Identification of Cancer-Specific Motifs in Mimotope Profiles of Serum Antibody Repertoire

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.
For fighting cancer, earlier detection is crucial. Circulating auto-antibodies produced by the patient's own immune system after exposure to cancer proteins are promising bio-markers 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 a random peptide phage display library (RPPDL). This opens the possibility to develop an early cancer detection test based on a set of peptide sequences identified by comparing cancer patients' and healthy donors' global peptide profiles of antibody specificities.
Due to the enormously large number of peptide sequences contained in global peptide profiles generated by next generation sequencing, the large number of cancer and control sera is required to identify cancer-specific peptides with high degree of statistical significance. To decrease the number of peptides in profiles generated by nextgen sequencing without losing cancer-specific sequences we used for generation of profiles the phage library enriched by panning on the pool of cancer sera. To further decrease the complexity of profiles we used computational methods for transforming a list of peptides constituting the mimotope profiles to the list motifs formed by similar peptide sequences.
We have shown that the amino-acid order is meaningful in mimotope motifs since they contain significantly more peptides then motifs among peptides where amino-acids are randomly permuted. Also the single sample motifs significantly differ from motifs in peptides drawn from multiple samples. Finally, multiple cancer-specific motifs have been identified.

Keywords: random peptide phage display library, early cancer detection, immune response, peptide motifs, mimotope profile

1 Introduction

Circulating autoantibodies produced by the patient's own immune system after exposure to cancer proteins are promising biomarkers for the early detection

of cancer. It has been demonstrated, that panels of antibody reactivities can be used for detecting cancer with high sensitivity and specificity [5].

The whole proteome can be represented by random peptide phage display libraries (RPPDL). For any antibody the peptide motif representing the best binder can be selected from the RPPDL. 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.

Recent studies tested whether immunosignatures correspond to clinical classifications of disease using samples from people with brain tumors [2]. The immunosignaturing platform distinguished not only brain cancer from controls, but also pathologically important features about the tumor including type and grade. These results clearly demonstrate that random peptide arrays can be applied to profiling serum antibody repertoires for detection of cancer

The profiles generated by next-gen sequencing following several iterative round of affinity selection and amplification in bacteria can consist of millions of peptide sequences. A significant fraction of these sequences is not related to the repertoires of antibody specificities, but produced by nonspecific binding and preferential amplification in bacteria. The presence of high amounts of these unspecific, quickly growing "parasitic" sequences can complicate the analysis of serum antibody specificities.

Considering that the affinity selected sequences can be clustered into the groups of similar sequences with shared consensus motifs, while the parasitic sequences are usually represented by single copies, we propose a novel motif identification method (CMIM) based on CAST clustering [1].

We have shown that the amino-acid order is meaningful in mimotope motifs found by CMIM – the CMIM motifs identified in observed samples contain significantly more peptides then motifs among the same peptides but with amino-acids randomly permuted. Also the single sample motifs are shown to be significantly different from motifs in peptides drawn from multiple samples.

CMIM was applied to case-control data and identified numerous cancer-specific motifs. Although no motif is statistically significant after adjusting to multiple testing, we have shown that the number of found motifs is much larger than expected and may therefore contain useful cancer markers.

2 Generating Mimotope Profiles of Serum Antibody Repertoire

The experiment for generating mimotope profiles of serum antibody repertoire is outlined in the flowchart in figure 1. The first step of the experiment was library enrichment, the second step was directly generating of mimotipe profiles and next-gen sequencing.

Library enrichment. Pooled serum from eight stage 0 breast cancer patients were used for enrichment of the library. The enrichment was performed as follows. Twenty μl of pooled serum and 10 μl of the Ph.D.7 random peptide library

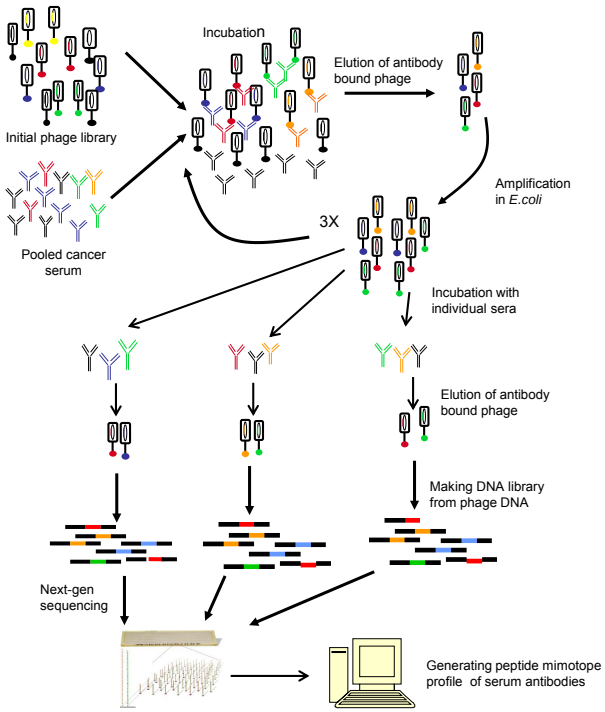

Fig. 1: A scheme for generating mimotope profiles of serum antibody repertoire.

(NEB) were diluted in 200 μl of the Tris Buffered Saline (TBST) buffer containing 0.1% Tween 20 and 1% BSA and incubated overnight at room temperature. The phages bound to antibodies were isolated by adding 20 μl of protein G agarose beads (Santa Cruz) to the phage –antibody mixture and incubating for 1 hour. To eliminate the unbound phage the mixture with beads was transferred to the well of 96-well MultiScreen-Mesh Filter plate (Millipore) containing 20 μm pore size nylon mesh at the bottom. The unbound phage was removed by applying vacuum to the outside of the nylon mesh using micropipette tip. The beads were washed 4 times by adding to the well 100 μl of TBST buffer and removing the liquid by applying vacuum to the outside of the nylon mesh using micropipette tip. The phage bound to the antibodies was eluted by adding to the beads of 100 μl of 100 mM Tris-glycine buffer pH 2.2 followed by neutralization using 20 μl 1 M Tris buffer pH 9.1. The eluted phages were amplified in bacteria by infecting 3 ml of an early log-phase culture . The amplified phages were isolated by precipitating phage with $^1/_6$ volume of 20% PEG, 05.M NaCl precipitation buffer. The cycle of incubation-bound phage isolation-amplification was repeated two more times and the isolated after the 3rd amplification library was used for analyzing antibody repertoires.

Generating peptide profiles Twenty μl of serum and 10 μl of the enriched library were diluted in 200 μl of the Tris Buffered Saline (TBST) buffer containing 0.1% Tween 20 and 1% BSA and incubated overnight at room temperature. The phages bound to antibodies were isolated using low pH buffer as described above for the enrichment of the library and the phage DNA was isolated using phenol-chloroform extraction and ethanol precipitation. The 21 nt long DNA fragments coding for random peptides were PCR-amplified using primers containing a sequence for annealing to the Illumina flow cell, the sequence complementary to the Illumina sequencing primer and the 4 nt barcode sequence for multiplexing. The PCR-amplified DNA library was purified on agarose gemultiplexed and sequenced by 50 cycle HiSeq 2500 platform.

The sequences were de-multiplexed to determine its source sample. The 21-base nucleotides were extracted between base position 29 and 49 and translated to 7-amino-acid peptide using the first frame. Any peptide containing stop codon was discarded.

3 CAST-based motif identification method

A motif was defined as a group of peptides having common sequence pattern. If we consider a motif as a cluster formed by peptides with a center represented by a consensus sequence then construction of a motif corresponds to a difficult clustering problem with many closely located centers. The radius of a cluster may exceed the distance from one cluster to another one. The standard clustering techniques (for example, k-means clustering [4]) are not applicable to current problem. Thence, for our purpose we modified more suitable CAST algorithm [1].

For motif finding we defined *similarity* measurement based on *Hamming distance*. The Hamming distance $HD(a, b)$ between sequences a and b of equal length is defined as the number of positions where the corresponding symbols are different. We extended the concept of Hamming distance to considering also shifts of sequences relative to each other. Thus the distance was computed on all sufficiently long overlaps between sequences a and b. We define *similarity* as following:

$$similarity(a, b) = l - minimum\{HD(a, b), HD(a_1, b_1), ...HD(a_n, b_n)\}, \quad (1)$$

where l - the length of a peptide, $HD(a_1, b_1)$,..., $HD(a_n, b_n)$ - Hamming distances on all possible shifts between a and b.

Since the minimal length of a peptide sequence that can mimic the epitope recognized by antibody is usually in the range from 4 to 6 amino acids, we assigned similarity threshold equal 4. So any two peptides in a motif should have approximately 4 common amino acids (diameter of a motif). As well as no more than 3 shifts between peptides to the right or left sides were allowed.

The Algorithm 1 describes the CAST-based motif identification method (CMIM).

Algorithm 1 CAST-based motif identification (CMIM)

Input: Set of peptides P, similarity matrix D, threshold θ
Set of seed peptides $S \leftarrow P$
while $S > \emptyset$ **do**
  Cluster set $M \leftarrow \{s_1, s_2\},\ s_1, s_2$ - the two most similar peptides in S
  Set of petides outside the cluster $R \leftarrow P \setminus M$
  $affinity(p) \leftarrow D(p, s_1) + D(p, p_2),$ for all $p \in M \cup R$
  while $(affinity(r)/size(M) \geq \theta,\ r \in R)$ OR
      $(affinity(m)/(size(M) - 1) < \theta,\ m \in M)$ **do**
    while $affinity(r)/size(M) \geq \theta$ for some $r \in R$ **do**
      $M \leftarrow M \cup \{r'\},\ r' \in R$ - peptide with the highest affinity
      $affinity(p) \leftarrow affinity(p) + D(p, r'),$ for all $p \in C \cup R$
    end while
    while $affinity(m)/(size(M) - 1) < \theta$ for some $m \in M$ **do**
      $M \leftarrow C \setminus \{m'\},\ m' \in M$ - peptide with the lowest affinity
      $affinity(p) \leftarrow affinity(p) - D(p, m'),$ for all $p \in C \cup R$
    end while
    $S \leftarrow S \setminus M$
    Add M to set of clusters M
  end while
end while
for $M_1 \in$ M **do**
  for $M_2 \in$ M **do**
    if $(intersection(M_1, M_2)/size(M_1) > 0.5)$ OR
      OR $(intersection(M_1, M_2)/size(M_2) > 0.5)$ **then**
      Collapse M_1 and M_2
    end if
  end for
end for
for $M \in$ M **do**
  align peptides in M
  count entropy in every position i of aligned M
  find consensus K for 7-mer window with the min entropy
end for
Output: Set of motifs M, represented by clusters M_i and consensus sequences K_i

The input for the algorithm was a list of distinct peptides from a serum sample, threshold and similarity matrix which stored similarity values between any pair of peptides. On every iteration of the algorithm two peptides with the highest similarity were chosen as the initial center of a cluster. Next the process of adding and removing of peptides from the cluster was performed while the similarity between every pair of petides in a final set were not less than the threshold. During that step initially assigned central peptides could be removed. Obtained cluster was saved removing its peptides from further consideration as initial centers. Then the procedure was repeated to find remaining motifs. Unlike CAST our algorithm allows intersection between clusters. As result some consensus sequences of motifs could be too close to each other. So the obtained clusters

were collapsed if they had more than 50% common peptides. The last step was to align all peptides in the cluster and compute entropy in every position. Seven positions with the smallest cumulative entropy (the most conserved part) were chosen, and the consensus amino acid sequence was found. The output of the algorithm was a set of finding motifs in a serum sample, each represented by a cluster and its consensus 7-mer sequence.

4 Results

Data set. We analyzed the profiles generated for the 15 serum samples of the stage 0 and 1 breast cancer patients and for the 15 serum samples of the healthy donors. For each serum sample the experiment was performed separately using the same enriched library on all samples. In average, for the experimental condition selected, the total number of distinct peptide sequences generated in one sample was 18450, and standard deviation σ was 6205. The average count value (expression) of a sample was $407335(\sigma = 252393)$.

After applying the motifs search separately to every sample, we obtained in average 3000(1073) motifs per a control sample and 3490(1315) motifs per a case sample. The average size of a motif in a case was 7.1(1.8) peptides, in a control it was 6.8(1.3) peptides. Every sample contained significant amount of large motifs. Thus, the average number of motifs consisting of 20 and more peptides was 154(71) and 131(53) for cases and controls respectively.

Motif validation. To validate found motifs we generated pseudo mimotope profiles using two strategies. The first strategy was random permutation of amino acides in a sample peptides. As result, we received 30 samples consisting of random 7-mer peptides. We ran our motif search method on the samples and obtained about 6639(1967) motifs with the average size 4.2(0.7). Although, the largest motif among all samples contained only 17 peptides. More than 95% of motifs in all samples had size no more than 4 peptides.The obtained motifs were significantly different from those found in real serum samples. This result proves the amino-acid order is meaningful in mimotope motifs found by CMIM.

The second strategy was random selection of peptides from existing samples and generating random samples. We collapse all original serum samples together assigning count value to each peptide. The more abundant and popular a peptide was among samples the more probable it would be selected to a new random sample. We generated 30 samples with 20k peptides each. We also applied motif search method to the random samples. In average we obtained 3890(34) motifs with the size of 5.71(0.04) peptides. To compare the group of random samples with the group of real serum samples we applied Kruskal–Wallis test [3]. This non-parametric method determines whether samples originate from the same distribution. The result p-value was $7.5 * 10^{-5}$ rejecting the null hypothesis that the population medians of both groups were equal. Thus, the single sample motifs are significantly different from motifs in peptides drawn from multiple samples.

Cancer-specific motifs. The cancer-specific motifs were defined as motifs significantly prevalent in cases. We compared motifs based on their consensus 7-

probability	observed	expected	FDR
<0.05	67	51.9	0.77
<0.04	27	20.5	0.76
<0.03	24	16.6	0.69
<0.02	10	8.1	0.81
<0.01	4	4.2	1.06

Despite the fact that no motif is statistically significant after FDR adjusting to multiple testing, we can see that their number is still larger than expected.

5 Conclusions

In current work we identified cancer-specific motifs by analyzing peptide profiles of serum samples from cancer patients and from healthy donors. These profiles were generated using a phage DNA sequencing following single selection without amplification on the serum samples with the library enriched by the cycles of affinity selection-amplification using a pool of serum samples from additional cancer patients.

A novel motif identification method based on CAST clustering (CMIM) was proposed. We found that for any real serum sample the number of peptides per a motif is significantly greater comparing with pseudo epitope repertoire consisting of a randomly permuted peptides. Also the single sample motifs are shown to be significantly different from motifs in peptides drawn from multiple samples.

Running on case-control data CMIM identified cancer-specific motifs. Although no motif is statistically significant after adjusting to multiple testing, the number of found motifs is larger than expected and may therefore contain useful cancer markers.

References

1. Amir Ben-Dor, Ron Shamir, and Zohar Yakhini. Clustering gene expression patterns. *Journal of computational biology*, 6(3-4):281–297, 1999.
2. 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.
3. William H Kruskal and W Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621, 1952.
4. Alkes L Price, Eleazar Eskin, and Pavel A Pevzner. Whole-genome analysis of alu repeat elements reveals complex evolutionary history. *Genome research*, 14(11):2245–2252, 2004.
5. 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.