History of the Cisplatin and Pemetrexed Predictors

Keith A. Baggerly

May 10, 2010

Contents

1 Outline

In this report, we outline the history of the cisplatin and pemetrexed signatures introduced by Hsu et al. [6]. Many details are taken from Case Study 2 of Baggerly and Coombes [2], and the corresponding supplementary report available from `http://bioinformatics.mdanderson.org/Supplements/ReproRsch-All`.

2 The Full History

Of the Duke genomic signatures being used in clinical trials, those for cisplatin and pemetrexed have been in use the longest (see `http://clinicaltrials.gov/` entries for 509366, contrasting cisplatin with pemetrexed, and 545948, contrasting pemetrexed with vinorelbine). To our knowledge, the only paper describing the signatures for these two drugs is Hsu et al. [6], which appeared in *JCO* in October of 2007.

In November of 2007, when the Coombes et al. [4] critique of the underlying Potti et al. [7] paper in *Nature Medicine* appeared, Potti and Nevins [8] cited Hsu et al. [6] in rebuttal. Potti and Nevins [8] claimed Hsu et al. [6] as one of two examples of high impact studies where they had gotten their approach to work again, and used the existence of such examples to argue that the work of Coombes et al. [4] must be flawed. (The other example cited by Potti and Nevins [8] was Bonnefoi et al. [3], which was in press at the time of the rebuttal and didn't appear until December of 2007.)

The initial Potti et al. [7] paper constructed genomic signatures from the NCI60 panel of cell lines, using U95A array profiles available at the time. This same approach was used by Hsu et al. [6] in constructing the signature for pemetrexed. However, Hsu et al. [6] noted that

> The collection of data in the NCI-60 data occasionally does not represent a significant diversity in resistant and sensitive cell lines to any given drug. Thus, if a drug screening experiment did not result in widely variable GI50/IC50 and/or LC50 data, the generation of a genomic predictor is not possible using our methods, as in the case of cisplatin.

Thus, Hsu et al. [6] assembled the cisplatin signature from a panel of 30 cell lines profiled by Györffy et al. [5]. Györffy et al. supply both U133A array quantifications and classifications of which cell lines were sensitive, intermediate or resistant in their response to various drugs, including cisplatin (their Figure 2). Hsu et al. [6] note that

> The cisplatin sensitivity predictor includes DNA repair genes such as ERCC1 and ERCC4, among others, that had altered expression in the list of cisplatin sensitivity predictor genes. Interestingly, one previously described mechanism of resistance to cisplatin therapy results from the increased capacity of cancer cells to repair DNA damage incurred, by activation of DNA repair genes.

Using the data available, we attempted to reconstruct the heatmaps and signatures reported by Hsu et al. [6]. This reconstruction made use of several files, described below.

- The NCI60 quantification data used to produce the pemetrexed signature is available from `http://dtp.nci.nih.gov/mtargets/download.html` (only quantifications with an "A" suffix are used).
- The Györffy et al. [5] quantification data used to produce the cisplatin signature is available as a supplementary file from the *International Journal of Cancer*, at `http://www.interscience.wiley.com/jpages/0020-0136/suppmat/2006/jws-ijc.21570.tbl1.xls`.
- The binreg software used by Potti et al. [7] was obtained from `http://data.genome.duke.edu/NatureMedicine.php` (that site is currently down, but copies are available from `http://bioinformatics.mdanderson.org/Supplements/ReproRsch-All`; other copies were independently acquired elsewhere and names of those having such copies will be made available upon request).
- Lists of the Affymetrix probesets comprising the two signatures were provided as supplements to Hsu et al. [6]; these are available from `http://jco.ascopubs.org/cgi/content/full/25/28/DC1`.
- Drug sensitivity data for pemetrexed (GI50 values for NSC 698037) is available from `http://dtp.nci.nih.gov/docs/cancer/cancer_data.html`.
- Drug sensitivity data for cisplatin was drawn from Figure 2 of Györffy et al. [5].

We reconstructed the heatmaps given in Figure 1 of Hsu et al. [6] using the above files as described in the supplementary reports to Baggerly and Coombes [2]. This reconstruction let us identify both the cell lines and genes involved in the signatures, showing several problems.

1. According to the drug sensitivity data available from the NCI, the sensitive/resistant labels for pemetrexed shown in Figure 1 of Hsu et al. [6] were reversed. The GI50 values for the two sets of cell lines involved actually overlap, because one cell line (TK-10) was placed in the wrong group.
2. All 85 probesets reported for pemetrexed are incorrect due to the same off-by-one indexing error that Coombes et al. [4] noted in the signatures reported by Potti et al. [7] (we note that Potti and Nevins [8] admitted the gene lists initially presented in Potti et al. [7] were incorrect).
3. All 45 probesets reported for cisplatin are incorrect, 41/45 due to the off-by-one error noted above, and 4/45 because they are not produced by their software at all.
4. The four cisplatin probesets their software does not produce are 203719_at (ERCC1), 210158_at (ERCC4), 228131_at (ERCC1) and 231971_at (FANCM, associated with DNA Repair), which Hsu et al. [6] explicitly mention as interesting components of the signature.

5. Checking annotation from Affymetrix (`http://www.affymetrix.com`) shows two of the four problematic probesets, 228131_at (ERCC1) and 231971_at, are not present on the U133A arrays used (they're on the U133B platform), so they weren't even measured.

We first presented the above problems in talks I gave at the NCI on November 7 of 2007 (presentation attached). The incorrect gene lists are still posted at *JCO* as of April 7, 2010.

We submitted a letter summarizing the above problems to *JCO* on November 5, 2007, when it was assigned the reference number JCO/2007/151985. On Friday, December 14, we received a note from *JCO* stating they had reviewed our correspondence, but "regret to inform you that we cannot accept your correspondence for publication." No further explanation was given. On Monday, December 17, we sent a followup note to the *JCO* editorial office asking for clarification. On January 4 of 2008, we received a rejection letter identical to the one from December 14, with no further explanation provided.

We then incorporated this objection into a list of problems we sent as a second correspondence to *Nature Medicine* as of May 30, 2008. On June 2, the editors requested permission to share our letter with Potti et al.; permission was granted on June 3. On June 11, *Nature Medicine* declined our correspondence, citing the "detailed response" given by Drs. Potti and Nevins.

In January of 2009, the cisplatin heatmap from Figure 1 of Hsu et al. [6] was later shown in Figure 4 of Augustine et al. [1], where it was listed as being derived from the NCI60 cell lines as a heat map for temozolomide. The latter assertion is wrong; this heatmap is not from the NCI60 quantifications, nor is it for temozolomide.

Later, around November 6 of 2009, the Duke group posted a supplementary web page for Hsu et al. [6] at `http://data.genome.duke.edu/JCO.php`. We're not sure exactly when this was first posted, but screen shots of pages higher in the hierarchy show no evidence of it as of November 1, 2009, so we assume it was posted after that. This timing places the posting after the online publication of Baggerly and Coombes [2] and the suspension of the clinical trials involved, and in the middle of an internal investigation commissioned by Duke University (as covered in *The Cancer Letter*, in issues from Oct 2, 9, and 23 of 2009).

Data provided on the above web page included lists of the cell lines nominally involved, numerical quantifications of these cell lines, new gene lists for the signatures, and samples labeled as validation data used in constructing the signatures. Our full report on these data is available from `http://bioinformatics.mdanderson.org/Supplements/ReproRsch-All/Modified`. Our main conclusions are quoted below:

> There are several problems present in the data now posted. In our assessment, three of these are fatal flaws with respect to building a signature:
>
> 1. The sensitive and resistant labels for the pemetrexed signature are reversed. If the method works as advertised, and they use this signature to guide treatment, then patients will be actively guided to the wrong therapy.
> 2. At least 49/59, and possibly all, of the validation samples are mislabeled. All claims about how well these signatures work clinically are based on how well they predict outcomes for patient samples, and if you scramble the labels, you're predicting the wrong things.
> 3. For 16/59 validation samples, the genes are mislabeled, and not in a manner that immediately suggests a simple fix (like an off-by-one error). As with point 2, this means that for these samples, they're predicting the wrong thing. Further, this discrepancy makes these samples "look different", and to the extent that one group is overrepresented in these samples (i.e., if they think these are all "responders"), this can make the classification problem inappropriately easy, and potentially bias the results.
>
> More concisely,
>
> 1. The sensitivity labels are wrong.

2. The sample labels are wrong.
3. The gene labels are wrong.

We submitted our report, code and documentation to the Duke investigation and to the NCI on November 9, 2009, and November 10, 2009, respectively. The data files were stripped from the web site within days. The overall data page, `http://data.genome.duke.edu`, and its subsidiary pages (e.g., the supplementary page for Potti et al. [7]) were also disabled in about a week. All of these pages were still disabled at the start of April 2010 (they weren't there April 1), but were reactivated (without raw data files) as of mid-April, 2010 (they were there by April 8).

We note in passing that the gene lists reported on the now-disabled page for Hsu et al. [6] had corrected the off-by-one error, so as of November 2009, their own computations showed several of the objections raised by Baggerly and Coombes [2] were valid. However, they evidently have not seen the need to correct the record with *JCO*, given the incorrect tables still posted there.

It would be an act of great chutzpah for the Duke group to submit a methodology paper on how to do things "right" to any of the journals where problems have been identified before fixing the errors in their previous submissions there.

3 Appendix

3.1 File Location

```
> getwd()
[1] "/Users/kabagg/ReproRsch/HistoryOfCisPem"
```

References

[1] Augustine CK, Yoo JS, Potti A, et al.: Genomic and molecular profiling predicts response to temozolomide in melanoma. *Clin Cancer Res*, **15**:502-10, 2009.

[2] Baggerly KA, Coombes KR: Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology. *Ann App Statist*, **3**:1309-34, 2009.

[3] Bonnefoi H, Potti A, Delorenzi M, et al.: Validation of gene signatures that predict the response of breast cancer to neoadjuvant chemotherapy: a substudy of the EORTC 10994/BIG 00-01 clinical trial. *Lancet Oncology*, **8**:1071-8, 2007.

[4] Coombes KR, Wang J, Baggerly KA: Microarrays: retracing steps. *Nat Med*, **13**:1276-7, 2007. Author reply, 1277-8.

[5] Györffy B, Surowiak P, Kiesslich O, et al.: Gene expression profiling of 30 cancer cell lines predicts resistance towards 11 anticancer drugs at clinically achieved concentrations. *Int J Cancer*, **118**:1699-712, 2006.

[6] Hsu DS, Balakumaran BS, Acharya CR, et al.: Pharmacogenomic strategies provide a rational approach to the treatment of cisplatin-resistant patients with advanced cancer. *J Clin Oncol*, **25**:4350-4357, 2007

[7] Potti A, Dressman HK, Bild A, et al: Genomic signatures to guide the use of chemotherapeutics. *Nat Med*, **12**:1294-1300, 2006.

[8] Potti A, Nevins JR: Reply to Microarrays: retracing steps. *Nat Med*, **13**:1277-8, 2007.