For STAT1, we use 200-bp windows around the peak centers to define the peak sequences. For RNA Polymerase 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 RNA Polymerase II peak sequence lengths lie between 200 and 4000 bp.

Evaluation of prediction performance

PeakRegressor predicts the peak scores and therefore, we have two different values for each peak. The "true" peak score is the score provided by [9], and is derived from the frequency of reads of ChIP-Seq data. The predicted score is computed by PeakRegressor using the peak sequence information. Ideally, the predicted score should be equal to the true score. We use correlation coefficients to evaluate the prediction quality of PeakRegressor.

Experimental protocol

For L1-norm log linear regression and ridge regression, 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-fold 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 results of L1-norm log linear regression and ridge regression are averaged over the 30-fold cross-validation.

For partial least squares regression and principal component regression, the experiments were limited by the slowness of both methods. First we have to set the number of components K used for regression. We tried $K = 1...10$, and performed a 30-fold cross-validation for each value of K. In each fold, we split the dataset into 50% for training and 50% for testing. All the results of partial least squares regression and principal component regression are averaged over the 30-fold cross-validation.

Acknowledgments

The authors thank the anonymous CAMDA reviewers for their helpful comments.

Author Contributions

Conceived and designed the experiments: JFP WF. Performed the experiments: JFP HH TT. Analyzed the data: JFP HC WF. Wrote the paper: JFP HC WF.

References

1. Bussemaker HJ, Li H, Siggia ED (2001) Regulatory element detection using correlation with expression. In: RECOMB '01: Proceedings of the fifth annual international conference on Computational biology. New York, NY, USA: ACM. 86 p. doi:http://doi.acm.org/10.1145/369133.369174.
2. Conlon EM, Liu XS, Lieb JD, Liu JS (2003) Integrating regulatory motif discovery and genome-wide expression analysis. PNAS.
3. Das D, Pellegrini M, Gray JW (2009) A primer on regression methods for decoding cis-regulatory logic. PLoS Comput Biol 5: e1000269.
4. Foat BC, Morozov AV, Bussemaker HJ (2006) Statistical mechanical modeling of genome-wide transcription factor occupancy data by matrixreduce. Bioinformatics 22: e141–e149.
5. Gao F, Foat BC, Bussemaker HJ (2004) Defining transcriptional networks through integrative modeling of mrna expression and transcription factor binding data. BMC Bioinformatics.
6. Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, et al. (2007) Genome-wide profiles of stat1 dna association using chromatin immunoprecipitation and massively parallel sequencing. Nat Meth 4: 651–657.
7. Butler JE, Kadonaga JT (2002) The rna polymerase ii core promoter: a key component in the regulation of gene expression. Genes Dev 16: 2583–2592.
8. Efron B, Hastie T, Johnstone I, Tibshirani R (2004) Least angle regression. The Annals of Statistics.
9. Rozowsky J, Euskirchen G, Auerbach RK, Zhang ZD, Gibson T, et al. (2009) Peakseq enables systematic scoring of chip-seq experiments relative to controls. Nat Biotech 27: 66–75.
10. Ameur A, Rada-Iglesias A, Komorowski J, Wadelius C (2009) Identification of candidate regulatory snps by combination of transcription-factor-binding site prediction, snp genotyping and haplochip. Nucleic acids research 37.
11. Tibshirani R (1996) Regression shrinkage and selection via the lasso. J Roy Statist Soc Ser B 58: 267–288.
12. Bishop CM (2006) Pattern Recognition and Machine Learning (Information Science and Statistics). Secaucus, NJ, USA: Springer-Verlag New York, Inc.
13. Frank IE, Friedman JH (1993) A statistical view of some chemometrics regression tools. Technometrics.

Cancers **2010**, *2*, 1717-1730; doi:10.3390/cancers2031717

OPEN ACCESS

cancers

ISSN 2072-6694

www.mdpi.com/journal/cancers

Review

5-FU Metabolism in Cancer and Orally-Administrable 5-FU Drugs

Koh Miura [1,*], Makoto Kinouchi [1], Kazuyuki Ishida [2], Wataru Fujibuchi [3], Takeshi Naitoh [1], Hitoshi Ogawa [1], Toshinori Ando [1], Nobuki Yazaki [1], Kazuhiro Watanabe [1], Sho Haneda [1], Chikashi Shibata [1] and Iwao Sasaki [1]

[1] Department of Surgery, Tohoku University Graduate School of Medicine, Sendai, Japan; E-Mails: kinouchi@surg1.med.tohoku.ac.jp (M.K.); hogawa@surg1.med.tohoku.ac.jp (H.O.); ando@surg1.med.tohoku.ac.jp (T.A.); n_yazaki@surg1.med.tohoku.ac.jp (N.Y.); k-wata@surg1.med.tohoku.ac.jp (K.W.); sho@surg1.med.tohoku.ac.jp (S.H.); cshibata@surg1.med.tohoku.ac.jp (C.S.); isasaki@surg1.med.tohoku.ac.jp (I.S.)

[2] Department of Pathology, Tohoku University Hospital, Sendai, Japan; E-Mail: musubi@patholo2.med.tohoku.ac.jp

[3] Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology, Tokyo, Japan; E-Mail: w.fujibuchi@aist.go.jp

* Author to whom correspondence should be addressed; E-Mail: k-miura@surg1.med.tohoku.ac.jp; Tel.: +81-22-717-7205; Fax: +81-22-717-7209.

Received: 23 August 2010; in revised form: 14 September 2010 / Accepted: 15 September 2010 / Published: 17 September 2010

Abstract: 5-Fluorouracil (5-FU) is a key anticancer drug that for its broad antitumor activity, as well as for its synergism with other anticancer drugs, has been used to treat various types of malignancies. In chemotherapeutic regimens, 5-FU has been combined with oxaliplatin, irinotecan and other drugs as a continuous intravenous infusion. Recent clinical chemotherapy studies have shown that several of the regimens with oral 5-FU drugs are not inferior compared to those involving continuous 5-FU infusion chemotherapy, and it is probable that in some regimens continuous 5-FU infusion can be replaced by oral 5-FU drugs. Historically, both the pharmaceutical industry and academia in Japan have been involved in the development of oral 5-FU drugs, and this review will focus on the current knowledge of 5-FU anabolism and catabolism, and the available information about the various orally-administrable 5-FU drugs, including UFT, S-1 and capecitabine. Clinical studies comparing the efficacy and adverse events of S-1 and capecitabine have been

reported, and the accumulated results should be utilized to optimize the treatment of cancer patients. On the other hand, it is essential to elucidate the pharmacokinetic mechanism of each of the newly-developed drugs, to correctly select the drugs for each patient in the clinical setting, and to further develop optimized drug derivatives.

Keywords: 5-FU metabolism; cell death; colon cancer; oral 5-FU drugs

1. Introduction

Since its introduction more than 50 years ago, 5-fluorouracil (5-FU) has become a key anticancer drug that has been used to treat various types of malignancies for its broad antitumor activity, as well as its synergism with other anticancer drugs. In 1957, Heidelberger *et al.* [1] reported the development of 5-FU, but several important findings had preceded their work. For example, in 1954 Rutman *et al.* [2] showed that uracil was incorporated into rat hepatomas more rapidly than normal tissues; and in 1956 Handschumacher *et al.* reported the tumor-inhibitory activity by 6-azauracil [3]. In recent chemotherapeutic regimens, the continuous intravenous infusion of 5-FU has been combined with oxaliplatin, irinotecan and other drugs. The continuous 5-FU infusion is based on an official report published in the US in 1964 [4], showing that 5-FU is a time-dependent antimetabolite. The meta-analysis of more than 1,200 colorectal cancer patients in six randomized clinical trials, which showed the efficacy of continuous 5-FU infusion compared with bolus 5-FU administration [5], also supported the importance of continuous 5-FU infusion. Based on these results, continuous 5-FU infusion regimens, such as FOLFOX or FOLFIRI, have been established and are widely utilized. On the other hand, recent clinical studies have shown that several of the chemotherapeutic regimens with oral 5-FU drugs are not inferior to those with continuous 5-FU infusion chemotherapy, and in some regimens it may be possible to replace continuous 5-FU infusion chemotherapies with oral 5-FU drugs. Historically, both the pharmaceutical industry and academia in Japan have contributed to the development of oral 5-FU drugs. This review will summarize the current knowledge about 5-FU metabolism, and the information about orally-administrable 5-FU drugs.

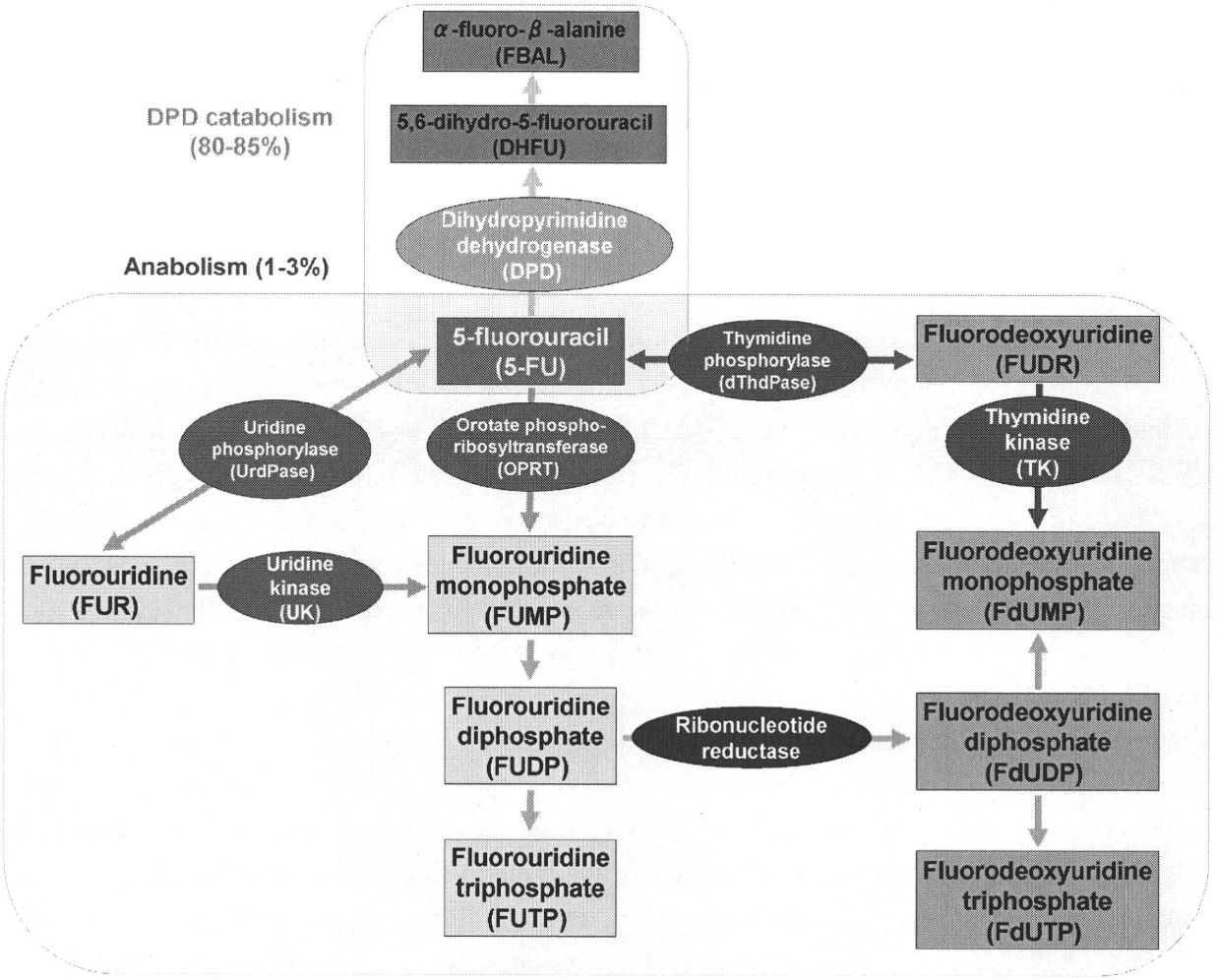
2. 5-FU Metabolism

It has been demonstrated that 80% to 85% of 5-FU is catabolized to inactive metabolites by dihydropyrimidine dehydrogenase (DPD), and only 1 to 3% of the original dose of 5-FU mediates the cytotoxic effects on tumor cells and normal tissues through anabolic actions [6], thereby inhibiting DNA synthesis and RNA processing and function (Figure 1). The 5-FU metabolite, fluorodeoxyuridine monophosphate (FdUMP), forms a ternary complex with thymidylate synthase (TS) and 5,10-methylene tetrahydrofolate (CH2THF), thereby inhibiting the synthesis of DNA.

2.1. 5-FU Anabolism

The chemotherapeutic compound 5-FU is a uracil analogue with a fluorine atom at the C-5 position. After intravenous administration of 5-FU, it rapidly enters cells using the same transport mechanism as uracil [7]. The processing mechanism of 5-FU in cells is as diverse as that of normal pyrimidines, and the current understanding of the metabolism is summarized in Figure 1. First, 5-FU is converted to the following active metabolites: 1) fluorouridine triphosphate (FUTP), which is incorporated into RNA instead of uridine triphosphate (UTP); 2) fluorodeoxyuridine triphosphate (FdUTP), which is incorporated into DNA instead of deoxythymidine triphosphate (dTTP); and 3) FdUMP, which inhibits the activity of TS in the ternary complex, as described in the previous section. FUTP causes alterations in RNA processing and function, and FdUTP and FdUMP cause DNA damage; both of these processes affect RNA and DNA and cause cell death.

Figure 1. 5-FU anabolism and catabolism.



As mentioned, a US report published in 1964 demonstrated 5-FU to be a time-dependent antimetabolite [4]. The main mechanism of 5-FU activation is conversion to fluorouridine monophosphate (FUMP), either directly by orotate phosphoribosyltransferase (OPRT) with phosphoribosyl pyrophosphate as a cofactor, or indirectly via fluorouridine (FUR) through the

sequential action of uridine phosphorylase (UrdPase) and uridine kinase (UK) [8]. The other 5-FU activation pathway involves thymidine phosphorylase (dThdPase), which catalyzes the conversion of 5-FU to fluorodeoxyuridine (FUDR), and FUDR is then phosphorylated by thymidine kinase (TK) to FdUMP. In this series of reactions, the phosphorylation reaction by the UrdPase requires ribose-1-phosphate as a cofactor, eventually synthesizing FUMP. In contrast, the phosphorylation reaction by dThdPase requires deoxyribose-1-phosphate as a cofactor, eventually leading to the synthesis of FdUMP. FUMP is further phosphorylated to fluorouridine diphosphate (FUDP), which is either further phosphorylated to the active metabolite FUTP, or converted to fluorodeoxyuridine diphosphate (FdUDP) by ribonucleotide reductase [8]. FdUDP is then either further phosphorylated to FdUTP, or dephosphorylated to FdUMP. Both FdUTP and FdUMP cause DNA damage.

The conversion of 5-FU to FdUMP in the gastrointestinal (GI) tract and bone marrow elicits GI toxicity and myelotoxicity, respectively. In 1979, an *in vivo* mouse study by Houghton *et al.* indicated that GI toxicity was caused by the incorporation of fluorinated pyrimidines, mainly FdUMP [9]. In 1984, Schuetz *et al.* analyzed the myelotoxicity of 5-FU using CF-1 mouse bone marrow cells under 5-FU exposure *in vitro* [10], and demonstrated that 5-FU incorporation into DNA was closely associated with toxicity and inhibition of DNA synthesis with FdUMP [10]. Interestingly, the meta-analysis of six randomized clinical trials performed in 1998 showed that the grade 3 or 4 hematologic toxicity was more frequent in patients assigned to bolus 5-FU infusion rather than in those assigned to continuous 5-FU infusion [11].

2.2. 5-FU Catabolism

DPD is an enzyme present in the liver, intestinal mucosa and various other tissues. DPD catabolizes 5-FU to 5,6-dihydro-5-fluorouracil (DHFU) [12], finally leading to the formation of α-fluoro-β-ureido-propionic acid and α-fluoro-β-alanine (FBAL) (Figure 1). In 1987, Heggie *et al.* investigated the kinetics of 5-FU and 5-FU metabolites in cancer patients following intravenous bolus administration of radio-labeled 5-FU [13], and revealed that approximately 60–90% of the administered 5-FU was excreted in urine as FBAL within 24 hours. While most patients tolerate 5-FU reasonably well, a number of cancer patients with DPD deficiency were shown to be at increased risk for severe toxicities, including diarrhea, mucositis, and neurotoxicity, as well as death, after administration of standard doses of 5-FU [6].

Since the 1970s, the neurotoxicity of FBAL as a 5-FU catabolite has been discussed quite extensively [14,15]. Okeda *et al.* investigated the mechanism of 5-FU neurotoxicity with *in vivo* experiments using cats [15]. The two 5-FU metabolites, monofluoroacetic acid and FBAL, were continuously administered into the left ventricle of the brain in cats. In their experiments, two types of neuropathological changes, vacuoles and necrosis/softening-like changes, were detected, and FBAL was more toxic than monofluoroacetic acid. Both of the neuropathological changes in the FBAL group were similar to those found in patients following orally-administered 5-FU [15].

The cardiotoxicity of 5-FU has also been attributed to FBAL. Matsubara *et al.* investigated the mechanism of cardiotoxicity for 5-FU and its derivatives using *in vivo* experiments with anesthetized open-chest guinea pigs [16], and proposed that the formation of fluoroacetate, an inhibitor of aconitase, from 5-FU via FBAL, caused cardiotoxicity during chemotherapy [16]. As described in later

publications, FBAL is also the main cause of hand-foot syndrome (HFS) acquired in cancer patients during 5-FU-based chemotherapy. In the 1998 meta-analysis HFS was more frequent in the continuous 5-FU infusion group than in the bolus 5-FU infusion group [5].

2.3. Ternary Complex

FdUMP forms a stable ternary complex with TS and CH2THF [17]. TS catalyzes the reductive methylation of deoxyuridine monophosphate (dUMP) to deoxythymidine monophosphate (dTMP) with the reduced folate CH2THF. The ternary complex blocks the access of dUMP to the nucleotide-binding site of TS by competition with FdUMP, which results in pool imbalances of deoxynucleotides, especially an increased level of deoxyuridine triphosphate (dUTP); leading to DNA damage. Depletion of dTMP results in the subsequent depletion of dTTP, which perturbs the levels of the other deoxynucleotides [18]. The pool imbalances of deoxynucleotides severely disrupt DNA synthesis and repair, again resulting in DNA damage [19]. As a result, the inhibition of TS results in the accumulation of dUMP, which leads to the increased levels of dUTP [20]. Thymidylate can be salvaged from thymidine through the action of TK, and this salvage pathway can also represent a mechanism of resistance to 5-FU [21]. Despite this information about the activity of 5-FU, the molecular mechanisms downstream of TS inhibition still have to be confirmed [8]. In addition, the clinical significance of TS needs to be demonstrated. In 2008, Showalter *et al.* investigated the connection between TS expression and 5-FU with a thorough literature survey, and in contrast to previous predictions, they found no connection between TS and the patient response to 5-FU [22]. To discuss this matter, we must remember that the influence of TS activity on 5-FU metabolism may change depending on the administration routes of 5-FU drugs, types of 5-FU drugs, the effects of LV, and other factors.

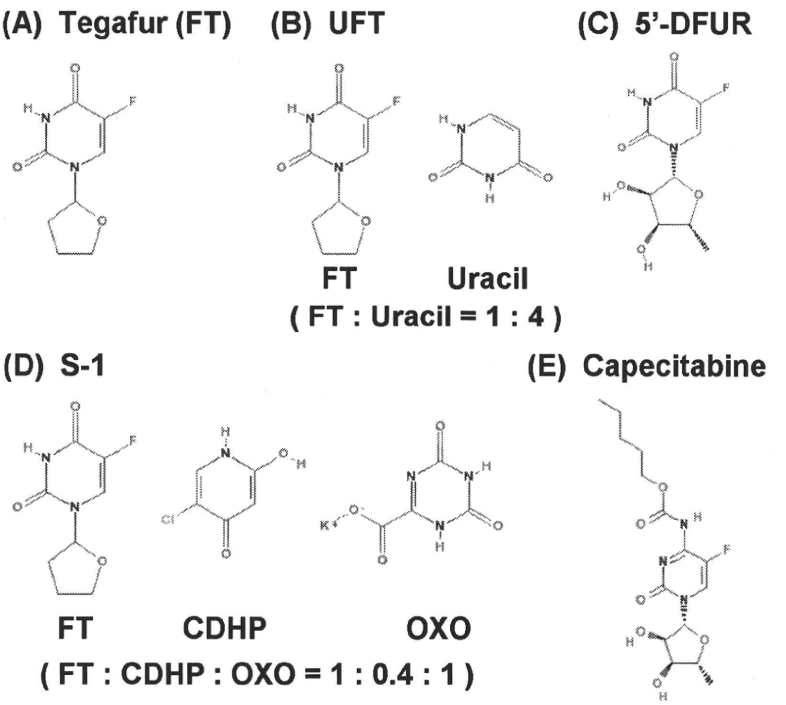
3. Oral 5-FU Drugs

As described in the "Introduction" section, 5-FU is a key anticancer drug for the treatment of various malignancies, and continuous 5-FU infusion regimens have been frequently used because of the apparent time-dependent effects of the drug. However, recent studies have shown that the continuous 5-FU infusion chemotherapies can be replaced with orally-administrable 5-FU drugs in some regimens, without any significant changes in either efficacy or adverse events [23,24]. In addition, oral administration of drugs allows several types of iatrogenic issues to be avoided. For the continuous infusion regimens such as FOLFOX or FOLFIRI, the implantation of a central venous port is required, but complications such as pneumothorax, hemothorax, or disconnection of the devices can occur. Furthermore, catheter-related infection or thrombosis is a serious problem for cancer patients [25,26]. The cost and benefit balance with the use of the central venous port system has been discussed [27], and recent clinical studies revealed that patients prefer oral administration rather than continuous infusion procedures. As such, orally-administered 5-FU regimens are likely to become more common in the clinical setting. Some fluoropyrimidines such as BOF-A2 (Emitefur) and Galocitabine (Ro 09-1390) are under development but not clinically available. In this section, we summarize the information currently available about orally-administrable 5-FU drugs (Table 1 and Figure 2).

Table 1. Orally-administrable 5-FU drugs.

Drug name	Structure (Composition)	Concept	Developer	Refs.
Tegafur	1-(2-Tetrahydrofuryl)-5-fluorouracil	Prodrug	National Institute for Organic Syntheses (Latvia)	[28]
UFT	FT:Uracil = 1:4	Prodrug, DPD inhibitor	Osaka University (Japan)	[30]
5'-DFUR	5'-Deoxy-5-fluorouridine	Prodrug	Hoffmann-La Roche (Switzerland); Nippon Roche Research Center (Japan)	[38,39]
S-1	FT:CDHP:OXO = 1:0.4:1	DPD inhibitor, OPRT inhibitor	Taiho Pharmaceuticals (Japan)	[40]
Capecitabine	N4-Pentyloxycarbonyl-5'-deoxy-5-fluorocytidine	Prodrug	Nippon Roche Research Center (Japan)	[44]

Figure 2. Structures of oral 5-FU drugs. (**A**) Tegafur; (**B**) UFT; (**C**) 5'-DFUR; (**D**) S-1; (**E**) Capecitabine.



3.1. Tegafur

1-(2-Tetrahydrofuryl)-5-fluorouracil (tegafur, FT, FT-207, Futrafur, Ftorafur, *etc.*) was developed as a 5-FU prodrug in the Soviet Union during the Cold War (as reported in 1967 by Giller *et al.* in a Russian record [28]). In 1970, the drug was introduced to Taiho Pharmaceuticals (Japan). Utilizing the benefits of FT, including: 1) its excellent absorbability from the GI tract and 2) its slight conversion to 5-FU in the GI tract, the development of orally-administrable FT was attempted, accomplished and reported in 1977 [29,30]. FT was shown to be gradually converted to 5-FU via cytochrome p450 enzymes in hepatic microsomes [31].

3.2. UFT

UFT consists of uracil and FT. Uracil competes with 5-FU for DPD activity [32,33], resulting in a prolonged 5-FU half-life. To optimize the molecular ratio of FT and uracil, Fujii *et al.*, at the Institute for Protein Research (Osaka University, Japan), analyzed *in vivo* rat models administered with the combination of drugs, and revealed the optimal molar ratio to be 1:4 [34], which led to the introduction of UFT in 1985. In 1978, Fujii *et al.* also reported that the antitumor activity of FT on sarcoma-180 and AH-130 tumors was enhanced by oral administration of uracil, deoxyuridine or uridine [30], and this enhancement of the antitumor activity of FT increased with uracil, which caused a more extensive enhancement than did deoxyuridine or uridine. Furthermore, biochemical modulation of 5-FU had been investigated [35] using methotrexate, trimetrexate, interferon-α, leucovorin (LV) [36], and *N*-(phosphonacetyl)-L-aspartic acid. The addition of LV to UFT regimens increases the available reduced folates, and thereby stabilizes the binding of FdUMP to TS, eventually inhibiting DNA synthesis. In 1997, Rustum *et al.* showed that LV increased the antitumor activity of UFT in the rat [32]; and Ichikura *et al.* showed that UFT with LV enhanced the inhibition of TS activity in gastric cancer patients [37]. In fact, the combination of 5-FU-based drugs with LV has been regarded as one of the standard treatments for colorectal cancer. These results eventually led to the development of S-1.

3.3. 5'-DFUR

In 1979, Cook *et al.* at Hoffmann-La Roche (Switzerland) [38] and Ishitsuka *et al.* in 1980 at the Nippon Roche Research Center (Japan) [39] reported the development of 5'-deoxy-5-fluorouridine (5'-DFUR, doxyfluridine, 5'-fluoro-5'-deoxyuridine, Ro 21-9738, Furtulon, *etc.*). The compound 5'-DFUR is parenterally and orally effective, and its activity was better than that of other fluorinated pyrimidines available at that time. A subline of L1210 leukemia cells was resistant to 5'-DFUR, and Ishitsuka *et al.* revealed that its resistance to 5'-DFUR was due to the lack of the UrdPase [39]. This is because 5'-DFUR is considered to be a depot form of 5-FU, which can be promptly activated by UrdPase [39]. Capecitabine (see below) was developed as the next generation of 5'-DFUR.

3.4. S-1

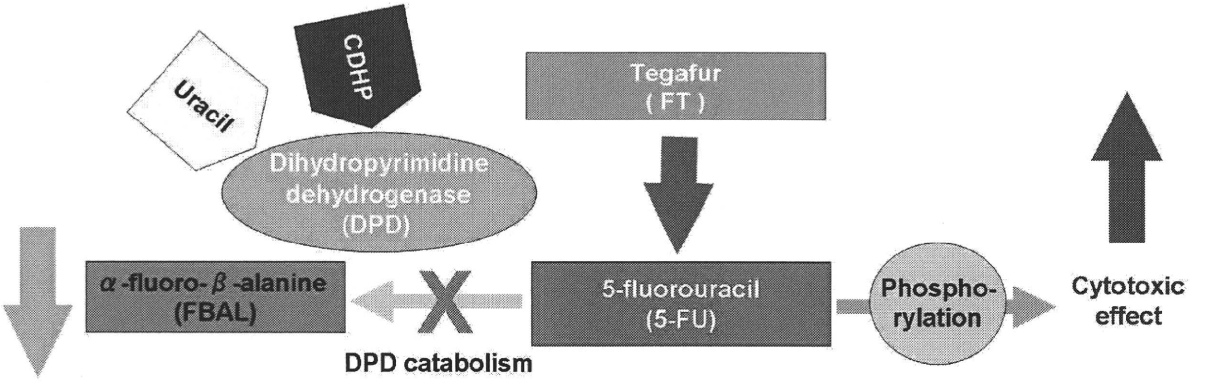
After the development of UFT, Shirasaka *et al.* focused on the development of a novel oral FT-based fluoropyrimidine agent. They developed the next-generation drug, S-1, which both enhances the anticancer activity of 5-FU and reduces its GI toxicity [40]. The development of S-1 was based on two important findings: 1) 5-chloro-2,4-dihydroxypyridine (CDHP, Gimeracil, gimestat, *etc.*) is a DPD inhibitor, and 2) potassium oxonate (OXO) is an OPRT inhibitor (Figure 3).

Tatsumi *et al.* at Otsuka and Taiho Pharmaceuticals (Japan) investigated about 30 compounds for their inhibitory effects of DPD, mainly focusing on pyrimidines, barbituric acid and pyridine derivatives [41]; and in 1987 they reported that 3-cyano-2,6-dihydroxypyrimidine (CNDP) and CDHP were the strongest inhibitors of DPD [41]. Next, Shirasaka *et al.* [42] investigated the possibility of decreasing the GI toxicity of 5-FU without reducing its antitumor activity in rats. OXO localizes in the GI mucosa and selectively inhibits the OPRT, which inhibits 5-FU phosphorylation to FUMP, limiting GI toxicity effects (diarrhea, nausea and vomiting) [42]. In 1993, they reported that OXO inhibited the

phosphorylation of 5-FU to FUMP catalyzed by pyrimidine phosphoribosyl-transferase, in a different manner from allopurinol. With experiments using Yoshida sarcoma-bearing rats, OXO was found to inhibit the formation of FUMP from 5-FU, with its subsequent incorporation into the RNA fractions of the small and large intestine, but not of the tumor and bone marrow tissues. This selective inhibition of 5-FU phosphorylation in the GI tract was due to the much higher concentrations of OXO in GI tissues than in other tissues and in the blood [42].

Based on these findings, CDHP and FT were simultaneously given orally to Yoshida sarcoma-bearing rats in various molar ratios, and then OXO was given orally during consecutive administration of the FT-CDHP mixture to find out the best condition to protect the animals from body weight loss without affecting the high antitumor efficacy of the FT-CDHP mixture [40]. Shirasaka *et al.* finally proposed a suitable formulation of the FT-based anticancer drug, called S-1, consisting of FT, CDHP and OXO at a 1:0.4:1 molar ratio and showed that it had tumor-selective cytotoxicity. S-1 is designed to reduce the GI toxicity of 5-FU; and in 2005 Muneoka *et al.* also reported that S-1 may be administered safely to patients with 5-FU-induced cardiotoxicity in whom FBAL is related to adverse events [43]. Recently, a combination granule version of S-1 has become commercially available.

Figure 3. The metabolism of S-1.



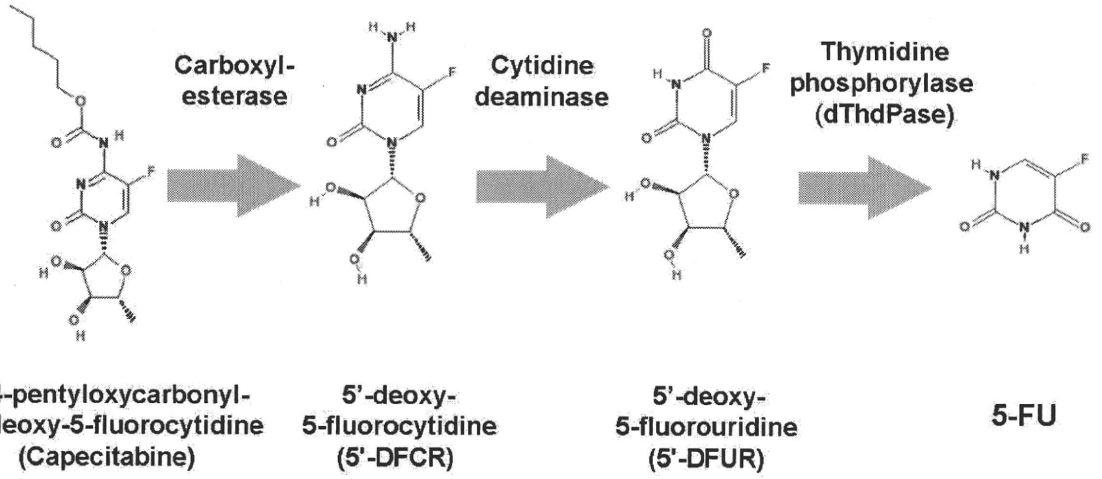
3.5. Capecitabine

Capecitabine (N4-pentyloxycarbonyl-5′-deoxy-5-fluorocytidine, Xeloda™, Ro 09-1978, *etc.*) is an oral fluoropyrimidine carbamate [44], which is selectively converted to 5-FU in tumors through a cascade of three enzymes: (1) carboxylesterase, which is almost exclusively located in the liver and hepatoma, but not in other tumors and normal tissues; (2) cytidine deaminase, which is located in the liver and various types of solid tumors, and 3) dThdPase, which is more concentrated in various types of tumor tissues than in normal tissues (Figure 4).

Miwa *et al.* investigated the tissue localization of the three enzymes in humans [44], and these unique tissue localization patterns enabled the design of capecitabine. Oral capecitabine passes intact through the intestinal tract, but is converted first by carboxylesterase to 5'-deoxy-5-fluorocytidine (5'-DFCR) in the liver, then by cytidine deaminase to 5'-DFUR in the liver and tumor tissues, and finally by dThdPase to 5-FU in tumors. To design the optimized fluoropyrimidine carbamate, a series of N4-alkoxylcarbonyl derivatives were screened for hydrolysis to 5'-DFCR, specifically by carboxylesterase [45]. During the screening process, derivatives having an N4-alkoxylcarbonyl moiety

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.

In Search of True Reads: A Classification Approach to Next Generation Sequencing Data Selection

Edward Wijaya*†, Jean-François Pessiot*, Martin C. Frith*, Wataru Fujibuchi*, Kiyoshi Asai*†, and Paul Horton*‡
*AIST, Computational Biology Research Center, 2-42 Aomi, Koutou-Ku, Tokyo 1350064
†Department of Computational Biology, Graduate School of Frontier Science,
the University of Tokyo, 5-1-5, Kashiwanoha, Kashiwa, Chiba 2778561
‡Email: horton-p@aist.go.jp

***Abstract*—Next generation sequencing (NGS) technology has increasingly become the backbone of transcriptomics analysis, but sequencer error causes biases in the read counts. In this paper we establish a framework for predicting true sequences from NGS data. We formulate this task as a classification problem. We define several features, such as log likelihood ratio of estimated true counts, error probability and observed count of the reads. Using a Support Vector Machine (SVM) classifier, we show that on simulated reads these features can achieve 96.35% classification accuracy in discriminating true sequences. Using this framework we provide a way for users to select sequences with a desired precision and recall for their analysis. The feature generation software and the simulated data set can be obtained from (`http://seq.cbrc.jp/NGSFeatGen`).**

***Keywords*-next generation sequencing; transcriptomics; Illumina; Solexa; expectation maximization; classification**

I. Introduction

Recent advances in DNA sequencing technologies enables transcripts to be measured with unprecedented accuracy and resolution. However these technologies also have a sequencing error rate that creates biases by yielding false sequences [6], [11], and therefore can significantly reduce the quality of the conclusions which can be drawn from the data.

The first batch of sequence count correction methods [1]–[4] were designed for SAGE data [16]. However the software created at that time is not able to handle the large datasets generated by next generation sequencers.

Recently several methods have been developed for correcting errors in DNA reads. One class of methods is aimed at genome sequencing: they assume that the genome is sequenced with high coverage, so that correct sequences tend to be present many times in the data, whereas incorrect sequences tend to be present fewer times [12], [14]. So these methods identify erroneous sequences by their rareness. This approach is not suitable for transcriptome or metagenome sequencing, where correct sequences are often rare. For these sequencing applications, more sophisticated error-correction methods, which consider the entire ensemble of reads when correcting are needed.

At least two such tools have been made for sequence count error correction in next generation sequencing data. The first work, FreClu, by Qu *et. al* [11] involves an iterative procedure to cluster reads and performs the sequencing error test for each cluster to assess the reads membership to the cluster. Finally the estimated true counts are computed from the total frequency of reads inside the cluster. The second is our tool RECOUNT [18]. RECOUNT adopts the method proposed by Beißbarth et al. [2] which models the sequencer error as a multinomial thinning process and applies the Expectation-Maximization (EM) framework [5] to infer the set of true counts which (locally) maximize the likelihood of the observed reads.

Despite the fact that these methods are able to give an estimate of true counts, to our knowledge there has been no quantitative evaluation of their ability to distinguish true from false sequences. In this paper we propose a framework for predicting true sequences from next generation sequencing data. We define true sequences as those that would have been output by an error-less sequencer. We approach this task as a binary classification problem, in which we learn a function mapping a feature set to 0 (sequences with zero real count) or 1 (sequences with non-zero real count). We defined six features for this task: observed count, estimated true count from error correction tool (RECOUNT), log likelihood with entropy penalty, log likelihood ratio, expectation matching score, and self-correctness coefficient (SCC).

Using an SVM classifier on simulated data sets, we show that these six features can achieve 96.35% accuracy. By making use of the predictor's posterior probability, we also suggest a way for users to determine sequences for their downstream analysis with an appropriate tradeoff between precision and recall.

This paper is organized as follows. First we describe our procedure for creating the simulated data set; second, we explain in detail our feature generation step; third, we show our experimental results, discuss the misclassified sequences, and conclude.

II. Data Simulator

In this section we describe our simulator to produce artificial datasets of sequence reads in which we know the

true sequence behind every read. Our simulator assumes a probability distribution of true sequences and of sequencer error, but rather than arbitrarily defining these distributions, we base them on real read datasets in a semi-empirical way described below. For lack of a more precise term, we call this procedure "training" the simulator on the given dataset. Our simulator first selects a pre-simulated dataset and then generates a post-simulated dataset, which simulates the reads which one might obtain in a real sequencing experiment.

A. Pre-simulated dataset

1) Extract 100K reads with high quality from a real experiment by randomly sampling reads with average Solexa quality greater than 33. [1]. Keep the sequences and discard their quality score information.
2) Randomly sample 100K reads of any quality from the same experiment, and retain only their quality data.
3) The final pre-simulated data set is obtained by randomly pairing the obtained sequences and quality score vectors.

B. Post-simulated dataset

We obtain this dataset by mutating bases in each sequence chosen above according to the probabilities stipulated by their assigned quality scores.

The *real* count of a sequence is its count in the *pre-error-simulated* set. Our classification procedure is to generate features from the *post-error-simulated* data set and use them to identify which sequences in that data set have zero or non-zero real count.

Subsequently we will simply denote the *pre-error-simulated* dataset as the *real* dataset and *post-error-simulated* set as the *simulated* dataset.

For the simulations described here, we trained our simulator on two real datasets: *B. vulgaris* genomic clone [6] reads (27bp) and *Drosophila* somatic cell mRNA [10] reads (36bp). Although read lengths generated by Illumina/Solexa recently have increased to 100bp or more, reads with shorter length are still often used in biological experiments.

III. Method

A. Features Generation

1) Observed Count: We define the observed sequence of a read in the usual way, as the sequence obtained by calling the base with highest probability according to the quality scores in each position. The observed count of a sequence s, is the number of reads whose observed sequence is s.

[1]Note that because of this threshold, depending on the type of data set and the quality score, the final number of sampled tags will be $\leq$ 100K.

2) Predicted True Count: We obtained the estimated true count from RECOUNT. Our RECOUNT software has been described elsewhere [18] and it closes follows the method of Beißbarth et al. [2] to infer true counts. For the readers convenience, we summarize the method here.

Mostly following the notation of Beißbarth et al. [2]; α_{ij} denotes the probability that a true sequence i generates an observed sequence j (when the read is called correctly $i = j$). u denotes the total number of unique sequences, which in principle includes all possible DNA sequences with the right length, but in practice we approximate by only considering sequences observed at least once). For the ith sequence, $1 \leq i \leq u$, n_i denotes its observed count and m_i its true count.

In forming a probability model, we assume the true sequence counts follow a Poisson distribution, namely given a true proportion p_j of a tag j, the true count is m_j with probability:

$$\frac{e^{-p_j\lambda}(p_j\lambda)^{m_j}}{m_j!} \tag{1}$$

for a fixed λ.

We adopt the Expectation Maximization algorithm [2], [5] to calculate the true counts given the observed counts and sequencing error rate estimates. The parameters we want to estimate are λ and the p_j's, under the constraint that the p_j's add to one. The log likelihood function is given by:

$$-\lambda + \sum_{j=1,\cdots,u} \hat{m}_j log(p_j\lambda) \tag{2}$$

The details of the EM algorithm are as follows:

1) E-step: Compute the likelihood and expected count of a sequence j given by:

$$\hat{m}_j = \sum_{i=1,\ldots,u} \left(\frac{\alpha_{ij}p_j}{\sum_{k=1,\ldots,u}\alpha_{ik}p_k} \cdot n_i \right) \tag{3}$$

2) M-step: Maximize the likelihood of the complete data given the expected values and re-calculate new estimates for the parameters: $\hat{\lambda} = \sum_{k=1,\ldots,u}\hat{m}_k$ and $\hat{p}_j = \hat{m}_j/n$, where n is the total read count. Note, the total read count is equivalent to the total sequence count, but different than the total *unique* sequence count.

We iterate these steps until the parameters converge. We initialize the expected values $\hat{m}_j$ with the observed count of read j. The following two features also make use of the log likelihood, formula (2).

3) Entropy Penalty to Log likelihood: In general, the sequences found in a biological sample are expected to cover only a small amount of fraction of the possible sequence space (e.g. the 4^{36} possible length 36 DNA sequences). Based on this prior knowledge, we introduce an extra entropy term to favor sparse solutions, i.e. solutions in which

the number of inferred unique sequences is relatively small. Our modified optimization function is:

$$-\lambda + \sum_{j=1,\cdots,u} \hat{m}_j log(p_j \lambda) + \beta \sum_{j=1,\ldots,u} p_j \log p_j \quad (4)$$

where $\beta \geq 0$ is a user-defined parameter which controls the tradeoff between the likelihood and the sparsity of the solution. When $\beta = 0$, we only focus on maximizing the log likelihood and ignore the sparsity constraint. The higher the value of β is, the more we focus on the model's sparsity.

Since there is no closed form expression for the parameters to maximize equation 4 with respect to the p_j's, we use the conjugate gradient method [9] for the "M step". The m_j's are then updated in the "E step" by equation 1 as in the standard EM procedure.

4) Zero-clamped Log Likelihood Ratio: Above we described the probabilistic model of RECOUNT, where the estimate true count for each observed sequence is determined by expectation maximization. At convergence we also obtain the log likelihood (from equation 2) of the observed data given the inferred true sequence counts. In other words, this is a kind of maximum likelihood estimation.

We conjectured that it would be useful to also compute the likelihood under the constraint that a sequence s of interest is forced to have a true count of zero. Our reasoning is that if clamping that sequence to zero significantly reduces the overall likelihood of the observed data, then it is likely that the sequence actually has a non-zero true count.

More formally, let L be the unconstrained likelihood and $L_{s=0}$ be the likelihood obtained when the true count of sequence s is clamped to zero. The zero-clamped log likelihood ratio for sequence s (or for brevity just *likelihood ratio*) is defined as $L - L_{s=0}$.

To compute this we modified RECOUNT to allow p_s for a specified sequence s to be clamped to zero during the calculation. Algorithm 1 illustrates the whole process.

Algorithm 1 Zero-clamped Log Likelihood Ratio

1: Run RECOUNT and compute log likelihood L
2: **for** Each sequence s in library **do**
3: Compute L_s by running RECOUNT with p_s clamped to zero.
4: **end for**

Time complexity: Unfortunately, the time complexity of this computation increases quadratically with the number of observed sequences N, which would make the procedure impractical if implemented naïvely. Fortunately, clamping a single probability p_s only affects a small portion of the likelihood computation and the rest can be efficiently reused.

To be more precise, we first define the *plausible misread graph* of an input dataset as an undirected graph whose nodes are the observed sequences and each edge connects two sequences s and r which are similar enough that a true sequence s could plausibly be misread as r (or vice versa). RECOUNT allows the user to stipulate this graph to be defined using either hamming distance or the probability of misreading s as r from the average quality scores of each read called as s. The results discussed here use one hamming distance edges.

In the plausible misread graph, consider C_s, the connected component of a sequence s. We note (without presenting a formal proof) that in the likelihood computation of equation 2, a change in p_s can only affect terms n_r, m_r for sequences r which are in C_s. Using this observation, we can compute $L_{s=0}$ efficiently by reusing the results from the unconstrained likelihood L and only recomputing the terms corresponding to C_s. Using this technique, the running time for generating $L_{s=0}$ for all observed sequence on 100K Solexa/Illumina reads was ~1.5 hours on a 2.9GHz 32 RAM Linux machine. Although we did not pursue this, we note in passing that the computation of the $L_{s=0}$'s is easily parallelizable.

5) Expectation Matching: We coin the term "Expectation Matching" to describe a simple alternative to maximizing an explicit likelihood as in equation 2. In short, expectation matching is an iterative heuristic procedure which takes advantage of the fact that, if we assume we know the true counts, it is easy to compute the expected observed counts.

More formally, the expected counts of the jth sequence are:

$$\text{sequence } j \text{ expected count} = \sum_{i=1,\cdots,u} m_i * \alpha_{ij} \quad (5)$$

where m_i is the assumed true count of the ith sequence and (as in previous sections) u denotes the total number of unique sequences, and α_{ij} denotes the probability that a read of sequence i is called as j. Let M denote an assumed true count vector and $E(M)$ denote the expected observed counts of M.

We define our task as finding an estimated true counts vector $\hat{M}$, such that $E(\hat{M})$ is close to the observed counts N. Algorithm 2 describes the procedure of expectation matching, which uses a form of gradient descent to improve an initial estimate of $\hat{M}$.

For the convergence criteria, we currently use the maximum (absolute value) sequence count change.

An advantage of this approach is its simplicity and speed. Also it conserves the total sequence count between N and M, which is reasonable since we only intend to model the miscalling of reads – not their loss or gain. Note that this property is not shared with the particular Expectation Maximization based approach described in previous sections. However this approach lacks an explicit probabilistic model and suffers from the fact that it sometimes infers negative counts for some sequences.

Algorithm 2 Expectation Matching

Require: γ holds learning rate
Require: η holds decay rate for γ
Require: N holds observed sequence counts
Let $M = N$
Let done = *false*
while !done **do**
  Let $\Delta = N - E(M)$
  Let $C = \Delta * \gamma$
  Let $\gamma = \gamma * \eta$
  Let $M = M + C$
  if change C small enough **then**
    Let done = *true*
  end if
end while

This approach is in the same spirit as the work by Colinge, et.al, [4], except that they use a sophisticated (and more time consuming) Lanczos numerical algorithm to compute $\hat{M}$.

6) Self-Correctness Coefficient (SCC): If an observed sequence s is false, it must be the case that every read which was called as s was in fact a misread. More formally, let Pc_{ij} denote the probability that the ith such read was called correctly in position j; a quantity indictaed by the quality scores of read i. Let l denote the read (and therefore sequence length) of the input data. The probability that all M_s reads called as sequence s, were miscalled is:

$$\mathrm{SCC}_s = 1 - \prod_{i=1..M_s} (1 - \prod_{j=1..l} Pc_{ij}) \qquad (6)$$

We call this the "Self-Correctness Coefficient", because it is a very simplistic measure which only considers reads called as s rather than the whole ensemble of reads. We speculate this coefficient should be able to compensate for a weakness of the ensemble methods as they are implemented. For efficiency reasons, we and others do not consider all possible sequences, but only those with non-zero observed counts. This approximation generally works well, but breaks down when there is a significant probability that a read comes from a sequence which has zero observed count. The result is that the ensemble methods are forced to accept all observed sequences which are isolated in the plausible misread graph.

IV. Results

For this study, we adopted the Support Vector Machine (SVM) classifier [15]. All the experiments were done with the SVM implementation under the klaR package for R [17], using a radial basis kernel function. The accuracy estimates were made by using leave-one-out cross-validation and selected from the best results from the following SVM's hyper-parameter ranges: C = (4, 8, 16, 32, 64, 128, 256) and γ = (0.5, 1, 2), yielding the combination: $C = 256$, $\gamma = 2$.

A. Prediction Accuracy

We measured the accuracy by using each feature by itself and with all six features. There are no hyper-parameters involved for generating the features except for the *entropy penalty*, for which we use $\beta = 100$. We use the simulator output when trained from *Beta vulgaris* dataset to test the accuracy. It contains 20,576 positive class and 25,632 negative class.

Table I shows the accuracy of prediction using single features. The best feature was the true count as inferred by the EM formulation. The observed count feature can be viewed as a sort of baseline. When we combined all six features we obtained a cross-validated accuracy of 96.35%.

Table I
ACCURACY USING ONLY SINGLE FEATURES

Feature	Accuracy
EM	94.73
EM + Entropy term	94.72
Observed Count	93.44
Expectation Matching	93.28
SCC	91.01
Log likelihood Ratio	89.70

B. Precision and Recall

One practical question biologists may ask when assessing next generation sequencing data is to ask how to select observed sequences which are correct at a given confidence level.

We attempt to answer this question by using the posterior probability given by the SVM classifier. Figure 1 shows a histogram of the posterior probability of true sequences. Although most sequences have probabilities near zero or one, some have probabilities near 0.9, which should perhaps be removed if the downstream analysis is highly sensitive to false positives.

We then determine two evaluation measures for the user to select the reads. Let θ be the user-defined confidence threshold. The precision and recall are defined as:

$$\text{Precision} = \frac{\#\text{true seqs with post. prob} \geq \theta}{\#\text{seqs with post. prob.} \geq \theta}$$

$$\text{Recall} = \frac{\#\text{true seqs with post. prob.} \geq \theta}{\#\text{true seqs}}$$

Precision gives the fraction of sequences classified as true that really are true sequences, while recall gives a fraction of true sequences that are classified as true.

Figure 2 shows two plots of precision and recall using simulations based on two different datasets. Both use six features for prediction. In these figures we can observe that

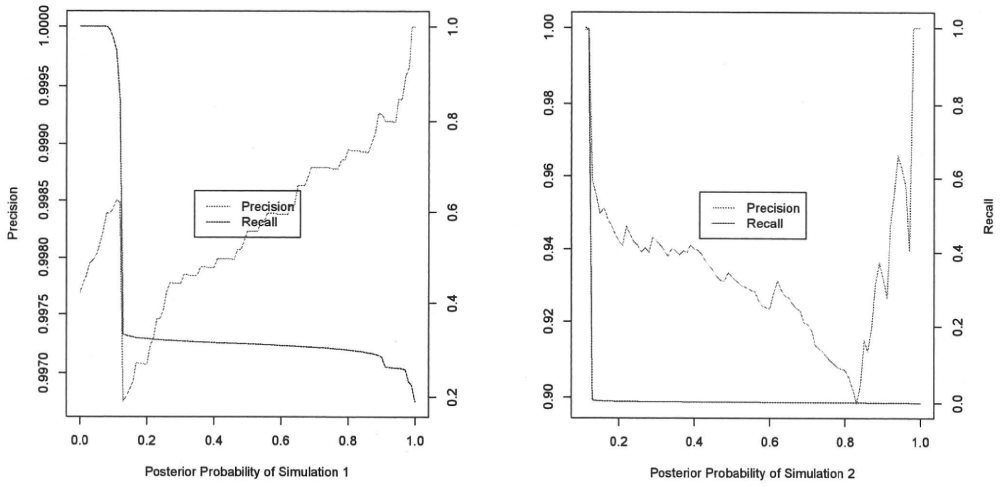


Figure 2. Precision and recall. The panel on the left (simulation 1) were generated using *Beta vulgaris* as dataset, and panel on the right (simulation 2) using a *Drosophila* dataset.

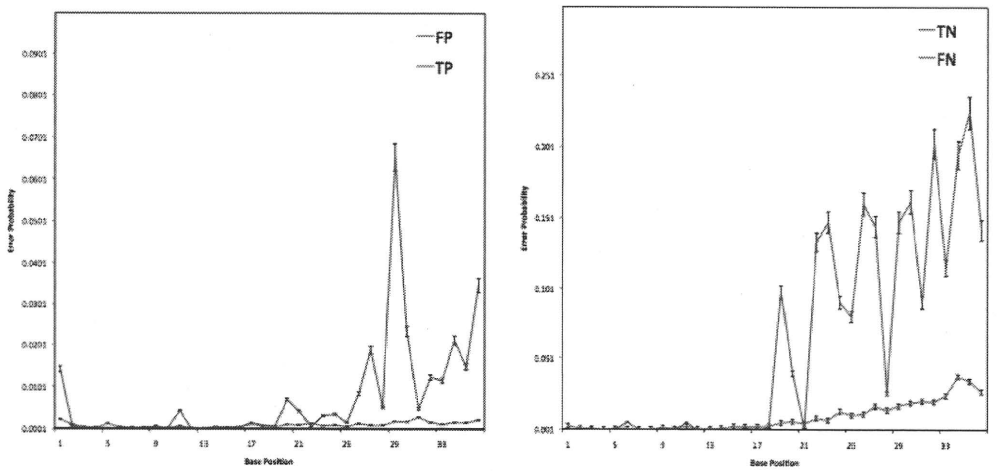


Figure 3. The average error probability by position (as given by the quality scores) is shown for correctly and incorrectly classified sequences

by using posterior probability ≤ 0.12 the user can obtain reads up to 99.98% precision. Also with the same posterior probability threshold the user can obtain reads with at least 99.6% recall.

C. Computation Time

The running time for generating all features on 100K Solexa/Illumina reads is ~2.5 hours with ~20MB memory, using 2.9GHz, 32GB RAM Linux machine. The most time consuming step was the computation of the Log likelihood-Ratio, which required ~1.5 hours.

D. Characteristics of Misclassified Reads

Figure 3 shows the error probabilities per position of correctly and incorrectly classified sequences (for sequences called from multiple reads, their harmonic average is used). It is not surprising that the negative data (false positive and true negatives) have a higher probability of error than the positive data. Amongst the positive sequences, the ones with low error probability tend to be correctly predicted – they claim they are correct with high confidence and the classifier accepts them. On the other hand, amongst the negatives sequences the ones with high error probability tend to be correctly predicted – they admit the may be misreads and the classifier judges that the are.