Using Random Peptide Phage Display Libraries for early Breast cancer detection

Ekaterina Nenastyeva[1], Yurij Ionov[2], Ion Mandoiu[3], and Alex Zelikovsky[1]

[1] Department of Computer Science, Georgia State University, Atlanta, GA 30303
Email: {enenastyeva1,alexz}@cs.gsu.edu
[2] Department of Cancer Genetics, Roswell Park Cancer Institute, Buffalo, NY 14263
Email: Yurij.Ionov@roswellpark.org
[3] Department of Computer Science & Engineering, University of Connecticut, Storrs, CT 06269 Email : ion@engr.uconn.edu

Abstract. Thousands of people beat cancer every year. Doing so is easier when cancer is diagnosed at an early stage as treatment is often simpler and more likely to be effective. Cancer cells starts out as normal body cells, but they begin to grow out of control because of an abnormal gene expression. The immune system plays a major role in limiting the development of these abnormalities. It elicits a detectable humoral immune response to changes in antigen profiles caused by growing cancer cells. Circulating autoantibodies produced by the patient's own immune system after exposure to cancer proteins are promising biomarkers for the early detection of cancer.

Since an antibody recognizes not the whole antigen but 4-7 critical amino acids within the antigenic determinant (epitope), the whole proteome can be represented by random peptide phage display libraries (RPPDL). To solve cancer detection problem we propose a new method based on RPPDL. We determined that peptides assigned to breast cancer serum samples better correlate with each other than peptides assigned to control serum samples. Thus, the cancer samples had common features in immune response.

We tested our method on the serum antibody repertoire profiles for 5 stage 0 breast cancer patients and for 5 cancer-free women. As result all samples were predicted correctly except one cancer which gave sensitivity equaled 0.8, specificity equaled 1 and accuracy equaled 0.9.

Keywords: Breast cancer. Random peptide phage display library. Early cancer detection. Immune response. Pearson correlation.

1 Introduction

Thousands of people beat cancer every year. Doing so is easier when cancer is diagnosed at an early stage as treatment is often simpler and more likely to be effective. So finding cancer early can make a real difference.

1.1 Motivation

Cancer cells starts out as normal body cells, but they begin to grow out of control because of an abnormal gene expression. The immune system plays a major role in limiting the development of these abnormalities. There are multiple lines of evidence that the immune system elicits a detectable humoral immune response to changes in antigen profiles caused by growing cancer cells [1], [2], [3], [4], [6].

Circulating autoantibodies produced by the patient's own immune system after exposure to cancer proteins are promising biomarkers for the early detection of cancer. An advantage of autoantibodies in cancer detection is their production in large quantities, despite the presence of a relatively small amount of the corresponding antigen. It has been demonstrated, that panels of antibody reactivities can be used for detecting cancer with high sensitivity and specificity [7].

1.2 Previous works

The current methods of analysis of antitumor humoral immune response, such as SEREX, SERPA, antigen microarrays, or ELISA are designed to detect high-affinity/high-titer IgG or IgM antibodies. However, the immune system can react to alterations in local antigenic compositions caused by growing tumors by producing a variety of low-affinity/low-titer antibodies. There is a need to develop more sensitive method.

Recently the authors tested whether immunosignatures correspond to clinical classifications of disease using samples from people with brain tumors. The immunosignaturing platform distinguished not only brain cancer from controls, but also pathologically important features about the tumor including type and grade [5]. These results clearly demonstrate that random peptide arrays can be applied to profiling serum antibody repertoires for detection of cancer. The important advantage of using peptide arrays instead of antigen arrays is that peptides can mimic not only the protein epitopes but also the carbohydrate epitopes that represent an essential part of the repertoire of cancer-associated autoantibodies.

1.3 Biological mechanism

Since an antibody recognizes not the whole antigen but 4-7 critical amino acids within the antigenic determinant (epitope), the whole proteome can be represented by random peptide phage display libraries (RPPDL). Also, it has been demonstrated that for any antibody the peptide motif representing the best binder can be selected from the RPPDL. The RPPDL are widely used for identifying the epitope specificity of monoclonal antibodies. The obstacle for using RPPDL as diagnostic tools was the necessity to sequence large number of individual phage DNA for identifying epitopes recognized by antibodies. The next generation (next-gen) sequencing technology makes possible to identify all the epitopes recognized by all antibodies contained in the human serum using one run of the sequencing machine. By screening human serum samples from breast cancer patients and healthy donors using RPPDL and next-gen sequencing we identify as biomarkers of cancer not the whole TAAs, but short peptide 7-mer sequences recognized by cancer associated autoantibodies.

1.4 Approach

To solve cancer detection problem we propose a new method based on RPPDL and Pearson correlation. We determined that peptides assigned to breast cancer serum samples better correlate with each other than peptides assigned to control serum samples. Thus, the cancer samples had common features in immune response. Using that property we calculated pairwise correlation between whole lists of peptides from different samples. As result we were able to identify a group of high correlated cancer samples.

2 Experiment setting

The serum antibody repertoire profiles for 5 stage 0 breast cancer patients and for 5 cancer-free women were generated by next-gen sequencing of peptide-coding DNA from phage selected from the RPPDL for binding to serum IgG antibodies. The flowchart in Figure 1 represents the experiment that was performed in duplicates for every human serum sample.

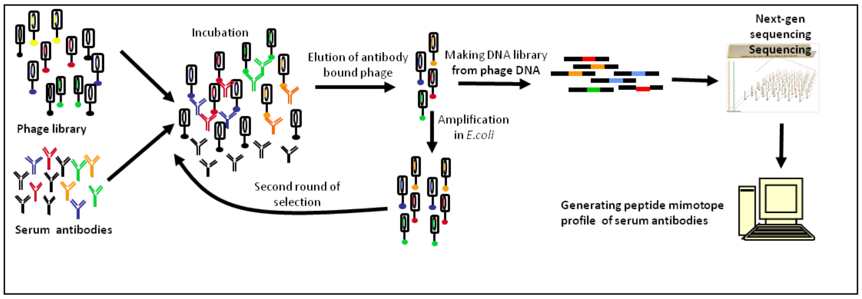
Phage library
Serum antibodies
Incubation
Elution of antibody bound phage
Making DNA library from phage DNA
Next-gen sequencing
Sequencing
Amplification in E.coli
Second round of selection
Generating peptide mimotope profile of serum antibodies

Fig. 1. A scheme for generating mimotope profiles of serum antibody repertoire.

Ph.D-7 phage displayed library of 7-mer random peptides was mixed with the serum and incubated overnight. Phage bound to antibodies was isolated using protein-G beads and eluted from the beads using low pH buffer. The eluted phage was amplified by propagation in E.coli and the amplified library was incubated with the same serum. The phage bound to antibodies was isolated using protein-G beads and the phage DNA was PCR amplified with the primers flanking the peptide coding insert. The library of peptide coding inserts was next-gen sequenced and the DNA sequences were translated into the peptide sequences.

3 Data preprocessing

After sequencing reads were translated into the peptide sequences for each peptide variant the number indicating how many times this peptide variant was encountered for each serum sample were assigned. Then obtained results were normalized. In average, for the experimental condition selected, the total number of distinct peptide sequences generated in one experiment was $3 * 10^6$. This number represents only 0.3 % of all possible 7-mer peptide sequences contained in the aliquot of the library used for the experiment.

As all 10 samples were done in duplicate totally we had 20 experiments (or replicas). We rationalized that as the measure of the reproducibility of the method we could consider the correlation of peptide abundance between the two replicas of the same profile generated by using the same serum sample and the same library independently. The low correlation between the replicas would indicate that the method detects only unspecific noise. In reality the correlation between the two replicas of the same serum for the all 10 serum samples was comparatively high, ranging between 0.68 and 0.99 with an average of 0.87. For the method we used Pearson correlation.

4 Method

As the result of the data preprocessing for every replica there was a list of assigned peptides. In turn, all those peptides had corresponding expression levels. We calculated the average correlation of the peptides' expression between the two replicas belonging to the different samples within the cancer set, the average correlation between the two replicas belonging to the different samples within the control set and the average correlation between the replicas from the cancer and the control sets.

The results showed the highest average correlation between the cancer samples equal 0.12. Between the cancer and control sets the correlation was 0.03. Finally, between control samples the correlation was the smallest and equal 0.02. Thus, the average correlation between the two replicas within the same sample was significantly higher than the overlap between the two replicas of the two different samples. This demonstrates that the method is reproducible and that the list of peptide sequences shared between the two replicas of the same serum sample is not the reduced list of random peptides but represents the profile of serum antibody repertoire. On the other hand, the average correlation between cancer samples was higher than the average correlation between the replicas from the cancer and the control sets. Based on the last property we proposed next algorithm for the breast cancer prediction:

- find the average correlation between known cancers (S)
- find the average correlation between known cancers and controls (L)
- find the average correlation between any unknown sample X and all known cancers (A)
- classify X to be the cancer if A >= (S+L)/2 and control otherwise

5 Results and discussion

To verify the accuracy of our method we used two techniques: cross-validation and permutation test. For cross-validation we applied described algorithm to available 10 serum samples trying to correctly predict cancers and controls. Every time we considered one of the ten samples as unknown X and used others to establish S, L, A and classified X as control or cancer. As result all samples were predicted correctly except one cancer which gave sensitivity equaled 0.8, specificity equaled 1 and accuracy equaled 0.9. To perform permutation test our method were run swapping control and cancer serum samples. Totally, there were 252 possible permutations or, in other words, ways of assigning case status to 5 out of 10 samples. The real result was in the top 2% of the best permutations according accuracy, sensitivity and specificity proving the reliability of the method.

In addition, we decided to analyzed the profiles of serum samples to identify the peptides associated with cancer. We use the following very stringent criteria for a peptide to be specific to breast cancer – its minimum expression level among the 10 replicas for breast cancer patients should exceed its maximum expression level among the 10 replicas for healthy donors. According to the above criteria, there is a single 7-mer peptide, 9 6-mer peptides, and 44 5-mer peptides specific to breast cancer. On the other hand, there are no 7-, 6-, and 5-mer peptide specific to healthy donors. Using permutation test, we found that the above property is statistically significant, namely, the p-value is less than 3%.

Although the number of serum samples used for our experiment is too low to have significant statistical power, identified peptides allow screening a large number of serum samples specifically for the reactivity against these peptides. The design of a quantitative PCR based immunoassay to screen large number of serum samples for analyzing reactivity against identified peptides is currently underway.

References

1. Gavin P Dunn, Allen T Bruce, Hiroaki Ikeda, Lloyd J Old, Robert D Schreiber, et al. Cancer immunoediting: from immunosurveillance to tumor escape. *Nature immunology*, 3(11):991–998, 2002.
2. Gavin P Dunn, Lloyd J Old, and Robert D Schreiber. The immunobiology of cancer immunosurveillance and immunoediting. *Immunity*, 21(2):137–148, 2004.
3. Sacha Gnjatic, Erika Ritter, Markus W Büchler, Nathalia A Giese, Benedikt Brors, Claudia Frei, Anne Murray, Niels Halama, Inka Zörnig, Yao-Tseng Chen, et al. Seromic profiling of ovarian and pancreatic cancer. *Proceedings of the National Academy of Sciences*, 107(11):5088–5093, 2010.
4. Mitchell Ho, Raffit Hassan, Jingli Zhang, Qing-cheng Wang, Masanori Onda, Tapan Bera, and Ira Pastan. Humoral immune response to mesothelin in mesothelioma and ovarian cancer patients. *Clinical cancer research*, 11(10):3814–3820, 2005.
5. Alexa K Hughes, Zbigniew Cichacz, Adrienne Scheck, Stephen W Coons, Stephen Albert Johnston, and Phillip Stafford. Immunosignaturing can detect products from molecular markers in brain cancer. *PloS one*, 7(7):e40201, 2012.

6. Arun Sreekumar, Bharathi Laxman, Daniel R Rhodes, Srilakshmi Bhagavathula, Jason Harwood, Donald Giacherio, Debashis Ghosh, Martin G Sanda, Mark A Rubin, and Arul M Chinnaiyan. Humoral immune response to α-methylacyl-coa racemase and prostate cancer. *Journal of the National Cancer Institute*, 96(11):834–843, 2004.
7. Li Zhong, Sarah P Coe, Arnold J Stromberg, Nada H Khattar, James R Jett, and Edward A Hirschowitz. Profiling tumor-associated antibodies for early detection of non-small cell lung cancer. *Journal of Thoracic Oncology*, 1(6):513–519, 2006.