S.; Ikenaka, K.; Fukushima, M.; Shirasaka, T. Effect of uracil and its derivatives on antitumor activity of 5-fluorouracil and 1-(2-tetrahydrofuryl)-5-fluorouracil. *Gann* **1978**, *69*, 763–772.
31. El Sayed, Y.M.; Sadée, W. Metabolic activation of R,S-1-(tetrahydro-2-furanyl)-5-fluorouracil (ftorafur) to 5-fluorouracil by soluble enzymes. *Cancer Res.* **1983**, *43*, 4039–4044.
32. Rustum, Y.M. Mechanism-based improvement in the therapeutic selectivity of 5-FU prodrug alone and under conditions of metabolic modulation. *Oncology* **1997**, *54* (Suppl. 1), 7–11.
33. Diasio, R.B. The role of dihydropyrimidine dehydrogenase (DPD) modulation in 5-FU pharmacology. *Oncology* **1998**, *12*, 23–27.
34. Fujii, S.; Kitano, S.; Ikenaka, K.; Shirasaka, T. Effect of coadministration of uracil or cytosine on the anti-tumor activity of clinical doses of 1-(2-tetrahydrofuryl)-5-fluorouracil and level of 5-fluorouracil in rodents. *Gann* **1979**, *70*, 209–214.
35. Hoff, P.M.; Cassidy, J.; Schmoll, H.J. The evolution of fluoropyrimidine therapy: From intravenous to oral. *Oncologist* **2001**, *6* (Suppl. 4), 3–11.

36. Poon, M.A.; O'Connell, M.J.; Wieand, H.S.; Krook, J.E.; Gerstner, J.B.; Tschetter, L.K.; Levitt, R.; Kardinal, C.G.; Mailliard, J.A. Biochemical modulation of fluorouracil with leucovorin: confirmatory evidence of improved therapeutic efficacy in advanced colorectal cancer. *J. Clin. Oncol.* **1991**, *9*, 1967–1972.
37. Ichikura, T.; Tomimatsu, S.; Okusa, Y.; Yahara, T.; Uefuji, K.; Tamakuma, S. Thymidylate synthase inhibition by an oral regimen consisting of tegafur-uracil (UFT) and low-dose leucovorin for patients with gastric cancer. *Cancer Chemother. Pharmacol.* **1996**, *38*, 401–405.
38. Cook, A.F.; Holman, M.J.; Kramer, M.J.; Trown, P.W. Fluorinated pyrimidine nucleosides. 3. Synthesis and antitumor activity of a series of 5'-deoxy-5-fluoropyrimidine nucleosides. *J. Med. Chem.* **1979**, *22*, 1330–1335.
39. Ishitsuka, H.; Miwa, M.; Takemoto, K.; Fukuoka, K.; Itoga, A.; Maruyama, HB. Role of uridine phosphorylase for antitumor activity of 5'-deoxy-5-fluorouridine. *Gann* **1980**, *71*, 112–123.
40. Shirasaka, T.; Shimamato, Y.; Ohshimo, H.; Yamaguchi, M.; Kato, T.; Yonekura, K.; Fukushima, M. Development of a novel form of an oral 5-fluorouracil derivative (S-1) directed to the potentiation of the tumor selective cytotoxicity of 5-fluorouracil by two biochemical modulators. *Anticancer Drugs* **1996**, *7*, 548–557.
41. Tatsumi, K.; Fukushima, M.; Shirasaka, T.; Fujii, S. Inhibitory effects of pyrimidine, barbituric acid and pyridine derivatives on 5-fluorouracil degradation in rat liver extracts. *Jpn. J. Cancer Res.* **1987**, *78*, 748–755.
42. Shirasaka, T.; Shimamoto, Y.; Fukushima, M. Inhibition by oxonic acid of gastrointestinal toxicity of 5-fluorouracil without loss of its antitumor activity in rats. *Cancer Res.* **1993**, *53*, 4004–4009.
43. Muneoka, K.; Shirai, Y.; Yokoyama, N.; Wakai, T.; Hatakeyama, K. 5-Fluorouracil cardiotoxicity induced by alpha-fluoro-beta-alanine. *Int. J. Clin. Oncol.* **2005**, *10*, 441–443.
44. Miwa, M.; Ura, M.; Nishida, M.; Sawada, N.; Ishikawa, T.; Mori, K.; Shimma, N.; Umeda, I.; Ishitsuka, H. Design of a novel oral fluoropyrimidine carbamate, capecitabine, which generates 5-fluorouracil selectively in tumours by enzymes concentrated in human liver and cancer tissue. *Eur. J. Cancer* **1998**, *34*, 1274–1281.
45. Shimma, N.; Umeda, I.; Arasaki, M.; Murasaki, C.; Masubuchi, K.; Kohchi, Y.; Miwa, M.; Ura, M.; Sawada, N.; Tahara, H.; *et al.* The design and synthesis of a new tumor-selective fluoropyrimidine carbamate, capecitabine. *Bioorg. Med. Chem.* **2000**, *8*, 1697–1706.
46. Ishikawa, T.; Utoh, M.; Sawada, N.; Nishida, M.; Fukase, Y.; Sekiguchi, F.; Ishitsuka, H. Tumor selective delivery of 5-fluorouracil by capecitabine, a new oral fluoropyrimidine carbamate, in human cancer xenografts. *Biochem. Pharmacol.* **1998**, *55*, 1091–1097.
47. Yen-Revollo, J.L.; Goldberg, R.M.; McLeod, H.L. Can inhibiting dihydropyrimidine dehydrogenase limit hand-foot syndrome caused by fluoropyrimidines? *Clin. Cancer Res.* **2008**, *14*, 8–13.
48. Milano, G.; Etienne-Grimaldi, M.C.; Mari, M.; Lassalle, S.; Formento, J.L.; Francoual, M.; Lacour, J.P.; Hofman, P. Candidate mechanisms for capecitabine-related hand-foot syndrome. *Br. J. Clin. Pharmacol.* **2008**, *66*, 88–95.

49. Hattori, K.; Kohchi, Y.; Oikawa, N.; Suda, H.; Ura, M.; Ishikawa, T.; Miwa, M.; Endoh, M.; Eda, H.; Tanimura, H.; *et al.* Design and synthesis of the tumor-activated prodrug of dihydropyrimidine dehydrogenase (DPD) inhibitor, RO0094889 for combination therapy with capecitabine. *Bioorg. Med. Chem. Lett.* **2003**, *13*, 867–872.
50. Lee, J.L.; Kang, Y.K.; Kang, H.J.; Lee, K.H.; Zang, D.Y.; Ryoo, B.Y.; Kim, J.G.; Park, S.R.; Kang, W.K.; Shin, D.B.; *et al.* A randomised multicentre phase II trial of capecitabine vs S-1 as first-line treatment in elderly patients with metastatic or recurrent unresectable gastric cancer. *Br. J. Cancer* **2008**, *99*, 584–590.
51. Seol, Y.M.; Song, M.K.; Choi, Y.J.; Kim, G.H.; Shin, H.J.; Song, G.A.; Chung, J.S.; Cho, G.J. Oral fluoropyrimidines (capecitabine or S-1) and cisplatin as first line treatment in elderly patients with advanced gastric cancer: a retrospective study. *Jpn. J. Clin. Oncol.* **2009**, *39*, 43–48.

PeakRegressor identifies composite sequence motifs responsible for STAT1 binding sites and their potential rSNPs

Jean-François Pessiot[1], Hirokazu Chiba[1], Hiroto Hyakkoku[2,1], Takeaki Taniguchi[3], Wataru Fujibuchi[1*]

[1]Computational Biology Research Center, Advanced Industrial Science and Technology (AIST),

[2]Waseda University, [3]Mitsubishi Research Institute, Inc.

Abstract

How to identify true transcription factor binding sites on the basis of sequence motif information (e.g., motif pattern, location, combination, etc.) is an important question in bioinformatics. We present "PeakRegressor", a system that identifies binding motifs by combining DNA-sequence data and ChIP-Seq data. PeakRegressor uses L1-norm log linear regression in order to predict peak values from binding motif candidates. Our approach successfully predicts the peak values of STAT1 and Pol II with correlation coefficients as high as 0.65 and 0.66, respectively. Using PeakRegressor, we are able to identify composite motifs for STAT1, as well as potential regulatory SNPs (rSNPs) involved in the regulation of transcription levels of neighboring genes.

1 Introduction

The experimental identification of *cis*-regulatory sites based on transcription factor binding motifs (TFBMs) is a difficult and time-consuming task. In this regard, *in silico* analysis of TFBMs has recently attracted attention as a promising tool for discovering true *cis*-regulatory sites. Previous works attempt to find TFBMs to model the mechanisms underlying the control of gene expression levels[2, 4]. They assume that the gene expression levels are determined by the presence of certain motifs in the upstream regions of the genes. Based on this assumption, they find TFBM candidates which show a strong correlation with changes in the gene expression levels.[5] Instead of modeling the expression levels, another solution is to model the binding affinities between a protein and its target genes based on the thermodynamics theory. However, the binding affinities are difficult to measure and related works use transcription factor occupancy to

*Corresponding Author - w.fujibuchi@aist.go.jp

approximate binding affinity[6, 7].

In this article, we present PeakRegressor, a new tool for the identification of functional TFBMs from ChIP-Seq data. As far as we know, this is the first attempt to perform peak signal regression based on candidate motif models. Our contribution is twofold. First, in contrast with previous approaches, we use the peak scores (provided by[9]) as a surrogate for the binding affinities. We argue that they provide more accurate approximations and therefore lead to better identification of functional TFBMs. Second, our approach identifies not only primary TFBM candidates but also secondary motifs that may often synergistically strengthen or weaken the binding. The rest of this paper is organized as follows. We describe PeakRegressor in section 2. In section 3, we illustrate the performance of our approach on two ChIP-Seq datasets and discuss its ability to identify the binding motifs of STAT1 and Pol II.

2 PeakRegressor System to Find Functional TFBMs

PeakRegressor is a system to find TFBMs that are statistically important for transcription factor binding signals, by taking ChIP-Seq data as input, and outputs a list of TFBM candidates. The workflow is summarized in Figure 1.

Step 1 First, we define the peak sequences as the 200-bp genomic regions centered around the peaks. Then, we sort the peak sequences according to their ascending scores. We group the peak sequences into clusters such that each cluster contains 200 peaks of consecutive scores. Then, we apply MEME[1] to each peak sequence cluster. For each sequence cluster, MEME is parameterized in ZOOPS mode to find 10 motifs of lengths $8--20$.

This strategy has two advantages. First, it allows us to identify motifs that may be associated with a given binding affinity level. If a cluster contains only low (resp. high) binding affinity peaks, the corresponding sequences may contain weak (resp. strong) binding motifs, i.e., motifs that are specific to low (resp. high) binding affinity. Second, it reduces computational time by parallelizing MEME computations.

Step 2 In order to predict the binding affinity of the peaks, we need to represent each peak as a vector in the motif space. Let seq^i be the DNA sequence of peak i. Let $\text{seq}^i_{j,\ell}$ be the ℓ-length sub-sequence of seq^i, starting from position j. Let S^d be the PSSM of motif d. Let ℓ_i be the length of seq^i and ℓ_d be the length of motif d. We represent peak i as

[1] http://meme.sdsc.edu/

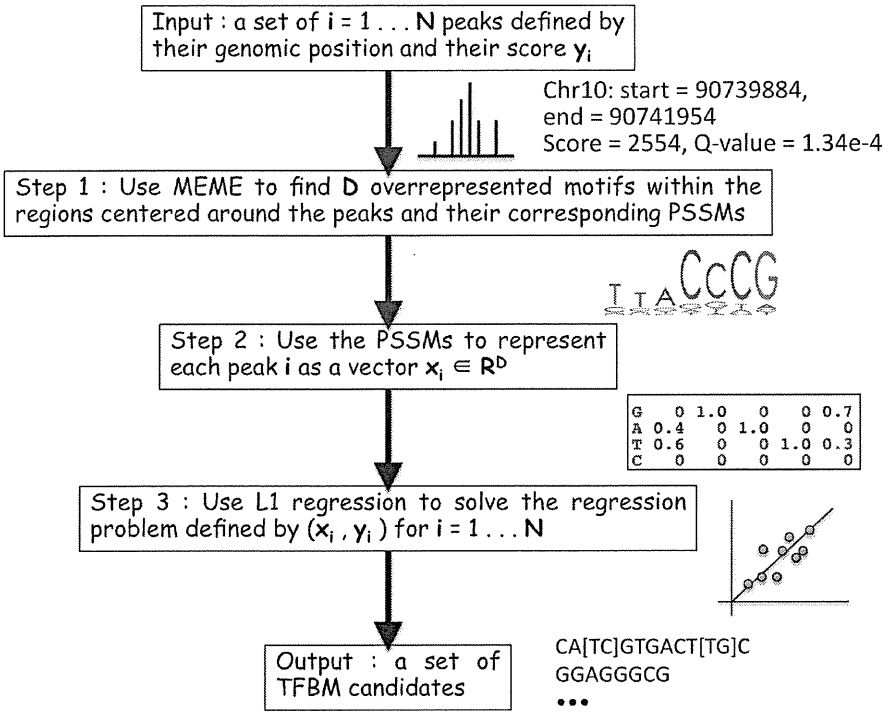


Figure 1: **Schematic view of the workflow of PeakRegressor.** PeakRegressor takes ChIP-Seq data as input and outputs a list of TFBM candidates and their weights that give the best regression accuracies.

vector $x_i \in R^D$, such that

$$x_{id} = \max_{j=1\ldots\ell_i-\ell_d+1} f(\text{seq}^i_{j,\ell_d}, S^d) - \max(S^d)$$

for $d = 1 \ldots D$. The quantity $f(\text{seq}^i_{j,\ell_d}, S^d)$ is a sum of log-odd scores, representing how well motif d matches sub-sequence seq^i_{j,ℓ_d}. Hence, the first term of the sum, x_{id}, corresponds to the best match when we slide motif d along sequence seq^i. The term $\max(S^d)$ is the maximum score achievable by any sequence matching with the motif d. Therefore we always have $x_{id} \leq 0$, with $x_{id} = 0$ for the best possible match.

Step 3 Quantities y_i to be fitted are the log values of the peak enrichment scores, as given by PeakSeq[9]. We can now solve the regression problem defined by (x_i, y_i) pairs for $i = 1 \ldots N$. Linear regression is a simple and popular approach, but is prone to overfitting. Hence, we choose to regularize the model with L1-norm, i.e., we want to minimize the sum of squared errors and the L1-norm of the regression coefficient vector:

$$\min_{b \in R^D} \beta \|b\| + \sum_{i=1}^{N} (b^T x_i - y_i)^2,$$

where $\beta > 0$ is a user-defined regularization coefficient. The L1-norm regression is able to select a small number of features that best explain the fitted quantity[10]. In our case, the features correspond to DNA motifs and hence, the result of this step is a set of motifs that best explain the binding signal values from ChIP-Seq dataset. We use Lasso, a popular algorithm for solving L1-norm regression. Lasso is available as part of the LARS package for R[2].

3 Results and Discussion

3.1 Input datasets

We use the ChIP-Seq data provided by[9]. For STAT1, we use 200-bp windows around the peak centers to define the peak sequences. For Pol II, the peak centers are not available and thus, we use the peak start and peak end coordinates to define the peaks. When the length of the resulting sequence is less than 200 bp, we enlarge it in both directions in order to reach 200 bp length. When the length is more than 4000 bp, we trim it in both directions in order to reach 4000 bp length. As a result, all the Pol II peak sequence lengths lie between 200 and 4000 bp.

For the regression analysis, we have to set the regularization parameter β. First, we define $\beta = 2^i$ for $i \in [-25, 25]$. Then for each value of β, we perform a 30 folds cross-validation. In each fold, we split the dataset into a training set and a test set, with a $90\% - 10\%$ ratio. The optimal value for β is the one which corresponds to the lowest prediction error on the test set. All the following results are averaged over the 30 folds cross-validation.

3.2 L1-norm log linear regression

We considered three settings before applying PeakRegressor. In the first setting, we considered all the peaks for regression. In the second setting, we excluded the peaks which showed no overlap with a promoter region (as defined by UCSC dataset[3]). In the third setting, we excluded the peaks which showed high Q-values ($> 10^{-3}$), as provided by [9]. Table 1 shows the averaged correlation coefficients between peak values and their predicted values in the test dataset. We can see that filtering peaks with their Q-values enhances the correlation coefficient for both STAT1 and Pol II. However, when filtering with promoter proximity, we observe than the correlation coefficient improves for Pol II but decreases for STAT1.

In Figure 2, we plot the STAT1 peak scores with two filtering methods such as Q-value $< 10^{-3}$ and promoter proximity in the test dataset against

[2]http://www-stat.stanford.edu/ hastie/Papers/LARS/

[3]http://hgdownload.cse.ucsc.edu/goldenPath/hg18/bigZips/upstream1000.fa.gz

Filtering method	*#Peaks (STAT1/Pol II)*	*STAT1*	*Pol II*
None	36,998 / 24,739	0.50	0.44
Promoter proximity	3,907 / 9,094	0.41	0.53
Q-value $< 10^{-3}$	16,639 / 17,580	0.65	0.66

Table 1: Influence of the peak filtering methods on the correlation coefficients between peak values and their predicted values in the test dataset. The correlation coefficients are averaged in 30-fold cross-validation.

their predictions by PeakRegressor. The correlation coefficient is as high as 0.65 between the peak and predicted values for the Q-value filtering, whilst it is as low as 0.41 for promoter proximity filtering. Interestingly, however, the data points that are selected by promoter proximity exist only in a biased region, leading to worse prediction.

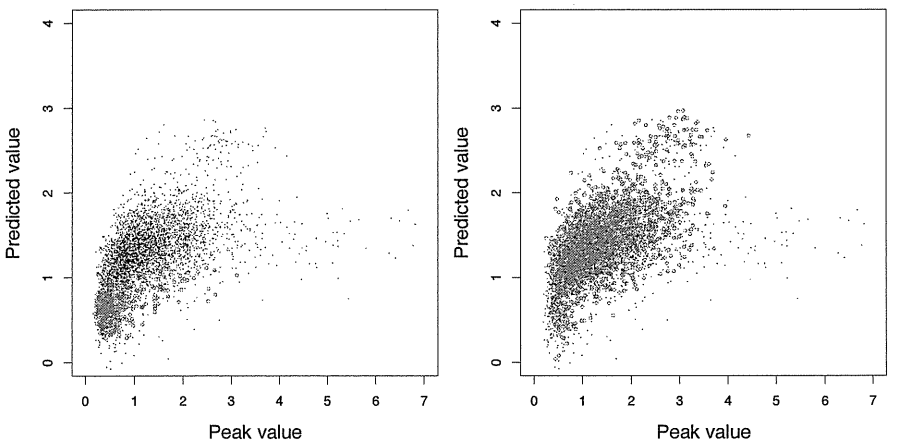


Figure 2: **The STAT1 regression results in test data with two filtering methods (shown by circle): promoter proximity (left) and Q-value (right).** The correlation coefficients between peak values and their predicted values are 0.45 and 0.65 for promoter proximity and Q-value filtering, respectively.

In Tables 2 and 3, we show the top 10 motifs for STAT1 and Pol II identified by PeakRegressor, respectively. The motifs are sorted according to the absolute values of their averaged regression coefficients. A motif with a positive (resp. negative) coefficient is thought to have a strengthening (resp. weakening) effect on the binding. In the case of STAT1, it is clear that our approach correctly identifies the classical GAS motif TTC[TC]N[GA]GAA as the main binding pattern[8]. Meanwhile, the Pol II binding motifs also contain Downstream Promoter Element [AG]G[AT][CT][GAC] and Initiator Site [TC][TC]AN[TA][TC][TC][3].

As the most important feature of PeakRegressor, it can give us a list of

STAT1	*Normalized coef.*
CA[TC]GTGACT[TG]C	1.
[TG]G[GTA][GC][AG]T**TT[CA]C[AGC][GA]GAA**[AC][TG]G[GA][GC]	0.96
TTC[CT][TG][GA]GAAAT[GC][CA][CA][CAT][AT][TCG][CG][CT]	0.72
[CT][TC]CA[GT]**TTCCAGGAA**[AT]T[CG][CAT]C[CT]	0.65
GGAGGGCG	-0.57
GGACGCCG	-0.56
A[CT]**TTC[TC][TG]GGAA**	0.56
TT[CA]C[TAG][GA]GAA[GA]T	0.55
A[TA]**TTCC[CT][GA]GAA**[AC]T[CG][AC]	0.48
TT[CA][TC][GA]GGAA[AG]	0.47

Table 2: List of putative STAT1 binding motifs. The classical GAS motifs are shown in boldface.

Pol II	*Normalized coef.*
T[AG]**A[GC][TAG]CA**[GCT]A[AC]AA	1.
A[GA]AA[AC][CA]AA[AC]AAA	0.78
C[ACT]**[GT][CG][CT][TA]C**C[AGT]CC[TA]	0.76
C[CT][CG][AT]GGCTGG[AG]G	0.68
TTTCTGC[CT][CT]TT[GT]	0.67
T[TA]T[TC]**[CA]CAGACT**[AT]	0.63
GGAGGGAGGC[AG]G	0.62
AC[AC][CA][AC][AT][AG]AGAAA	0.61
TTTG**T[CT][TA]T[TG][AC][AT]**T	0.54
AAA[AT][GC]AAA[AT]A[GA]A	0.54

Table 3: List of putative Pol II binding motifs. The Downstream Promoter Element and Initiator site motifs are shown in boldface.

putative composite motifs. Basically, it is difficult to evaluate whether a composite motif consists of the same motif or multiple (different) motifs. In order to identify the composite motifs, we proceed as follows. First, we consider the best set of motifs according to PeakRegressor (i.e. the set which corresponds to the best prediction accuracy). Among these, we select 136 motifs which have a normalized coefficient higher than 0.1. We use these motifs to represent each peak sequence as a binary vector, indicating whether a motif is present or not in the peak sequence. Then we cluster the resulting peak vectors using the K-means algorithm. Thus each cluster contains peak vectors which show similar motif patterns, i.e. sequences containing potential composite motifs.

Here we show an example of a composite motif that is responsible for STAT1 binding signals:

TCACA[TG]G[ACG] + [TC]TT[CA]C[CA][AG][GC][AC]A.

3.3 Candidate motifs and their potential rSNPs

Single or composite motifs found in the PeakRegressor system may reflect actual transcription factor binding sites. If a single nucleotide polymorphism (SNP) occurs within the sites, regulatory control of nighboring gene transcription will be perturbed, thus leading to genetic diseases in some cases[1]. Therefore, true binding sites may have SNPs less frequently than the non-binding sites. As an important verification, we check the number of known SNPs to be found within the STAT1 positions presented by PeakRegressor by using dbSNP database[4]. We find that 0.39% (138 for 35,156 bp) of mapped positions with 7 GAS-like motifs in Table 2 on the whole genome contains SNPs, while as much as 0.54% (18,097 for 3,344,439 bp) of all positions contains SNPs on the whole genome sequences. The statistical difference between the above two ratios (0.39 % vs. 0.54 %) is highly significant such as $p < 7.8 \times 10^{-5}$ by Fisher's exact test. These sites are possible candidates of rSNPs because the slight change within the motif may affect the change of gene expression level and might cause diseases.

References

[1] A. Ameur, A. Rada-Iglesias, J. Komorowski, and C. Wadelius. Identification of candidate regulatory snps by combination of transcription-factor-binding site prediction, snp genotyping and haplochip. *Nucleic acids research*, 37(12):e85+, July 2009.

[2] Harmen J. Bussemaker, Hao Li, and Eric D. Siggia. Regulatory element detection using correlation with expression. In *RECOMB '01: Proceedings of the fifth annual international conference on Computational biology*, page 86, New York, NY, USA, 2001. ACM.

[4] http://www.ncbi.nlm.nih.gov/SNP/

[3] J. E. Butler and J. T. Kadonaga. The rna polymerase ii core promoter: a key component in the regulation of gene expression. *Genes Dev*, 16(20):2583–2592, October 2002.

[4] Erin M. Conlon, X. Shirley Liu, Jason D. Lieb, and Jun S. Liu. Integrating regulatory motif discovery and genome-wide expression analysis. *PNAS*, 2003.

[5] Debopriya Das, Matteo Pellegrini, and Joe W. Gray. A primer on regression methods for decoding cis-regulatory logic. *PLoS Comput Biol*, 5(1):e1000269, 01 2009.

[6] Barrett C. Foat, Alexandre V. Morozov, and Harmen J. Bussemaker. Statistical mechanical modeling of genome-wide transcription factor occupancy data by matrixreduce. *Bioinformatics*, 22(14):e141–e149, 2006.

[7] Feng Gao, Barrett C. Foat, and Harmen J. Bussemaker. Defining transcriptional networks through integrative modeling of mrna expression and transcription factor binding data. *BMC Bioinformatics*, 2004.

[8] Gordon Robertson, Martin Hirst, Matthew Bainbridge, Misha Bilenky, Yongjun Zhao, Thomas Zeng, Ghia Euskirchen, Bridget Bernier, Richard Varhol, Allen Delaney, Nina Thiessen, Obi L. Griffith, Ann He, Marco Marra, Michael Snyder, and Steven Jones. Genome-wide profiles of stat1 dna association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Meth*, 4(8):651–657, August 2007.

[9] Joel Rozowsky, Ghia Euskirchen, Raymond K. Auerbach, Zhengdong D. Zhang, Theodore Gibson, Robert Bjornson, Nicholas Carriero, Michael Snyder, and Mark B. Gerstein. Peakseq enables systematic scoring of chip-seq experiments relative to controls. *Nat Biotech*, 27(1):66–75, January 2009.

[10] Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.

Progress on Post-genome Technologies
——Proceedings of the 6'th International Forum of
Post-genome Technologies(IFPT'6)

DESIGNING PYRO-PRIMER SEQUENCES USING A SIMULATED ANNEALING ALGORITHM, TO CRITICALLY TARGET MRNAS IN QUANTITATIVE CELL ANALYSIS

Wataru Fujibuchi[1], Hirokazu Chiba[1], Hideo Akiyama[2], Hitoshi Shiku[3]

[1] Computaional Biology Research Center, Advanced Industrial Science and Technology (AIST), 2-42 Aomi, Koto-ku, Tokyo 135-0064, Japan
[2] Toray Industries, Inc., 10-1 Tebiro 6-chome Kamakura, Kanagawa 248-8555, Japan
[3] Graduate School of Environmental Studies, Tohoku University, 6-6-11-604 Aramaki Aoba, Sendai 980-8579, Japan

ABSTRACT

Exhaustive quantitation of mRNAs in a single cell by giga-sequencers is one of the key technologies used to measure the molecular states of cells in the Life Surveyor project in Japan. We have developed a computational method for designing multiplex primers for pyro-sequencers based on a combinatorial optimization algorithm called "simulated annealing." Here, we present a system to optimize a combined set of primers that binds to as many target genes as possible while minimizing cross-hybridization to off-target genes. The optimization of the primer combination for target genes is a challenging problem because of the large number of candidate combinations and the various factors that must be considered. Our system consists of three steps: (1) ultrafast evaluation of primer specificity with an accelerator; (2) T_m calculation by the nearest-neighbor method; and (3) optimization of the primer combination with a simulated annealing algorithm. For 1,000/1,000 target/off-target genes, our system can already successfully find primers that cover more than 85% of the targets and suppresses cross-hybridization to 10% of off-target genes under the most stringent conditions in just 1 day.

KEYWORDS Single-cell analysis, exhaustive quantitation, multiplex primers, pyro-sequencing, simulated annealing, FPGA, combinatorial optimization

INTRODUCTION

'Life Surveyor' is one of the leading projects that scrutinize single cells by the accurate quantitative and digital analysis of biomolecules, to achieve a comprehensive understanding of single cellular functions. With new technologies such as giga-sequencers[1], it has become more realistic to exhaustively measure mRNA quantities inside cells by directly reading their sequences. To achieve an exhaustive and accurate reading, the design of good primers is one of the most important issues in sequence analysis. However, optimizing the primer set is a challenging problem because of the large number of candidate primer combinations[1], together with various factors that must be considered in primer selection: (i) the minimum rate of cross-hybridization to off-target genes; (ii) primer selection by T_m; and (iii) the removal of primer pairs that form dimers when multiplex primers are used. Under these conditions, we must maximize the number of target genes recognized by a limited number of primers. Several methods have already been proposed to solve this problem[2~11]. However, these methods mainly focus on selecting the target genes and none of them minimizes the cross-hybridization rate while removing primer dimer formation among multiplex primers.

In this paper, we present a new method with which to design a multiplex primer set by solving the problems cited above using powerful informatics techniques. Our system consists of three steps: (1) ultrafast evaluation of primer specificity by an accelerator board called the "Field Programmable Gate Array" (FPGA)[12]; (2) T_m calculation by the nearest-neighbor method[13]; and (3) the optimization of primer combinations by a simulated annealing algorithm[14]. Our preliminary results indicate that our system can produce better candidate multiplex primer sets than can a simple greedy search. These sets cover 85% of target genes and cross-hybridize to only 10% of off-target genes, when 150 primers are used under stringent conditions, such as low T_m variances, with no

[1] To design 100 primers for 1,000 human cDNAs, there exist $\sim 10^6$ C $10^2 = \sim 10^{442}$ candidate combinations.

dimer formation.

ALGORITHMS AND METHODS

A schematic view of the entire system is shown in Figure 1.

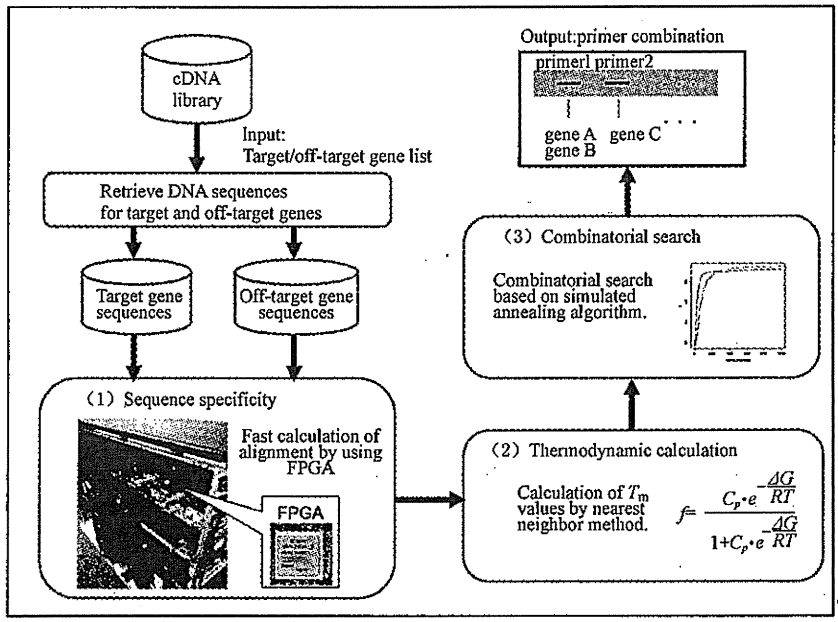


Figure 1 Schematic view of the entire system used to design multiplex primers by simulated annealing.

A list of target and off-target genes is first given by the user. Then candidate primers are extracted from the target genes and their hybridization specificity for target and off-target genes is checked with fast computation by FPGA. The primers are then screened according to their T_m values, and a set of multiplex primers, which bind to as many target genes as possible while minimizing cross hybridization to off-target genes, is finally optimized using a simulated annealing algorithm.

The candidate primers are usually taken from the 3 ends of cDNA sequences because anti-poly-A oligonucleotides (i. e. , oligo-Ts) are usually used to trap mature mRNAs and 3(sequences are more frequently amplified than 5 sequences. Our system currently uses ~1,000 candidates[2] in the 1,000 - nucleotide region upstream from the 3 end of each human cDNA.

Evaluation of Primer Specificity By FPGA Accelerator

As the input, a list of target and off-target genes is given by the user. Once ~1,000 candidate primers per target gene have been identified, the primer hybridization potential must be checked against other target and off-target genes by sequence alignment. There exist various sequence alignment algorithms, such as Smith-Waterman[15] and FASTA[16]. We used a simple gapless alignment between primers and cDNAs. However, this process requires a tremendous calculation time. For example, it takes more than 20,000 s (5. 6 h) to align each primer against all human cDNAs. Therefore, we used the FPGA board and developed a hardware circuit specialized for gapless alignments run in parallel. We coded the hardware logic in a C-like programming language called Verilog-HDL, with an implementing environment that automatically generates actual circuit maps and hardwires on the FPGA board.

T_m Calculation by the Nearest-Neighbor Method

Because the hybridization of DNA fundamentally obeys the laws of thermodynamics, it should be treated as a stochastic process of molecules in solution. There is a temperature, T_m, at which 50% of the DNA molecules of a target gene are probabilistically hybridized by a primer. In the hybridization reactions, it is important to select primers with similar T_m values. The calculation of T_m requires a prior calculation of the Gibbs' free energy change

[2] Mire precisely, 1,000−b+1 (where b is the length of the primer) candidates are extracted.

(ΔG_T) at standard temperature T for each primer based on the nearest-neighbor method[13]. Once ΔG_T is identified, the hybridization ratio f is estimated[17] by:

$$f=\frac{C_p \cdot e^{\frac{\Delta G_T}{RT}}}{1+C_p \cdot e^{\frac{\Delta G_T}{RT}}}$$

where Cp is the concentration of primers, R is the gas constant, and T is the absolute temperature. To determine the accurate $T_m(f=0.5)$, we used a binary search algorithm[18] to rapidly calculate the temperature as the average performance $O(\log n)$.

Optimization of Multiplex Primers By Simulated Annealing

To obtain an optimal set of multiplex primers that bind to as many target genes as possible while minimizing the cross-hybridization to off-target genes, the following three steps were followed:

(i) A matrix was constructed of the hybridization potentials of all candidate primers against target/off-target genes.

(ii) An initial set of multiplex primers was generated, whose number was specified by the user.

(iii) The new set of primers was rearranged with the simulated annealing algorithm to maximize the optimization score, which basically increases the number of bindings to target genes while reducing the number of bindings to off-target genes.

Multiplex primers, or a mixture of different primers, are added together if the reaction wells are not separated. This sometimes happens in sequencers and RT PCR experiments. If multiple different primers are added together, some primers may hybridize with each other and aggregate. To avoid this phenomenon, a conflict matrix must be constructed whose elements are either one or zero, indicating whether the primer pair hybridizes or not, respectively. Therefore, we included a further step:

(iv) Any conflict between the pairs of selected primers was checked and the new arrangement was discarded if an interaction was identified.

We repeated (iii) and (iv) until the optimization score reached a given threshold.

Step (iii) can be optimized by the standard simulated annealing algorithm with an optimization function. In this study, we used a difference-based function, such as

$$s=C_t-C_n,$$

where C_t and C_n are target coverage (%) and off-target coverage by cross-hybridization (%), respectively. We maximized s with the following steps: (1) one primer was selected as a replacement and the current score, s_t, was calculated at time t; (2) $s_{t+1,i}$ was calculated for each primer replacement candidate i at time $t+1$; (3) one primer was selected for which the probability was proportional to its score improvement Δs:

$$\Delta s=\min(1,\ \exp\{(s_{t+1,i}-s_t)/T_s\}),$$

where T_s is the temperature parameter that decreases to $T_s=T_s/1.05$ only if $\Delta s<1$. For comparison, we also used an alternative greedy approach, in which the best primer, with the highest $(s_{t+1,i}-s_t)$ difference, is always selected.

RESULTS AND DISCUSSION

Alignment by the FPGA Accelerator

We developed a gapless alignment code that works on FPGA in 32 parallel streams. The rate of alignment achieved by the FPGA accelerator is shown in Figure 2.

The calculation time for a gapless alignment by FPGA of 488 candidate primers to the human *ERBB* gene against a cDNA database was about 651 and 31 times faster than those of na ve and sophisticated algorithms (LAST[19]), respectively, on a general central processing unit (CPU).

The benchmark test using an actual human gene and a cDNA database showed that the total time for the calculation using na ve and sophisticated software (LAST[19]) on a general CPU, and its implementation with FPGA were

13 331.67, 637.43, and 20.32 s, respectively. Therefore, FPGA was 651 - and 31 - fold faster than the na ve and LAST methods, respectively. If we compare these methods based only on their alignment times, these numbers increase to 845 - fold faster and decrease to 3.7 - fold faster than the na ve and LAST methods, respectively. Therefore, the na ve method requires the shortest preprocessing time, whereas LAST requires a very long preprocessing time, although preprocessing is required only once before the program is run. FPGA requires a slightly longer preprocessing time than the na ve method but requires a much shorter alignment time than the other two methods. Importantly, the FPGA board used in this test costs ~ $US5,000, although the price will decrease and the calculation power increase in the future.

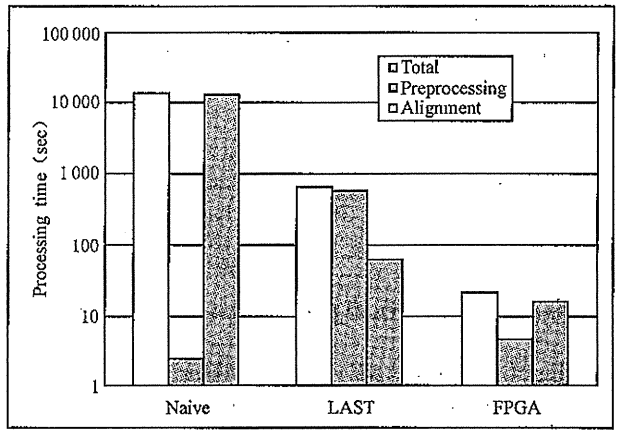
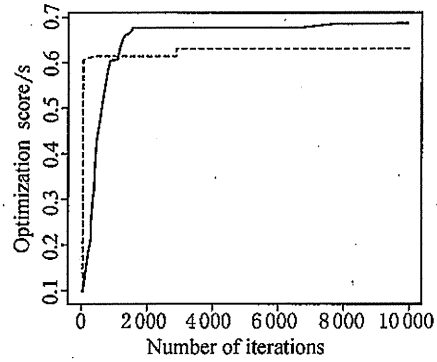


Figure 2 Increased rate of calculation of hybridization potential by FPGA.

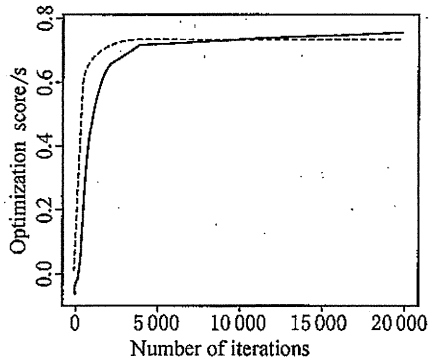
Searching for Global Optimal Solutions By Simulated Annealing

Simulated annealing (SA) is one of the most popular methods of searching a huge combinatorial space for global optimal solutions. To compare the capacity of SA with that of a greedy algorithm, we first created artificial data in which was embedded a global optimal solution for which the primer coverage (see ALGORITHMS and METHODS) was 70 of 100 targets and 5 of 400 off-targets. Figure 3(a) plots the results of searches by the SA and greedy algorithms. These results show that SA easily achieved the embedded solution (70, 5) with around 7,000 iterations. Surprisingly, SA found another unexpected and better solution (71, 5), whereas the greedy algorithm did not even find the embedded solution (66, 6) after 10,000 iterations.

After SA had been demonstrated to work better than the greedy algorithm on artificial data, we also tested it with more realistic data containing 1,000/1,000 target/housekeeping (off-target) gene sequences. All the sequence data were extracted from the Ensembl database[20]. As shown in Figure 3(b), the optimization scores, s, for the SA and greedy algorithms using 150 nine-nucleotide primers under the constraint 20℃ $\leqslant T_m \leqslant$ 30℃ after 20,000 iterations were s=0.750 (853, 103) and s=0.734(817, 83), respectively.



(a) Optimization with 100/400 target/off-target artificial sequences that contain primer coverage of (70, 5) for target/off-target genes as the solution. SA found this combination and an even better solution (71, 5) in 10,000 iterations, whereas the greedy algorithm identified only (66, 6).



(b) Optimization with real data containing 1,000/1,000 target/housekeeping (off-target) gene sequences. SA found a better solution (853, 103) than did the greedy approach.

Figure 3 Optimization of multiplex primers. The optimization scores achieved with the SA and greedy algorithms are plotted by solid and dashed lines, respectively.

Table 1 summarizes the optimization scores (s) for the SA and greedy algorithms under various conditions for the same 1,000/1,000 target/housekeeping gene dataset. In this experiment, SA achieved better optimization scores than did the greedy algorithm in all cases, with the same number of 20,000 iterations. Importantly, however, SA is a stochastic algorithm and generally takes a long time to converge on the global optimum, so the greedy algorithm sometimes gives better scores with limited iterations.

1,000/1,000 target/housekeeping (off-target) genes were the input sequences. The multiplex primers were optimized at three different T_m. Optimization with 100 (upper) and 150 (lower) primers. Note that the SA

algorithm found better solutions than did the simple greedy algorithm in all cases in this experiment.

Table 1 Summary of optimization scores under various conditions.

C=100primers	b=8nt, 20≤*Tm*≤30	b=9nt, 25≤*Tm*≤35	b=10nt, 30≤*Tm*≤40
greedy	0.573(777/204)	0.607(680/73)	0.565(589/24)
SA	0.601(798/197)	0.644(707/63)	0.567(597/30)
C=150primers	**b=8nt, 20≤*Tm*≤30**	**b=9nt, 25≤*Tm*≤35**	**b=10nt, 30≤*Tm*≤40**
greedy	0.545(826/281)	0.734(817/83)	0.676(714/38)
SA	0.596(875/279)	0.750(853/103)	0.677(704/27)

Experimental Validation

Although this is just a preliminary result and the primer conflict matrix (see ALGORITHMS and METHODS) was not considered because it was with the multiplex primers, the primers designed using our method were validated in wet experiments. We designed primers to three genes (*ERBB2*, *ABL1*, and *PR6SKB1*) that are important in human MCF-7 cells[21] and primers for three transcription factors (*Oct3/4*, *Dax1*, and *Rex1*) that are important in mouse embryonic stem (ES) cells[22], and tested their detection capacities using RT PCR. All the remaining cDNAs derived from the Ensembl database were used as the off-target genes. Although we did not show the gel electrophoresis data because the number of permissible figures is limited, the experimental data clearly showed DNA bands at exactly the expected nucleotide lengths for all six cDNAs.

CONCLUSION

In this study, we have established a fast calculation method using FPGA and a simulated annealing algorithm to optimize multiplex primers, which are important in next-generation sequencers. Currently, we are designing a large number of primers for genes expressed in mouse ES and early embryonic cells to better understand the mechanisms underlying the multipotency of stem cells. With the newly designed primers, we expect to establish associated pyrosequencing methods[23] for the exhaustive quantitation of whole mRNA species in single cells in the near future.

ACKNOWLEDGMENTS

We thank Dr Yutaka Akiyama, Dr Haruko Takeyama, Dr Yoshiko Okamura, and Dr Hideki Kambara for useful discussions and helpful comments. This work was partially supported by a Grant-in-Aid for Scientific Research on Priority Areas, number 19021049, from the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan.

REFERENCE

[1] Ronaghi, M., Genome Res 2001, 11, 3-11.
[2] Bekaert, M., Teeling, E. C., Nucleic Acids Res 2008, 36(10), e56.
[3] Chen, Y. F., Chen, R. C., Chan, Y. K., *et al.* Comput Biol Chem 2009, 33(2), 181-8.
[4] Fredslund, J., BMC Genomics 2008, 9, 140.
[5] Fu, Q., Ruegger, P., Bent, E., *et al.* J Microbiol Methods 2008, 72(3), 263-7.
[6] Li, K., Brownley, A., Stockwell, T. B., *et al.* BMC Bioinformatics 2008, 9, 191.
[7] Qu, W., Shen, Z., Zhao, D., *et al.* Bioinformatics 2009, 25(2), 276-8.
[8] Yao, J., Lin, H., Van Deynze, A., *et al.* BMC Microbiol 2008, 8, 185.
[9] You, F. M., Huo, N., Gu, Y. Q., *et al.* BMC Bioinformatics 2008, 9, 253.
[10] Yuryev, A., Methods Mol Biol 2007, 402, 93-104.
[11] Zhang, Q., Wu, G., Richards, E., *et al.* Virol J 2007, 4, 92.
[12] Yamaguchi, Y., Maruyama, T., Konagaya, A., Pac Symp Biocomput 2002, 271-82.
[13] SantaLucia Jr., J., Proc Natl Acad Sci USA 1998, 95, 1460-1465.
[14] Kirkpatrick, S., Gelatt Jr., C. D., Vecchi, M. P., Science 1983, 220, 671-680.
[15] Smith, T. F., Waterman, M. S., J Mol Biol 1981, 147, 195-197.
[16] Pearson, W. R., Lipman, D. J., Proc Natl Acad Sci USA 1988, 85, 2444-2448.
[17] Yamada, T., Soma, H., Morishita, S., Nucleic Acids Res 2006, 34, W665-W669.
[18] Cormen, T. H., Leiserson, C. E., Rivest, R. L., *et al.* Introduction to Algorithms, 2nd ed., 2001, The MIT Press, Cambridge, MA, U. S. A.
[19] http://last.cbrc.jp/
[20] Flicek, P., Aken, B. L., Beal, K., *et al.* Nucleic Acids Res 2008, 36, D707-D714.
[21] de Reynies, A., Geromin, D., Cayuela, J. M., *et al.* BMC Genomics 2006, 7, 51.
[22] Takahashi, K., Yamanaka S., Cell 2006, 126(4), 663-676.
[23] Taniguchi, K., Kajiyama, T., Kambara, H., Nature Methods 2009, 6, 503-506.

Down-regulation of cIAP2 enhances 5-FU sensitivity through the apoptotic pathway in human colon cancer cells

Hideaki Karasawa,[1] Koh Miura,[1,4] Wataru Fujibuchi,[2] Kazuyuki Ishida,[3] Naoyuki Kaneko,[1] Makoto Kinouchi,[1] Mitsunori Okabe,[1] Toshinori Ando,[1] Yukio Murata,[1] Hiroyuki Sasaki,[1] Kazuhiro Takami,[1] Akihiro Yamamura,[1] Chikashi Shibata[1] and Iwao Sasaki[1]

[1]Division of Biological-Regulation and Oncology, Department of Surgery, Tohoku University Graduate School of Medicine, 1-1 Seiryo-machi, Aoba-ku, Sendai, Miyagi, Japan 980-8574; [2]Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology, 2-42 Aomi, Koto-ku, Tokyo, Japan 135-0064; [3]Department of Pathology, Tohoku University Hospital, 1-1 Seiryo-machi, Aoba-ku, Sendai, Miyagi, Japan 980-8574

(Received August 31, 2008/Revised December 28, 2008/Accepted January 7, 2009/Online publication March 2, 2009)

Currently 5-fluorouracil (5-FU) plays a central role in the chemotherapeutic regimens for colorectal cancers and thus it is important to understand the mechanisms that determine 5-FU sensitivity. The expression profiles of human colon cancer cell line DLD-1, its 5-FU-resistant subclone DLD-1/FU and a futher 21 types of colon cancer cell lines were compared to identify the novel genes defining the sensitivity to 5-FU and to estimate which population of genes is responsible for 5-FU sensitivity. In the hierarchical clustering, DLD-1 and DLD-1/FU were most closely clustered despite over 100 times difference in their 50% inhibitory concentration of 5-FU. In DLD-1/FU, the population of genes differentially expressed compared to DLD-1 was limited to 3.3%, although it ranged from 4.8% to 24.0% in the other 21 cell lines, thus indicating that the difference of 5-FU sensitivity was defined by a limited number of genes. Next, the role of the *cellular inhibitor of apoptosis 2* (*cIAP2*) gene, which was up-regulated in DLD-1/FU, was investigated for 5-FU resistance using RNA interference. The down-regulation of cIAP2 efficiently enhanced 5-FU sensitivity, the activation of caspase 3/7 and apoptosis under exposure to 5-FU. The immunohistochemistry of cIAP2 in cancer and corresponding normal tissues from colorectal cancer patients in stage III revealed that cIAP2 was more frequently expressed in cancer tissues than in normal tissues, and cIAP2-positive patients had a trend toward early recurrence after fluorouracil-based chemotherapy. Although the association between drug sensitivity and the IAP family in colorectal cancer has not yet been discussed, cIAP2 may therefore play an important role as a target therapy in colorectal cancer. (*Cancer Sci* 2009; 100: 903–913)

5-fluorouracil (5-FU) is an anticancer drug that has been mainly used in the treatment of colorectal cancers. Recently, 5-FU has been combined with oxaliplatin or irinotecan as the first-line treatment for advanced colorectal cancers and these have significantly improved the response rates to 40–50% and prolonged overall survival.[1,2] Furthermore, novel biological agents including monoclonal antibodies such as cetuximab, which is an antibody against epidermal growth factor receptor (EGFR), and bevacizumab, which is an antibody against vascular endothelial growth factor, have been shown to provide additional clinical benefit for patients with metastatic colorectal cancers.[3–5] However, there are still a large number of patients who do not benefit from the present treatments because of anticancer drug resistance. Elucidating the mechanisms by which 5-FU resistance arises in colorectal cancer therefore remains an important issue for either overcoming or predicting such resistance.

5-FU is an analog of uracil and is rapidly incorporated into the cells using the same transport system as uracil.[6] Subsequently, 5-FU is converted into active metabolites which disrupt the action of thymidylate synthetase (TS) and RNA synthesis. TS and 5-FU-metabolizing enzymes such as dihydropyrimidine dehydrogenase (DPD) and thymidine phosphorylase (TP) have been analyzed to elucidate 5-FU resistance.[7] However the resistance to 5-FU has not been sufficiently explained by the metabolic pathway of 5-FU alone, because multiple factors participate in chemoresistance.[8] Recently, complementary DNA (cDNA) microarray technology has been used to identify novel genes regulating 5-FU resistance, and the potential biomarkers of 5-FU resistance other than pyrimidine metabolism-related enzymes have been proposed.[9,10]

Apoptosis is found to be one of the primary mechanisms of the cytotoxic effect of chemotherapeutic agents and inhibition of the apoptotic pathway is one of the factors that may be responsible for drug resistance.[11–13] In the process of apoptosis, the caspase cascade plays a central role,[14,15] and the inhibitor of apoptosis protein (IAP) family is thought to prevent apoptosis through direct caspase and pro-caspase inhibition (primarily caspase 3 and 7). The IAPs have been described to be abnormally regulated in various types of cancers,[16,17] and recently they have been regarded as therapeutic targets of cancer.[18,19] Although the association between IAPs and drug resistance has been discussed in cancers of some organs such as lung, pancreas and kidney,[20–22] it has not been fully analyzed in human colorectal cancer.

The present study compared the messenger RNA (mRNA) expression profiles between the human colon cancer cell line DLD-1 and its 5-FU-resistant subclone DLD-1/FU by cDNA microarray to investigate the novel genes regulating 5-FU resistance. To estimate which population of genes are responsible for regulating 5-FU sensitivity or resistance, the expression profiles of DLD-1 and DLD-1/FU were also compared to another 21 types of colon cancer cell lines. Next, the role of the cellular IAP 2 (cIAP2) gene, which is most highly expressed among genes of the IAP family in DLD-1/FU, was investigated using RNA interference (RNAi) on the sensitivity to 5-FU, the activation of caspase 3/7, and apoptosis in human colon cancer cells. Finally, to identify the association between cIAP2 expression and 5-FU resistance in human primary colorectal cancer, immunohistochemistry for

[4]To whom correspondence should be addressed.
E-mail: k-miura@surg1.med.tohoku.ac.jp

Abbreviations: 5-FU, 5-fluorouracil; TS, thymidylate synthetase; DPD, dihydropyrimidine dehydrogenase; TP, thymidine phosphorylase; IAP, inhibitor of apoptosis protein; RNAi, RNA interference; FBS, fetal bovine serum; MTS, 3-(4,5-dimethylthiazol-2-yl)-5-(3-carboxymethoxyphenyl)-2-(4-sulfenyl)-2H-tetrazolium, inner salt; IC_{50}, inhibitory concentration 50%; SD, standard deviation; RT-PCR, reverse transcription polymerase chain reaction; GAPDH, glyceraldehyde-3-phosphate dehydrogenase; siRNA, small interfering RNA; DW, distilled water; PI, propidium iodide; FITC, fluorescein isothiocyanate; CTNNA1, catenin alpha 1; HBEGF, heparin-binding EGF-like growth factor; PLA2G2A, phospholipase A2 group IIA; OPRT, orotate phosphoribosyl transferase; EGFR, epidermal growth factor receptor; NCBI, National Center for Biotechnology Information.

doi: 10.1111/j.1349-7006.2009.01112.x