J. Jedediah Smith

BIFX 545

Discussion Questions Week 5

2/22/2022

1. Before reading any of the new papers for this week, first read back through "Genomic signatures to guide the use of chemotherapeutics" and make a list of the experiments, leaving room for comments after each one.

Built a model using in vivo drug responses together with microarray gene response data.

* Goal was to use gene expression data to predict how tumors might react to a certain type of chemotherapy drug, and whether it would help destroy them.
* Their approach seems solid. Using how genes respond to a particular drug as a metric for how the tumor might respond to said drug seems like it would work.
* Could not reproduce their selection of cell lines. List of genes reported in supplemental materials were wrong due to an off-by-one error.
* Training and test data was mislabeled for multiple drugs.
* Important genes were not derived from training data.
* Some of the samples assayed were duplicated.

Applied model to lots of other drugs.

* After getting the model to work with one specific drug, it was applied to lots of other drugs. If one drug wasn’t predicted to work well for a tumor, maybe another would.
* They seem to have a suspiciously high accuracy rate, and their sample sizes were rather small. Their approach still looks okay, but their results seem too good to be true.
* After correcting for off-by-one error, only 3 of 7 drug predictions still held up.
* Software does not maintain the independence of training and test sets. When same methods are repeated but with maintained separation, prediction results were very poor.
* Article retracted due to the inability to reproduce critical aspects of the experiment with several of the drugs tested.

Applied model to multiple drugs at the same time.

* After getting the model to work with different types of drugs, it was applied to combination drug remiges to see of certain drugs would be better or worse together.
* This seems to be potentially stretching the limit of their model. I would want to look over their source code to see if they made any changes from their initial model.

Tried to find other pathways that contain the predictor genes studied.

* Would potentially allow them to identify novel targets.
* Although some of their previous results were concerning, it seems reasonable to conclude that novel targets might be identified through studying the pathways of predictor genes.
* Cell line sensitive/resistant designations are reversed.

1. For each experiment, indicate what the experiment was intended to determine in the space that you left for comments.

(See bullet points under the experiments listed for Question 1.)

1. To the extent that you can tell based on what was written in "Genomic signatures to guide the use of chemotherapeutics", was each experiment performed and reported correctly? Write down whether you believe the outcomes of each experiment in the space that you left for comments and explain why or why not.

(See bullet points under the experiments listed for Question 1.)

1. Now go ahead and read through this week's papers in the following order:

* File 1. MicroarraysRetracingSteps.pdf
* File 2. PottiResponse1.pdf
* File 3. PottiCorrection1.pdf
* File 4. natMedLetter\_RetracingStepsAgain.pdf
* File 5. PottiCorrection2.pdf
* File 6. PottiRetraction.pdf

As you read through, use the room that you left for comments after each experiment to keep track of which comments had errors and what they were.

(See bullet points under the experiments listed for Question 1.)

1. In class we discussed that the following questions were asked by the authors of "Genomic signatures to guide the use of chemotherapeutics":
2. Can gene expression and drug sensitivity data from cell lines be used to predict which used should be used to treat patients?
3. Can the data be used to determine combinations of drugs that can be used to treat patients?
4. Can the data be used to identify pathways of genes that are up/down regulated in response to treatment that represent potential new therapeutic targets?

Considering your own observations as well as the observations made in the follow-up papers, which of the original questions (if any) were NOT satisfactorily answered in "Genomic signatures to guide the use of chemotherapeutics"? Be specific as to why the question was not satisfactorily answered i.e. what were the problems with the experiments and or data used to answer each question?

The approaches they tried to use to answer these questions seemed very promising. But the way their handled their data and model building was inadequate to definitely answer most of their questions. While it stands to reason that all three of the points above might be possible, the researchers in this paper really failed to conclusively prove it owing to a myriad of technical errors and oversights that should not have occurred.

* 1. There were major problems with their model, from failure to maintain independence between the test vs training data to off-by-one errors. When these things were fixed, the accuracy of predictions dropped catastrophically.
  2. Presumably used the same model as the single drug tests. Therefore, it will probably subjected to the same problems. The results can be discounted.
  3. It was shown that the researchers mislabeled their test vs training data in some places, calling into question whether their identified pathways are even relevant.

1. Were there any other problems with "Genomic signatures to guide the use of chemotherapeutics"?

Some of their important genes were not derived from training data. It was also shown that some of their samples assayed were duplicated. Ultimately, it proved impossible to replicate their results using corrected techniques and the paper was retracted.

1. Who (if anyone) was at fault for the fact that "Genomic signatures to guide the use of chemotherapeutics" was published despite its problems? This may be multiple people, so be specific and state why.

Ultimately, the people most responsible was the researchers who wrote this paper, for failing to recognize the problems with their experiment and continuing to submit it for publication anyways. It feels like they were too hasty and rather negligent. But there should also be some blame attributed to the publishers for not being more vigilant, and ultimately making the decision to publish it despite the results appearing too good to be true. I don’t think it is fair to hold any peer reviewers responsible since the initial approach of the paper seemed quite promising, and it would take someone familiar with data science and or microarray techniques to notice the full extent of errors made in this paper.

1. What should the person or people at fault (if any) have done differently?

Researchers should have been more diligent with their data keeping. Publishers should have taken note of the bombshell claims and suspiciously good results, then done more rigorous peer reviews of the paper in question to ensure it veracity.

1. What consequences (if any) should that person or those people face?

I’m genuinely not sure what the right answer is here. If no malice or deliberate fraud can be ascribed to the researchers and they were just incompetent, then there should be no criminal charges. Perhaps some of the scientists involved will lose funding or their job. The biggest blow by far will be to their reputation. As for the publishers, maybe someone loses their job for negligence, or people don’t buy their paper as much.