Research Article

In silico design and performance of peptide microarrays for breast cancer tumour auto-antibody testing

Parvez Syed[1], István Gyurján[1], Albert Kriegner[1], Klemens Vierlinger[1], Christian F Singer[2], Christine Rappaport-Fürhauser[2], Johannes Zerweck[3], Johannes Söllner[4] and Andreas Weinhäusel[1]

[1]Austrian Institute of Technology – AIT, Health & Environment, Molecular Diagnostics, Vienna, Austria
[2]Department of Obstetrics and Gynaecology, Medical University of Vienna, Vienna, Austria
[3]JPT Peptide Technologies GmbH, Berlin, Germany;
[4] Emergentec Biodevelopment GmbH, Vienna, Austria

Received on April 22, 2012; Accepted on June 2, 2012; Published on June 16, 2012

Correspondence should be addressed to Andreas Weinhäusel; Phone: +43 50550445, Fax: +43 505504450, E-mail: andreas.weinhaeusel@ait.ac.at

Abstract

The simplicity and potential of minimally invasive testing using sera from patients makes auto-antibody based biomarkers a very promising tool for use in cancer diagnostics. Protein microarrays have been used for the identification of such auto-antibody signatures. Because high throughput protein expression and purification is laborious, synthetic peptides might be a good alternative for microarray generation and multiplexed analyses.

In this study, we designed 1185 antigenic peptides, deduced from proteins expressed by 642 cDNA expression clones found to be sero-reactive in both breast tumour patients and controls. The sero-reactive proteins and the corresponding peptides were used for the production of protein and peptide microarrays. Serum samples from females with benign and malignant breast tumours and healthy control sera (*n=16* per group) were then analysed. Correct classification of the serum samples on peptide microarrays were 78% for discrimination of 'malignant versus healthy controls', 72% for 'benign versus malignant' and 94% for 'benign versus controls'. On protein arrays, correct classification for these contrasts was 69%, 59% and 59%, respectively.

The over-representation analysis of the classifiers derived from class prediction showed enrichment of genes associated with ribosomes, spliceosomes, endocytosis and the pentose phosphate pathway. Sequence analyses of the peptides with the highest sero-reactivity demonstrated enrichment of the zinc-finger domain. Peptides' sero-reactivities were found negatively correlated with hydrophobicity and positively correlated with positive charge, high inter-residue protein contact energies and a secondary structure propensity bias. This study hints at the possibility of using *in silico* designed antigenic peptide microarrays as an alternative to protein microarrays for the improvement of tumour auto-antibody based diagnostics.

Introduction

Breast cancer is the leading tumour type in women, with an estimated 1 million new cases worldwide each year (Pisani *et al.* 2002; Sturgeon *et al.* 2008). An increased survival rate is highly correlated with an early detection of malignancy, making diagnostics a critical tool in cancer prevention. Over the past decades several diagnostic tools have been developed and used in screening programmes, such as mammography, ultrasound imaging and magnetic resonance imaging (MRI) (Piura & Piura 2011). These are all able to detect probable malignancies; however, definitive answers still require biopsy and histopathological examination.

Blood-based biomarker discovery is an emerging field of cancer research which seeks to identify specific and sensitive markers, enabling clinicians to make decisions with great accuracy and reliability.

Detection of tumour-associated auto-antibodies from a few drops of blood may provide a possibility to screen patients with the suspicion of breast cancer or even before, through periodical examination. Tumour-associated antibodies can be identified through selective binding to special antigens, called 'tumour-associated antigens' (TAAs). TAAs derived from aberrantly expressed proteins during the onset and progression of cancer development, display 'non-self' epitopes which trigger the immune system to remove them. The observed antigenicity has been attributed to multiple features of cancer growth, including accumulated mutations in cancer cells (*e.g.* point mutations, translocations), overexpression and translation of 'differentiation genes' or improper post-translational modification (Backes *et al.* 2011). These molecules usually possess important functions in tumourigenesis, such as regulation of the cell cycle, cell proliferation and apoptosis (Ullah & Aatif 2009). Previous studies have already elucidated several TAAs from the sera of breast cancer patients, such as MUC1, HSP90, HER2/neu, c-myc, NY-ESO1/LAGE1 and Lipophilin B (Carter *et al.* 2003, Chapman *et al.* 2007, Conroy *et al.* 1995, Disis *et al.* 1994). Auto-antibodies against p53 tumour suppressor proteins were also detected in the sera of 9-26% of women with breast cancer (Montenarh 2000). However, it has been shown that through assaying of sera, reactivity for a single TAA is neither sensitive nor specific enough to discriminate between healthy individuals and cancer patients. Thus a combination of multiple TAAs would be preferred to generate a diagnostic classification tool.

Several methods have been developed to identify, screen and validate discriminative TAAs. SEREX (Serological Analysis of Recombinant Expressed cDNAs) and SERPA (Serological Proteomics Analysis) are such methods, employed to identify *de novo* TAAs directly from tumour cells (Lu *et al.* 2008). Although these methods have been used successfully to uncover new antigens (Hamrita *et al.* 2008, Qian *et al.* 2005, Stempfer *et al.* 2010), the drawback of these technologies is that they are labour intensive and only applicable on a small scale. Higher throughput methods such as protein macro- and microarrays allow for simultaneous quantification of serum reactivity of thousands of proteins. One of the major challenges of these applications is the requirement of a huge number of in-frame cDNA clones and the subsequent expression and purification of the cognate proteins from them. The physicochemical properties (e.g. length vs. hydrophobic domains) of expressed proteins are usually highly variable and displaying the associated reactive epitopes upon immobilisation can be hardly controlled.

Peptide microarrays represent another alternative solution as shorter peptide sequences may recapitulate the biological function (*i.e.* the antigenic epitope) of the corresponding protein (Cretich *et al.* 2006; Uttamchandani & Yao 2008). Production of synthetic peptides is a well established technique and using peptide arrays as a potential alternative to protein arrays would have several advantages. The concept of the peptide array was first proposed by Southern in 1988 (Southern 1988). Techniques like photolithographic peptide synthesis on a glass surface (Fodor *et al.* 1991) and the SPOT-synthesis technology (Frank 2002) have accelerated the applications of synthetic peptides in microarray experiments (Shin *et al.* 2005).

In this report we evaluate the performance of a SPOT-synthesized peptide microarray. This technology utilizes the traditional *fmoc* chemistry to synthesize peptides in single droplets immobilized on the surface of slides. Based on a semi-empirical method developed by Kolaskar and Tongaonkar (Kolaskar & Tongaonkar 1990), we deduced antigenic peptides from a set of previously identified, protein microarray-derived, antigenic proteins. We probed these peptides with sera from breast cancer patients and individuals with benign breast nodules, whilst compared them with samples from healthy donors. We further evaluated the identified sero-reactive peptides using bioinformatics tools and defined panels of TAAs, which are able to discriminate between samples of healthy control, malignant and benign tumours.

Materials and Methods

Serum Samples

Serum samples were obtained after the consent of the breast cancer patients and healthy female volunteers. The samples were then stored at -80°C. The study was approved by the Ethics Committee of the Medical University of Vienna, the General Hospital of Vienna (study number: 143/2007) and all procedures were carried out in compliance with the Helsinki Declaration. For the protein and peptide microarray analysis of breast cancer serum biomarkers, 48 serum samples (malignant n=16; benign n=16; healthy n=16) were used. The clinical and the pathological cohorts of the serum samples are described in Table 1. All the 16 malignant samples were collected from patients diagnosed with invasive ductal carcinoma and tested positive to HER2/neu. Furthermore, the benign samples were collected from patients diagnosed with fibroadenoma. Healthy control serum samples (n=16, mean age 76.9±7.15), were collected from healthy volunteers who presented no personal or familial history of breast or ovarian cancer.

Table 1. Clinical and pathological data of the patient-study cohort. Benign and maligant samples were collected from patients with fibroadenomas and invasive ductal carcinoma, respectively.

	Benign (n=16)	**Malignant (n=16)**
Age (years)[a]	52.5±4.9	53.75±8
Grading[b]		
G1	-	1
G2	-	5
G3	-	10
Oestrogen receptor positive	-	9
pT stage[c]		
pT1; pT1b; pT1c; pT1mic; pT2	-	3; 3; 7; 1; 2
pN stage[d]		
pN0; pN1; pN1a; pN2; pN2a; pN3	-	7; 1; 1; 1; 3; 2
Metastasis stage[e]**: M0**	-	6
Menopause status[f]		
Pre-menopause	3	4
Post-menopause	8	11

[a]The age of the patients is represented as a mean (age±standard deviation).
[b]G1 (low-grade), G2 (intermediate grade) and G3 (high-grade). Low-grade tumours are usually slow growing and are less likely to spread. High-grade tumours are likely to grow more quickly and are more likely to spread.
[c]pT1: Tumour 2.0 cm or less in dimension; pT1b: Tumour between 0.5 and 1 cm in dimension; pT1c: Tumour between 1.0 and 2.0 cm in dimension; pT1mic: Microinvasion 0.1 cm or less in dimension; pT2: Tumour between 2.0 and 5.0 cm in dimension.
[d]pN stage: information available for 15 patients. pN0: No regional lymph node metastasis; pN1: Metastasis to movable ipsilateral axillary lymph node(s); pN1a: Only micrometastasis (none larger than 0.2 cm); pN2: Metastasis to ipsilateral axillary lymph node(s) fixed to each other or to other structures; pN2a. Metastasis in 4-9 axillary lymph nodes, including at least one that is larger than 2 mm; pN3: Metastasis to ipsilateral internal mammary lymph node(s).
[e]Metastasis stage: information from 6 patients. M0: No distant metastasis.
[f]Information from 11 benign and 15 malignant samples.

Protein extraction and purification

In an earlier study, 642 clones were identified from a collection of 38,016 cDNA expression *E. coli* clones (hEx1 library (Bussow *et al.* 2000)), which reacted positively to the sera from the breast cancer patients and the healthy control individuals. For the recombinant protein expression in *E. coli* and protein purification, the procedure developed by Stempfer *et al.* was followed (Stempfer *et al.* 2010). In brief, the cDNA expression clones were cultured in 96 deep well plates and were induced by an autoinduction strategy for recombinant protein production. The expressed His-tagged proteins were then purified using Ni-NTA agarose and eluted in microarray spotting buffer (50 mM KH_2PO_4 and 50 mM K_2HPO_4, pH 8.0, 500 mM imidazole, 0.01% SDS and 0.01% NaN_3).

Design of Antigenic Peptides

Peptides corresponding to the 642 reactive proteins were designed as an alternative to the recombinant proteins found reactive in the initial membrane screening. To predict the antigenic peptides, the EMBOSS tool "Antigenic" (http://liv.bmc.uu.se/cgi-bin/emboss/antigenic) was used. The minimum length of the predicted peptide sequences is 6 amino acids (aa). The "Antigenic" tool employs a semi-empirical method developed by Kolaskar and Tongaonkar for the selection of antigenic peptide sequences. This method uses the physicochemical properties of amino acid residues and their frequencies of occurrence in experimentally known segmental epitopes to predict antigenic determinants on proteins (Kolaskar & Tongaonkar 1990).

The DNA sequence was available for 596 of the 642 clones. Of those, 581 clones were unique and used for antigenic peptide prediction. The default settings of the "Antigenic" tool were used, and for each unique clone sequence, 2-3 different peptides were selected based on antigenicity score and peptide-length. In trying to achieve uniform synthesis, peptides sized 8-10 aa were selected. Based on the maximum antigenicity score, antigenic peptides which were longer than 10 aa were shortened. For antigenic motifs shorter than 8 aa peptides, N terminal aa's corresponding to the template sequence were added. In addition, tetanus specific antigenic peptides were designed for the NCBI reference sequence NP_783831; 56 tetanus specific peptides were selected from all potential antigenic peptides based on their maximum antigenicity score. Furthermore, peptides of 10 aa in length were selected for synthesis as described above.

In order to find over-represented motifs in the peptide set, sequences were submitted to MEME motif search web-based tool (http://meme.nbcr.net). The motif was considered as 'enriched' if it had at least 5 sequences (sites) with an E-value less than 0.001. Motif searching was also performed on peptide sequences with high sero-reactivity (defined as median log2 intensities >13 of all 48 samples analysed; min.: 6.21; max.: 15.84).

Microarray production

The procedure for the protein microarray production has been described in our previous study (Stempfer *et al.* 2010). In brief, the protein microarrays were generated using the purified recombinant proteins obtained from the cDNA expressing *E. coli* clones. These purified proteins were spotted using an Omnigrid arrayer (GeneMachines, San Carlos, CA) with SMP 3 pins (TeleChem International Inc., Sunnyvale, CA) under adjusted air humidity; between 55% and 60%. Spots were printed in duplicates on ARChip Epoxy slides (Preininger *et al.* 2004) and each microarray contained 4 identical subarrays. The crude protein extract of the *E.coli* host was used for positive control spots, and plain buffer spots were used as negative controls.

For the generation of peptide microarrays, 1212 clone-specific and 56 tetanus specific short peptides were synthesized using SPOT synthesis technology (JPT Peptide Technologies GmbH, Berlin, Germany). Aminooxy-acetylated peptides were synthesized in parallel on cellulose membranes. Once the de-protection of the side chain was achieved, the solid phase-bound peptides were transferred to 96 well microtitre filtration plates (Millipore, Bedford, USA). These peptides were cleaved from the cellulose membranes using 200 ml of aqueous triethylamine (0.5% v/v). The triethylamine-peptide solution was filtered and evaporated under reduced pressure to remove the solvent. This was followed by re-dissolving the resultant peptide derivatives (50 nmol) in 25 mL of spot buffer (70% DMSO, 25% 0.2 M sodium acetate pH 4.5, 5% v/v glycerol). The re-dissolved peptide solution was then transferred into 384 well microtitre plates and used for the generation of the peptide arrays. Two droplets of 0.5 nL peptide solution (1 mM) were immobilized in triplicates on ARChip Epoxy slides (Preininger *et al.* 2004), containing 4 identical sub-arrays on each slide. For the immobilization of the peptide solution, a non-contact printer Nanoplotter (GESIM, Groberkmannsdorf, Germany) fitted with a piezoelectric NanoTip (GESIM) was used. Apart from the peptides derived from the cDNA clone-proteins, human Immunoglobulins (Igs) (IgA, IgE, IgG and IgM) and 56 tetanus toxin (TT) specific peptides were also immobilized on the peptide microarrays. The human Igs and TT specific peptides were used as positive controls, while the empty buffer spots were used as negative controls.

Microarray processing

The microarrays were blocked with DIG easy Hyb (Roche Applied Science, Vienna, Austria) for 30 min and then washed twice in Phosphate Buffered Saline with 0.1% Tween 20 (PBST) for 5 min. Breast cancer serum samples (benign; n=16 and malignant; n=16) and control sera (n=16) diluted in a 1:10 ratio with PBST were applied onto the microarrays and incubated for 2 hours. The microarrays were then washed twice in PBST for 5 min. This was followed by incubation for 30 min with goat anti human IgG detection antibody, fluorescently labelled with Alexa647 dye (Invitrogen, Vienna, Austria), diluted 1:500 in PBST+3% non-fat dry milk powder. Later, the microarrays were washed twice in PBST for 5 min. The array images of the processed slides were then captured using an Axon Genepix 4000A microarray scanner (Molecular Devices, Union City, CA).

Data analysis

Fluorescent intensity values (median after subtraction of the local background) were calculated from the scanned images using the Genepix software (Molecular Devices). Statistical analysis of the microarray experiments was performed using the BRB-ArrayTools software 3.8.1 [http://linus.nci.nih.gov/BRB-ArrayTools.html] developed by Dr. R Simon and Amy Peng Lam (Simon *et al.* 2007). The $\log_2$-transformed values of the signal intensities obtained from the scanned images of the processed microarrays were used for the analysis. The peptide microarray data were normalized using the "house-keeping gene" normalisation option within BRB-ArrayTools using the "Tetanus peptides" and "Igs" spots as normalisation features. For the data from protein microarrays, a global normalization was used to normalize each array using the relative median over all the log intensity values within one experiment. To identify the proteins/peptides expressed differentially between classes, a random-variance t-test was applied to the data sets (Wright & Simon 2003). Significance of differentially expressed proteins/peptides were then ranked using the p-value (0.05 and 0.01 for protein and peptide microarray data, respectively) from the univariate test. Further statistical data analysis was performed using R version 2.6.2 (R Development Core Team 2005).

For defining a classifier set of antigenic proteins and peptides, the class prediction tools implemented in BRB-ArrayTools were used and leave-one-out cross validation (LOOCV) was conducted. Different classification algorithms (compound covariate, k nearest neighbour (k=1 and k=3), nearest centroid, support vector machines, diagonal linear discriminant analyses and Bayesian compound covariate prediction) were run for model generation. The model incorporated the features that were differentially expressed among all the microarray-features at the 0.01 and 0.05 significance level as assessed by the random variance t -test, respectively (Wright & Simon 2003). We esti-

mated the prediction error for each model using LOOCV, as described by Simon and colleagues (Simon *et al.* 2003). For each LOOCV training set, the entire model building process was repeated, including the peptide and protein selection process. We also evaluated whether the cross-validated error rate estimate for a model was significantly less than one would expect from random prediction. The class labels were randomly permuted and the entire LOOCV process was repeated. The significance level is the proportion of the random permutations that gave a cross-validated error rate no greater than the cross-validated error rate obtained with the real data. Cross-Validation receiver operating curve (ROC) analyses from the Bayesian Compound Covariate Predictor were conducted and the 'area under the curve' (AUC) values were calculated as implemented in 'BRB-ArrayTools' class prediction tools.

Over-representation analysis

An over-representation analysis (ORA) of the classifiers derived from the microarray experiments was performed using the gene set enrichment analysis tool "GeneTrail" (Keller *et al.* 2008). The classifiers from the peptide microarray analysis were traced back to the proteins they were derived from and the ORA was performed using the corresponding gene Ids. Similarly, the classifiers from the protein array analysis were used for the ORA. For ORA, a reference set was compared to the test sets (genes corresponding to the classifiers). All annotated human genes (NCBI GeneIDs) were used as reference set and a 'hypergeometric distribution test' was performed for computing P-values. The significance value of 0.05 (Benjamini and Hochberg adjusted) was chosen.

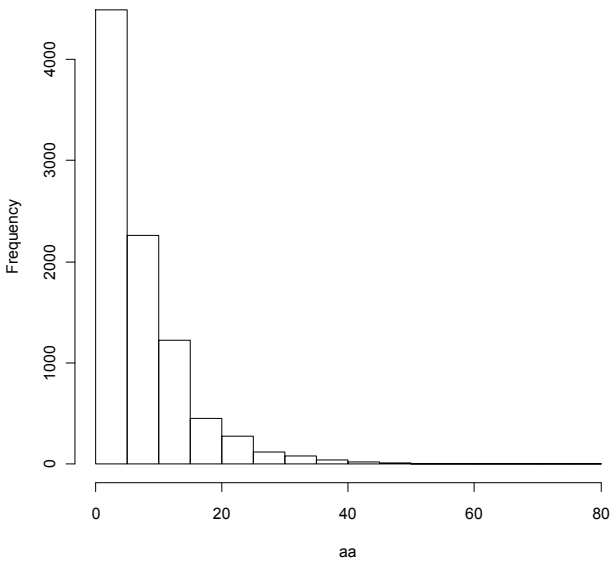

Figure 1. Length distribution of 4492 peptides. The figure shows the frequency of occurrence (Y-axis) of the peptides with regards to the length of the antigenic motif (X-axis). A relatively high frequency of occurrence was observed for the short-length peptides.

Results

Antigenic motif search

Out of 642 clone-proteins, which were used for the protein microarray production, sequences of 596 proteins were available. All 3 possible reading frames of DNA sequences coding for proteins were collected and checked for the longest uninterrupted ORF sequence. After eliminating the duplicates, we found 581 unique sequences which were used for antigenic motif search. Using the "Antigenic" tool we obtained 4492 antigenic peptides for these 581 clone-sequences, resembling an average of 7.73 peptides per clone. When the length of the 4492 antigenic peptides were plotted against frequency of occurrence, a high frequency of occurrence was observed with peptides of length ranging from 6 to 20 amino acids (Figure 1) and also, a uniform distribution of antigenic motifs were found along the 581 clone-sequences subjected to peptide design (Figure 2).

Of the 4492 antigenic motifs 2866 were unique motifs. From the latter, 2-3 peptides per clone

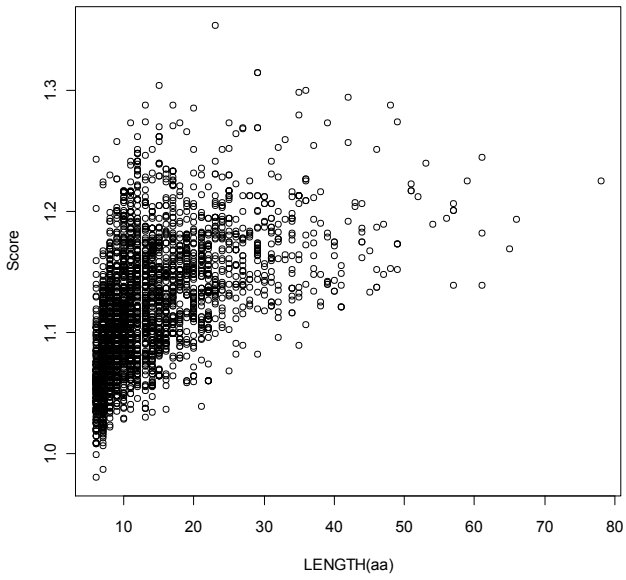

Figure 2. Distribution of 4492 antigenic motifs along the 581 clone sequences. The x-axis depicts the start amino acid position within the targeted clone-sequences; on the y-axis, the length of the antigenic-peptides is depicted. A uniform distribution of the antigenic motifs was observed along the clone sequences. Density of plotted antigenic motifs is highest for short peptides (<20mers; y-axis).

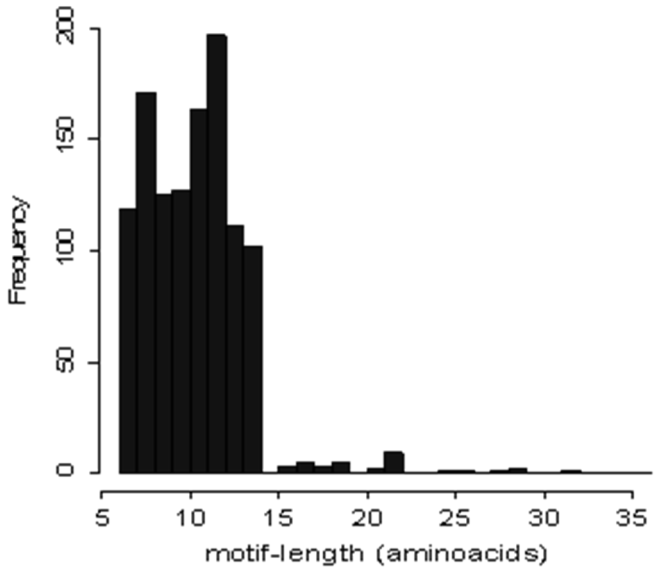

Figure 3. Length distribution of 1185 single peptides. The frequency of occurrence (Y-axis) of the peptides with regards to the length of the antigenic motif (X-axis) is shown.

were selected which had maximum scores (1.3 to 0.987) and thus identified 1212 peptides. Out of these 1212 peptides, 53%, 33.7% and 13.2% of the peptides were 8-10, 7-14 and more than 14 aa long, respectively (Table 2). Peptides with lengths ranging 6 to14 were present at highest frequency compared to the longer peptides (Figure 3). These 1185 single peptides, including the human Igs and the 56 TT specific peptides were used for the peptide array production.

Serum reactivity of 'antigenic' peptide arrays

Median intensities of each duplicate peptide spot from the 48 microarray analyses were calculated and used to evaluate the correlation of serum-reactivity towards the 'antigenicity score' and the influence of aa addition or removal from the antigenic motif (peptide lengths were adjusted in order to synthesize and spot 8-10 aa peptides; see Methods). We could not find any correlation of microarray signal intensities with 'length adjustment of peptides' and 'antigenicity scores' (Figure 4).

Length (aa)	Number of peptides	Number of clones
8-10	643	329
7-14	409	79
>14	160	55

Table 2. Number of peptides with regards to the length of the antigenic motifs and the number of the corresponding clones.

Furthermore peptides were converted into numerical representations and subjected to linear correlation analysis with median and maximum intensities. Overall 3697 amino acid and sequence parameters were used for this alternative representation. Of these, the majority was derived from AAINDEX (http://www.genome.jp/aaindex/), augmented by a few feature descriptors commonly used in QSAR analysis and basic amino acid statistics, including simplified alphabets (Söllner 2006). B-cell antigenicity was estimated using a previously presented regression model and sequence entropy using a composition biased method (Sollner *et al.* 2008). Susceptibility to proteasomal processing was assessed using netChop (Keşmir *et al.* 2002). Affinity for 43 MHC alleles was predicted using netMHC (Lundegaard *et al.* 2008) and highest affinities mapped to respective supertypes. For all single

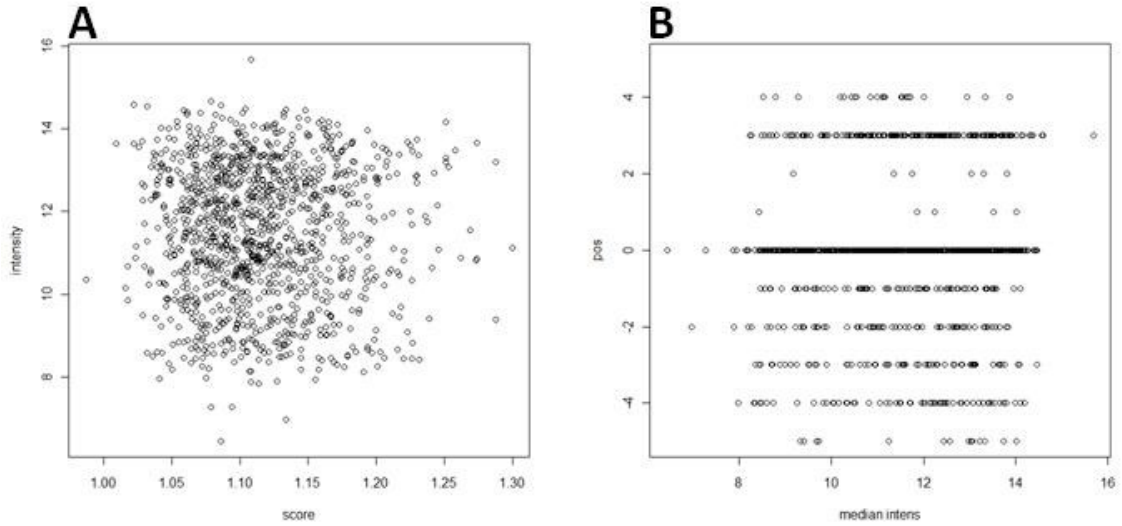

Figure 4. Antigenic reactivity derived from 48 samples. Median peptide array intensities (log2 transformed) were plotted versus the 'antigenicity score' (A), and the 'length adjustment' (denoted "pos"). Positive values correspond to the number of aa added, negative values to aa removed from the antigenic motifs for the generation of 8-10aa peptides for array spotting (B).

aa parameters, averages over the entire peptide, N- and C-terminal residues were computed.

The features with the highest positive or negative correlation (in the order of -0.5 and +0.45, respectively) all originate among physico-chemical properties, in particular hydrophobicity, inter-residue contact energy, secondary structure and charge related properties. A scatter plot of the maximally correlated feature, a hydrophobicity scale by Wilce *et al.* (Wilce *et al.* 1995), is shown in Figure 5. The scatter plot clearly supports a linear dependency and a selection of other substantially correlated peptide properties is listed in Table 3. Measured median intensities are particularly negatively correlated with hydrophobicity and positively correlated with both positive charge and high inter-residue protein contact energies. The existence of a possible secondary structure propensity bias is also apparent.

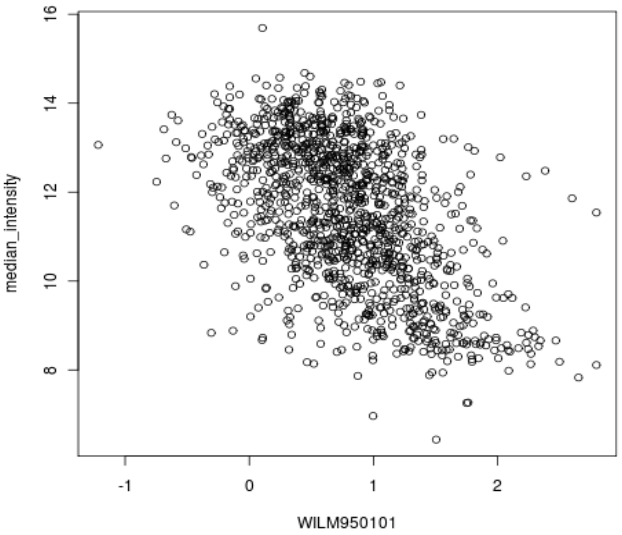

Figure 5. Scatter plot of the peptides parameterized using the hydrophobicity scale by Wilce *et al.* (WILM950101 on x-axis) versus the median peptide array intensities (log2 transformed on y-axis).

Motif enrichment analysis

Motif enrichment analysis was performed on our microarray peptide set (1185 peptides) using the MEME motif discovery tool (Bailey & Elkan, 1994; http://meme.sdsc.edu/meme/). The most significant and highly represented motif found was similar to Zn-finger domains of Zn-H2C2-type (Fig. 6A, see pfam13465: zf-H2C2_2). The motif logo consisted of 26 sequences and the diagram clearly depicts the highly weighted two central cysteines, separated by two other amino acids. Seemingly the first two amino acids (proline and tyrosine) also have a conserved role to constitute these domains (Figure 6A). In a second screen, only those peptides were considered in the analysis that gave high intensity values (median log2 >13). The analysis of highly reactive peptides elucidated similar results: the only significantly enriched motif was again the previously identified Zn-finger

Table 3. Physico-chemical parameters maximally correlated with the median intensity derived from the peptide arrays processed with 48 serum samples. Correlation coefficients scale between -1 and 1, indicating negative and positive correlation, respectively.

Parameter	Correlation to median intensity	Type	URL
WILM950101	-0.53	negative	http://www.genome.jp/dbget-bin/www_bget?aaindex:WILM950101
COWR900101	-0.53	negative	http://www.genome.jp/dbget-bin/www_bget?aaindex:COWR900101
GUOD860101	-0.52	negative	http://www.genome.jp/dbget-bin/www_bget?aaindex:GUOD860101
NAKH920108	-0.52	negative	http://www.genome.jp/dbget-bin/www_bget?aaindex:NAKH920108
JURD980101	-0.51	negative	http://www.genome.jp/dbget-bin/www_bget?aaindex:JURD980101
MONM990101	0.48	positive	http://www.genome.jp/dbget-bin/www_bget?aaindex:MONM990101
FAUJ880111	0.46	positive	http://www.genome.jp/dbget-bin/www_bget?aaindex:FAUJ880111
CHOP780207	0.46	positive	http://www.genome.jp/dbget-bin/www_bget?aaindex:CHOP780207
MIYS990102	0.45	positive	http://www.genome.jp/dbget-bin/www_bget?aaindex:MIYS990102
MIYS990101	0.45	positive	http://www.genome.jp/dbget-bin/www_bget?aaindex:MIYS990101

(Figure 6B). Since only 385 sequences were used in this analysis these motifs were "shorter" (8 amino acids) and again clearly depicted the Zn-finger domain characteristics for the superfamily. The finding that Zn -finger domains presented in our peptide array analyses as highly sero-reactive, confirms their antigenicity. These findings are concordant with previous reports, which found several members of Zn-finger proteins as tumour-associated antigens (Ludwig *et al.* 2012).

Microarray analysis

The data obtained upon processing the protein and the peptide microarrays with the breast cancer (n=16), benign fibroadenomas (n=16) and healthy control (n=16) sera was subjected to statistical evaluation. The class prediction of the samples was performed using BRB-ArrayTools and the performance of algorithms with the highest correct classifications is depicted.

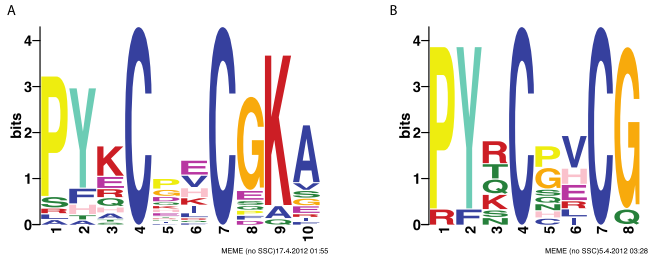

Figure 6. Sequence logos of enriched motifs. (A) Sequence logo depicting the most significant motif (E-value: 1.0-e100, 26 sites). (B) Analysis of peptides with high experimental signal intensity (median log2>13) giving very similar results (E-value: 9.8e-23, 11 sites). MEME sequence logos represent probability matrices that specify the probability of each letter in all possible positions.

We elucidated a marker-set of 54 peptides (Table 1S; see supplementary data) which enabled 78% correct classification using the "compound covariate classifier" of malignant samples and healthy controls with 75% sensitivity and 81.2% specificity (Table 4). The ROC curve derived from this class prediction (Figure 7A) demonstrated AUC values of 0.758. For the prediction of the same classes on protein array, a marker-set of 57 proteins was deduced (Table 2S; see supplementary data). These proteins enabled the 69% correct classification of the malignant samples and healthy controls with 62.5% sensitivity, 75% specificity (Table 4) and an AUC value of 0.68 (support vector machine classifier) (Figure 7B).

For class prediction of the benign and malignant samples on peptide arrays, we elucidated 9 peptides (Table 3S; see supplementary data) which enabled correct classification of 72% (3-Nearest Neighbours classifier) with 62.5% sensitivity and 81.2% specificity (Table 4) An AUC value of 0.6 was observed for this classification (Figure 7D). Similarly on the protein array, 17 proteins (Table 4S; see supplementary data) enabled 59% correct classification (1-Nearest neighbour) of benign and malignant samples with 87.5% sensitivity, 31.2% specificity (Table 4) and an AUC value of 0.461 (Fig. 7E) (Table 4).

The class prediction between the benign and the control samples yielded 17 peptides (Table 5S; see supplementary data) which gave 93.8% sensitivity and specificity with 94% correct classification (1-Nearest Neighbour classifier method) (Table 4). The observed AUC value of the ROC curve for this class prediction

Table 4. Class predication of benign, malignant and control samples using peptide and protein arrays.

Classes	Microarray	Classification method	Correctly classified (%)	Sensitivity (%)	Specificity (%)	PPV	NPV	AUC
Malignant vs. Control	Peptide	Compound covariate predictor	78	75	81.2	0.8	0.765	0.758
	Protein	Support vector machine	69	62.5	75	0.714	0.667	0.68
Benign vs. Malignant	Peptide	3-Nearest neighbours	72	62.5	81.2	0.769	0.684	0.6
	Protein	1-Nearest neighbour	59	87.5	31.2	0.56	0.714	0.461
Benign vs. Control	Peptide	1-Nearest neighbour	94	93.8	93.8	0.938	0.938	0.852
	Protein	Compound covariate predictor	59	62.5	56.2	0.588	0.6	0.648

PPV: positve predictive value. NPV: negative predictive value. PPV and NPV correspond to the proportion of samples with positive and negative test results, respectively, which are correctly diagnosed.

was 0.852 (Figure 7F). From the protein microarray data, we elucidated a panel of 35 proteins (Table 6S; see supplementary data) which enabled 59% correct classification (compound covariate classifier) of benign and control samples; sensitivity was 62.5%, specificity was 56.2% (Table 4) and the corresponding AUC value was 0.648 (Figure 7G). Upon comparing the classifiers derived from all the class predictions performed on peptide and protein arrays, we identified 9 overlapping proteins corresponding to the genes:, *PCSK1, DGKK, ZNF598, TBC1D9, TMEM199, EPB41L3, SAMD6, PRPF38A* and *C1orf9.*

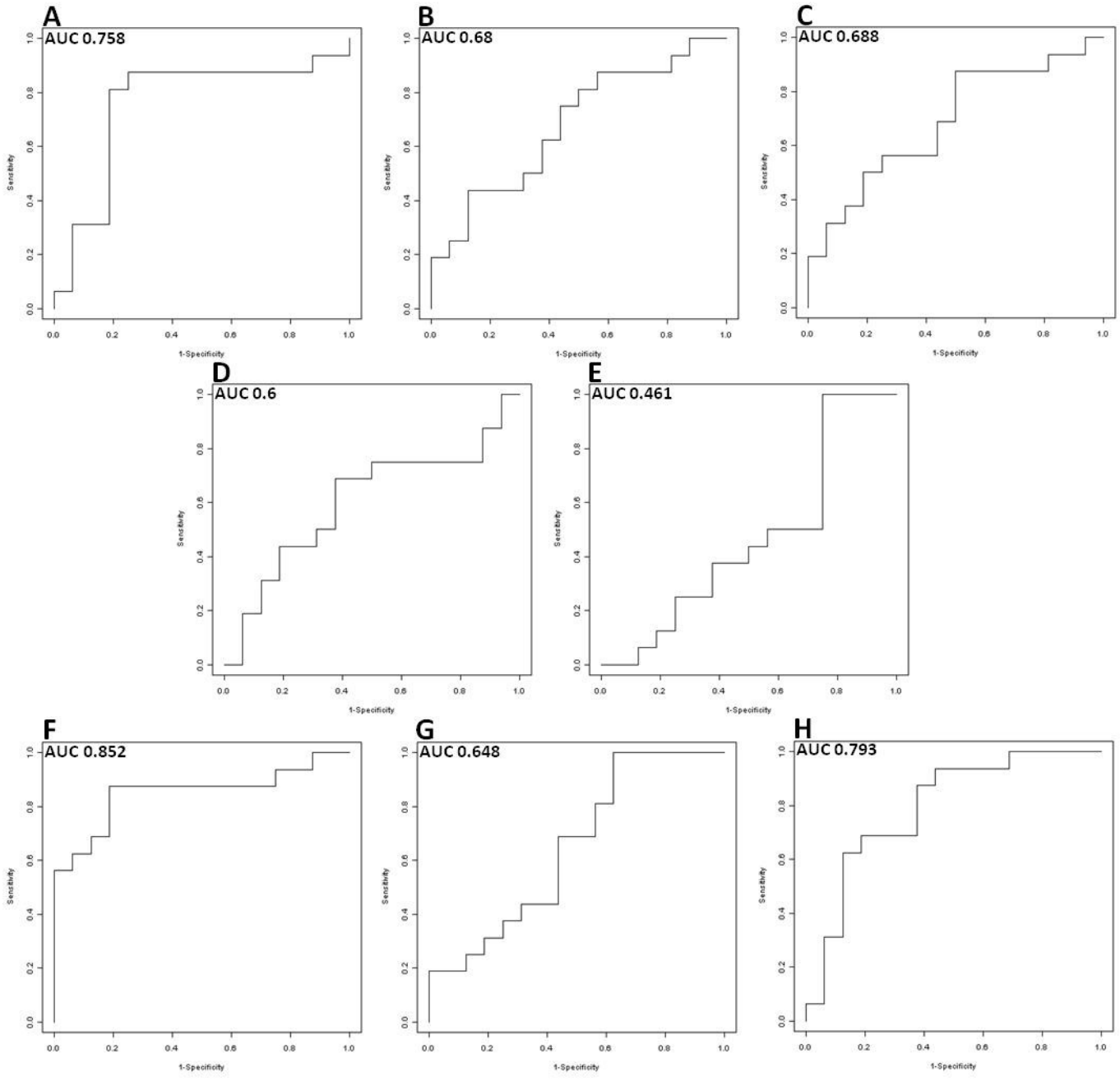

Figure 7. Cross-Validation ROC curves from the Bayesian Compound Covariate Predictor. The figures A, D and F represent the ROC curves obtained from the class predictions performed using the data from the peptide arrays. The figures B, E and G represent the ROC curves from the class predictions obtained from the protein microarray analysis. The figures C and H represent the ROC curves from the class predictions obtained from the protein array analyzed using the proteins corresponding to the respective peptide array classifiers. The x-axes and y-axes represent the false positive rate (1-specificity) and true positive rate (sensitivity), respectively. The ROC curves A, B and C represent the class prediction of malignant and control samples. The class predictions of benign and malignant samples are represented by the ROC curves D and E. Similarly, the ROC curves F, G and H represent the class prediction of the benign and control samples.

Using the clone-proteins (recombinantly expressed proteins from the cDNA expression clones) corresponding to the classifier peptides derived from the class prediction of malignant and control samples, a set of 4 proteins (Table 7S; see supplementary data) were deduced which enabled a correct classification of 66% (compound covariate classifier), with 56.2% sensitivity, 75% specificity and an AUC value 0.688 (Figure 7C). The class prediction of malignant and benign samples was not possible using the clone proteins corresponding to the classifier peptides. The clone proteins corresponding to the classifier from the class prediction of benign and the control samples enabled a 72% correct classification (compound covariate classifier) of the same classes with 2 clone proteins (Table 8S; see supplementary data). For this class prediction, 68.8% sensitivity, 75% specificity and an AUC value of 0.793 (Figure 7H) were observed.

Higher percentages of the correct classification were observed on peptide arrays compared to the same contrasts on protein microarray data (Table 4). For example, when malignant samples were compared to healthy controls, peptide arrays gave 9% more correct classification compared to protein arrays. Similarly, peptide arrays gave a 13% and 35% increasingly correct classification for the contrasts between 'benign vs malignant' and 'benign vs controls', respectively, when compared to the protein microarrays.

The signal intensities from all the peptide array classifiers were compared with the ones from the corresponding proteins on the protein arrays. Similarly, the signal intensities from all the protein array classifiers were compared to the ones from the corresponding peptides. These comparisons failed to give any correlation between the peptide and protein array data (Figures 1S and 2S; see supplemental file).

Table 5. Over-represented genes from the peptide array classifier and the corresponding KEGG pathways

KEGG pathways	p-value	Expected number of genes	Observed number of genes	Genes
Pentose phosphate pathway	0.003	0.06	2	*PGLS, ALDOA*
Ribosome	0.003	0.2	3	*RPL7A, RPL24, RPL6*
Spliceosome	0.04	0.3	2	*ISY1, PRPF38A*

Over-representation analysis

Peptide array classifiers

Over-representation analysis of the genes encoding the classifier peptides from all class predictions on peptide arrays was performed using the gene set analysis tool "GeneTrail" (Keller *et al.* 2008). Out of 57 genes representing the peptide classifiers, 3 genes, namely *RPL7A, RPL24* and *RPL6* were involved in the KEGG ribosome pathway. *PGLS* and *ALDOA* were found to be involved in the pentose phosphate pathway, while *ISY1* and *PRPF38A* were involved in the spliceosome pathway (Table 5). Out of the 57 genes, 7 genes contain a Zn-finger domain and among these 7 genes, 4 genes contain Zn-finger domains of the Zn-H2C2-type.

Protein array classifiers

Similarly, ORA was performed using the genes encoding the classifiers from all the class predictions on the protein array. 2 genes (*RPS3A* and *RPS13*) out of a total of 59 representing the protein classifiers were found to be involved in the KEGG ribosome pathway. *GIT1, CHMP4C, EHD2* and *GRK1* were involved in the KEGG endocytosis pathway (Table 6). We found that 28% and 32% of the genes represented by the classifier proteins contained sequence motifs such as coiled coils and ELR motifs at p-values equal to 0.0004 and 0.003, respectively. An enrichment of the protein family domains such as the UBA/TS-N domain (ubiquitin associated domain found on the N terminus of EF-TS (elongation factor thermo stable)) and the TBC domain was also observed.

Table 6. Over-represented genes from the protein array classifier and the corresponding KEGG pathways

KEGG pathways	p-value	Expected number of genes	Observed number of genes	Genes
Endocytosis	0.006	0.6	4	*GIT1, CHMP4C, EHD2, GRK1*
Ribosome	0.04	0.2	2	*RPS3A, RPS13*

Discussion

Peptide microarrays displaying synthetic peptides can be used for the detection of antibodies in serum apart from their utility in epitope-mapping, substrate profiling and probing peptide-ligand interactions (Andresen *et al*. 2006; Uttamchandani & Yao 2008). In the context of serodiagnostics, peptide arrays have been used for the detection of Hepatitis B and C viruses, human immunodeficiency virus (HIV), Epstein-Barr virus and syphilis (Duburcq *et al.* 2004). Li *et al*. (Li *et al.*

2010) used peptide arrays with an extracellular domain of epidermal growth factor receptor (EGFR) protein and detected auto-antibodies against an EGFR domain in the sera of non-small cell lung cancer patients. In this study, we designed peptides representing sero-reactive antigenic proteins using the antigenic motif search tool. We then used peptide microarrays to identify the auto-antibody signatures against these peptides in the sera of patients with breast cancer and benign fibroadenomas as well as healthy females.

The prediction of the antigenic peptides was performed based on the occurrence of hydrophobic residues (cysteine, leucine and valine) in a given protein sequence (Kolaskar & Tongaonkar 1990). This prediction method predicts the antigenic sites with approximately 75% accuracy. Using the Antigenic tool (Selak *et al.* 2003) we identified antigenic sites within the sequence of the early endosome antigen 1 (EEA1) protein with antigenic scores ranging from 1.135 to 1.09. In our study, we used the predicted antigenic motifs with the antigenic scores ranging from 1.3 to 0.987. The short peptides, with lengths ranging from 4 to 15 aa, are effective enough to identify antibody epitopes (Reineke & Sabat 2009). In this study, 86.7% of the predicted antigenic peptides had varying lengths from 7 to 14 amino acids. These peptide sequences were used for deducing the individual peptides and for the generation of peptide arrays. Peptides with the highest sero-reactivity of all the 48 samples showed enrichment of motifs similar to Zn-finger domains, which can be explained due to the central cysteine being highly hydrophobic. Moreover, many Zn-finger proteins contain variable numbers of Zn-finger domains (Iuchi 2001). These features allowed the Antigenic tool to label many of these peptides as antigenic.

We have tested another method introduced by Wilce and colleagues (1995) to see the correlation between experimental signal intensities and physic-chemical properties. The physicochemical parameters maximally correlated with the median intensity derived from the peptide arrays. The peptides' reactivities were found particularly negatively correlated with hydrophobicity while conversely they were positively correlated with positive charge, high inter-residue protein contact energies and possibly a secondary structure propensity bias.

The peptide microarrays were generated using synthetic peptides designed with the Antigenic tool using cDNA sequences of seroreactive proteins. At the same time, protein microarrays were produced using the recombinantly expressed proteins from the human cDNAs expressed in *E. coli*. The peptide and protein microarrays were used for the evaluation of the same set of serum samples. On peptide arrays, classification success for distinguishing the 3 classes of malignant, benign and control serum samples outperformed protein arrays during the class prediction analyses. Apart from the better sensitivities and specificities, ROC analyses on peptide array data provided higher AUC values compared to that of protein microarrays (Table 4 and Figure 7).

The binding ability of an antibody to a protein largely depends on the conformation at the region of binding. The antibodies specific to the proteins have the same specificity as long as the binding site is located on the surface of the molecule (Geysen *et al.* 1985). Expression of recombinant proteins in *E. coli* often leads to the production of misfolded proteins (Baneyx & Mujacic 2004). Furthermore, microarray immobilization of proteins will dramatically change their conformation and accessibility. These effects may, in turn, lead to low reproducibility and controversial findings when array-platforms are changed. In our study, the possibility of having misfolded proteins immobilized onto protein microarray may have attributed to the identification of classifier proteins which did not correlate with the results from the peptide microarrays.

As performed here, sero-reactive clones were identified by a macro-membrane based screening. On those membranes, *E. coli* clones were grown, protein expression was induced and the proteins were immobilized directly on the site of clone growth. For elucidation of the diagnostic value of identified antigenic proteins, microarrays provide today's best option for confirmation and validation of thousands of proteins in parallel. Biomarker-validation requires analyses of many patient samples that would thus be best performed on microarrays. However, switching from the macro-membranes used in biomarker identification to microarrays requires the isolation of proteins and subsequent spotting on the microarray surfaces. Although array-variability due to clone cultivation, protein expression and immobilization directly on each membrane is omitted when microarrays are generated from purified proteins, conditions are dramatically changed when moving from macro- to micro-arrays. This might also be the reason why protein microarrays in our study have shown up with moderate to low classification success of malignant breast cancer, benign breast nodules and controls. In addition, it has to be noted that when sero-reactive clones are discovered from e.g. patients and controls, the different TAA-discovery technologies like SEREX, macromembranes or phage display are currently performed using pooled samples rather than many single samples processed in parallel. Consequently, even when "differential" TAA profiles for different pools of sample classes are discovered, these findings are no warranty for any classification

success during analyses of single samples.

Since the prediction of the antigenic peptides was solely based on the protein sequences and each peptide presented a single antigenic site, there may be a better chance for auto-antibodies to bind specifically to a single feature on the microarray. Single spots of purified proteins however, might present with multiple antigenic sites and would thus enable binding of multiple antibodies. As already mentioned, this might specifically be the case for multiplexed protein analyses when high numbers of different proteins are processed under "one" condition. Conformational changes and thus presentation and accessibility upon protein immobililzation are hardly controllable and will result in a potential mixture of the antigenic sites presented by each protein on the protein arrays. This might explain why the classifiers from the peptide and protein arrays were so different at elucidating varying results. However, upon comparing the classifiers from peptide and protein array class prediction analyses, 9 genes, namely *PCSK1, DGKK, ZNF598, TBC1D9, TMEM199, EPB41L3, SAMD6, PRPF38A* and *C1orf9,* were found. Among these proteins, EPB41L3 (Dal1) is a tumour suppressor molecule which is often lost in various cancers, including breast cancer (Heller *et al.* 2007). Zn-finger proteins (represented here as ZNF598) are also frequently found to be antigenic (Backes *et al.* 2011). These proteins are usually localised in the nucleus and many of them are expressed only during embryogenesis. Thus, overexpression in various cancers might be able to elicit immune responses. Another protein that might be relevant in tumour biology is Dyacilglycerol-kinase-kappa (DGKK). Diacylglycerol kinases catalyze the phosphorylation of diacylglycerol, which is a key intracellular signalling molecule able to activate protein kinase C pathways, one of the most important targets of oncotherapy (Ron & Kazanietz 1999).

Using the genes corresponding to all the classifiers obtained from peptide and protein arrays, an over-representation analysis (ORA) was performed. An over-representation of the genes associated with spliceosomes was observed in the classifiers from both peptide and protein arrays. A plausible explanation for this can be deduced from the hypothesis put forward by Tan (1989) and Hardin (1986). These authors hypothesise that auto-antibodies often target protein complexes rather than a single protein. One conceivable explanation might be that cancer growth and invasion releases cell debris into circulation and as a consequence, evokes an immune response. Spliceosomes which are involved in alternative splicing may have a role in tumourigenesis. Processes like cell cycle control, signal transduction, angiogenisis, metastasis and apoptosis may be affected, as alternative splicing affects the majority of the human genes. Two-thirds of all the human gene transcripts are known to undergo alternative splicing. Although the function of the encoded protein does not alter in most of the cases, some may exhibit a malignant phenotype (van Alphen *et al.* 2009).

ORA of the classifiers from the peptide array revealed an over-representation of genes associated with the ribosome and the pentose phosphate pathway. Like spliceosomes, ribosomes are frequently targeted by auto-antibodies (Backes *et al.* 2011). Apart from playing a pivotal role in translational regulation, the ribosomal proteins are also associated with processes like cellular transformation, tumour growth, aggressiveness and metastasis (Zhu *et al.* 2001). Similarly, the pentose phosphate pathway plays an important role in tumour proliferation by supplying reduced levels of nicotinamide adenine dinucleotide phosphate (NADP) and carbons for intracellular anabolic processes in cancerous cells (Boros *et al.* 1998).

Over-representation of the genes associated with endocytosis was also found among genes corresponding to the protein array classifiers. Deregulated expression of the endocytosis proteins may play a role in human cancers by affecting the control of cell proliferation. The enhancement of cell replication may be promoted through impaired endocytosis as a result of prolonged signalling by growth-factor receptors (Floyd & De 1998). The genes from the protein array classifier also showed enrichment of the sequence motifs such as coiled coils and ELR motifs. These sequence motifs may have autoantigenic potentiality (Backes *et al.* 2011; Dohlman *et al.* 1993). Chemokines with the ELR motifs activate the leukocytes, which in turn, trigger an immune response (Strieter *et al.* 2004).

Although recombinant protein expression in *E. coli* has been a method of choice, the process is riddled with problems, such as the amount, length and different forms of the desired protein to be expressed (Baneyx 1999). Expression of recombinant proteins in *E. coli* often leads to the formation of biologically inactive inclusion bodies (Singh & Panda 2005). Above all else, the process of high-throughput recombinant protein expression and purification is both time consuming and cumbersome. Shorter peptide sequences of the protein can recapitulate its biological activity and can therefore act as an alternative to a full-length recombinant protein (Min & Mrksich 2004). Synthetic peptides can mimic the biological activity of a protein and present a simple means for synthesis and manipulation. These peptides are also inexpensive to synthesize and are highly stable (Cretich *et al.* 2006; Uttamchandani & Yao 2008). In addition, purified proteins

from expression clones may contain a host protein background. When using proteins on arrays expressed in *E. coli*, one may encounter the problems associated with the *E. coli* specific reactivity for the evaluation of patient sera. With the usage of short synthetic peptides, the problem of *E. coli* or host specific reactivity can be avoided. These salient features make them a desirable candidate to replace protein arrays.

Conclusion

Protein microarrays were generated using 642 expression clones found sero-reactive with breast cancer, benign breast tumours and healthy controls in a TAA macroarray screen. Antigenic peptides were deduced from clone sequences and corresponding peptide microarrays were produced. Both protein and peptide arrays were then processed with serum samples from individuals with breast cancer, benign breast tumours and healthy controls. Classification success of the 3 sample groups was moderate using protein microarrays. The peptide arrays enabled classification of the serum samples with reasonable sensitivities and specificities. Through the use of peptide arrays, the difficulties associated with the protein arrays can be circumvented and thus provide a robust platform for early diagnosis of cancer. However, in order to establish peptide arrays as a potential breast cancer diagnostic tool, test sensitivities and specificities should be increased through additional antigenic peptides which then have to be thoroughly validated on larger sets of serum samples. This study shows that *in silico* designed peptides improve the classification success and peptide microarrays can thus be a good alternative to protein arrays for auto-antibody based biomarker development.

Conflicts of interest

The authors declare that they have no financial conflicts of interest.

Acknowledgements

We want to thank Ronald Kulovics (AIT) for his help with the *E. coli* cultivation, protein isolation and printing microarrays. This work was funded partially by the *Jubiläumsfonds der Österreichischen National Bank* (Project number: 12551) and the *Vienna Science and Technology Fund* (Project number LS11-026).

References

Andresen H, Grötzinger C, Zarse K, Kreuzer OJ, Ehrentreich-Förster E & Bier FF 2006 Functional peptide microarrays for specific and sensitive antibody diagnostics. *Proteomics* **6** 1376-1384.

Backes C, Ludwig N, Leidinger P, Harz C, Hoffmann J, Keller A, Meese E & Lenhof HP 2011 Immunogenicity of autoantigens. *BMC.Genomics* **12** 340.

Baneyx F 1999 Recombinant protein expression in Escherichia coli. *Curr.Opin.Biotechnol* **10** 411-421.

Baneyx F & Mujacic M 2004 Recombinant protein folding and misfolding in Escherichia coli. *Nat.Biotechnol* **22** 1399-1408.

Boros LG, Lee PW, Brandes JL, Cascante M, Muscarella P, Schirmer WJ, Melvin WS & Ellison EC 1998 Nonoxidative pentose phosphate pathways and their direct role in ribose synthesis in tumors: is cancer a disease of cellular glucose metabolism? *Med.Hypotheses* **50** 55-59.

Bussow K, Nordhoff E, Lubbert C, Lehrach H & Walter G 2000 A human cDNA library for high-throughput protein expression screening. *Genomics* **65** 1-8.

Carter D, Dillon DC, Reynolds LD, Retter MW, Fanger G, Molesh DA, Sleath PR, McNeill PD, Vedvick TS, Reed SG, Persing DH & Houghton RL 2003 Serum antibodies to lipophilin B detected in late stage breast cancer patients. *Clin.Cancer Res* **9** 749-754.

Chapman C, Murray A, Chakrabarti J, Thorpe A, Woolston C, Sahin U, Barnes A & Robertson J 2007 Autoantibodies in breast cancer: their use as an aid to early diagnosis. *Ann.Oncol* **18** 868-873.

Conroy SE, Gibson SL, Brunstrom G, Isenberg D, Luqmani Y & Latchman DS 1995 Autoantibodies to 90 kD heat-shock protein in sera of breast cancer patients. *Lancet* **345** 126.

Cretich M, Damin F, Pirri G & Chiari M 2006 Protein and peptide arrays: recent trends and new directions. *Biomol Eng* **23** 77-88.

Disis ML, Calenoff E, McLaughlin G, Murphy AE, Chen W, Groner B, Jeschke M, Lydon N, McGlynn E, Livingston RB & . 1994 Existent T-cell and antibody immunity to HER -2/neu protein in patients with breast cancer. *Cancer Res* **54** 16-20.

Dohlman JG, Lupas A & Carson M 1993 Long charge-rich alpha-helices in systemic autoantigens. *Biochem Biophys Res Commun* **195** 686-696.

Duburcq X, Olivier C, Malingue F, Desmet R, Bouzidi A, Zhou F, Auriault C, Gras-Masse H & Melnyk O 2004 Peptide-protein microarrays for the simultaneous detection of pathogen infections. *Bioconjug Chem* **15** 301-316.

Floyd S & De CP 1998 Endocytosis proteins and cancer: a potential link? *Trends Cell Biol* **8** 299-301.

Fodor SP, Read JL, Pirrung MC, Stryer L, Lu AT & Solas D 1991 Light-directed, spatially addressable parallel chemical synthesis. *Science* **251** 767-773.

Frank R 2002 The SPOT-synthesis technique. Synthetic peptide arrays on membrane supports--principles and applications. *J Immunol Methods* **267** 13-26.

Geysen HM, Barteling SJ & Meloen RH 1985 Small peptides induce antibodies with a sequence and structural requirement for binding antigen comparable to antibodies raised against the native protein. *Proc Natl Acad Sci U.S.A*

82 178-182.
Hamrita B, Chahed K, Kabbage M, Guillier CL, Trimeche M, Chaieb A & Chouchane L 2008 Identification of tumor antigens that elicit a humoral immune response in breast cancer patients' sera by serological proteome analysis (SERPA). *Clin Chim Acta* **393** 95-102.
Hardin JA 1986 The lupus autoantigens and the pathogenesis of systemic lupus erythematosus. *Arthritis Rheum* **29** 457 -460.
Heller G, Geradts J, Ziegler B, Newsham I, Filipits M, Markis-Ritzinger EM, Kandioler D, Berger W, Stiglbauer W, Depisch D, Pirker R, Zielinski CC & Zochbauer-Muller S 2007 Downregulation of TSLC1 and DAL-1 expression occurs frequently in breast cancer. *Breast Cancer Res Treat* **103** 283-291.
Iuchi S 2001 Three classes of C2H2 zinc finger proteins. *Cell Mol.Life Sci* **58** 625-635.
Keller A, Backes C, Al-Awadhi M, Gerasch A, Kuntzer J, Kohlbacher O, Kaufmann M & Lenhof HP 2008 GeneTrailExpress: a web-based pipeline for the statistical evaluation of microarray experiments. *BMC.Bioinformatics* **9** 552.
Keşmir C, Nussbaum AK, Schild H, Detours V & Brunak S 2002 Prediction of proteasome cleavage motifs by neural networks. *Protein Eng* **15** 287-296
Kolaskar AS & Tongaonkar PC 1990 A semi-empirical method for prediction of antigenic determinants on protein antigens. *FEBS Lett* **276** 172-174.
Li YX, Yue W, Wang Y, Zhang L, Gu M & Xu S 2010 Detecting EGFR autoantibodies in serums of NSCLC patients with peptide array. *Zhongguo Fei Ai Za Zhi* **13** 727-730.
Lu H, Goodell V & Disis ML 2008 Humoral immunity directed against tumor-associated antigens as potential biomarkers for the early diagnosis of cancer. *J Proteome Res* **7** 1388-1394.
Ludwig N, Keller A, Leidinger P, Harz C, Backes C, Lenhof HP & Meese E 2012 Is there a general autoantibody signature for cancer? *Eur J Cancer* doi:10.1016/ j.ejca.2012.01.017
Lundegaard C, Lamberth K, Harndahl M, Buus S, Lund O & Nielsen M 2008 NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8-11. *Nucleic Acids Res* **36** W509-512.
Min DH & Mrksich M 2004 Peptide arrays: towards routine implementation. *Curr.Opin.Chem.Biol* **8** 554-558.
Montenarh M 2000 Chapter 20 - Humoral Immune Response Against the Growth Suppressor p53 in Human Malignancies. In *Cancer and Autoimmunity*, pp 193-203. Eds S Yehuda & MG Eric. Amsterdam: Elsevier.
Pisani P, Bray F & Parkin DM 2002 Estimates of the worldwide prevalence of cancer for 25 sites in the adult population. *Int.J.Cancer* **97** 72-81.
Piura E & Piura B 2011 Autoantibodies to tailor-made panels of tumor-associated antigens in breast carcinoma. *J Oncol.* doi:10.1155/2011/982425.
Preininger C, Bodrossy L, Sauer U, Pichler R & Weilharter A 2004 ARChip epoxy and ARChip UV for covalent on-chip immobilization of pmoA gene-specific oligonucleotides. *Anal Biochem* **330** 29-36.
Qian F, Odunsi K, Blatt LM, Scanlan MJ, Mannan M, Shah N, Montgomery J, Haddad F & Taylor M 2005 Tumor associated antigen recognition by autologous serum in patients with breast cancer. *Int J Mol Med* **15** 137-144.
R Development Core Team 2005 R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*.
Reineke U & Sabat R 2009 Antibody epitope mapping using SPOT peptide arrays. *Methods Mol Biol* **524** 145-167.
Ron D & Kazanietz MG 1999 New insights into the regulation of protein kinase C and novel phorbol ester receptors. *FASEB J* **13** 1658-1676.
Selak S, Mahler M, Miyachi K, Fritzler ML & Fritzler MJ 2003 Identification of the B-cell epitopes of the early endosome antigen 1 (EEA1). *Clin Immunol* **109** 154-164.
Shin DS, Kim DH, Chung WJ & Lee YS 2005 Combinatorial solid phase peptide synthesis and bioassays. *J Biochem Mol Biol* **38** 517-525.
Simon R, Lam A, Li MC, Ngan M, Menenzes S & Zhao Y 2007 Analysis of gene expression data using BRB-ArrayTools. *Cancer Inform* **3** 11-17.
Simon R, Radmacher MD, Dobbin K & McShane LM 2003 Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J.Natl.Cancer Inst.* **95** 14-18.
Singh SM & Panda AK 2005 Solubilization and refolding of bacterial inclusion body proteins. *J Biosci Bioeng* **99** 303-310.
Sollner J, Grohmann R, Rapberger R, Perco P, Lukas A & Mayer B 2008 Analysis and prediction of protective continuous B-cell epitopes on pathogen proteins. *Immunome Res* **4** 1.
Southern EM 1988 Analysing polynucleotide sequences. Great Britain patent application GB 8810400.5
Stempfer R, Syed P, Vierlinger K, Pichler R, Meese E, Leidinger P, Ludwig N, Kriegner A, Nohammer C & Weinhausel A 2010 Tumour auto-antibody screening: performance of protein microarrays using SEREX derived antigens. *BMC.Cancer* **10** 627.
Strieter RM, Belperio JA, Phillips RJ & Keane MP 2004 CXC chemokines in angiogenesis of cancer. *Semin Cancer Biol* **14** 195-200.
Sturgeon CM, Duffy MJ, Stenman UH, Lilja H, Brunner N, Chan DW, Babaian R, Bast RC, Jr., Dowell B, Esteva FJ, Haglund C, Harbeck N, Hayes DF, Holten-Andersen M, Klee GG, Lamerz R, Looijenga LH, Molina R, Nielsen HJ, Rittenhouse H, Semjonow A, Shih I, Sibley P, Soletormos G, Stephan C, Sokoll L, Hoffman BR & Diamandis EP 2008 National Academy of Clinical Biochemistry laboratory medicine practice guidelines for use of tumor markers in testicular, prostate, colorectal, breast, and ovarian cancers. *Clin Chem* **54** e11-e79.
Söllner J 2006 Selection and combination of machine learning classifiers for prediction of linear B-cell epitopes on proteins. *J Mol Recognit* **19** 209-214.
Tan EM 1989 Antinuclear antibodies: diagnostic markers for autoimmune diseases and probes for cell biology. *Adv Immunol* **44** 93-151.
Ullah MF & Aatif M 2009 The footprints of cancer development: Cancer biomarkers. *Cancer Treat Rev* **35** 193-200.

Uttamchandani M & Yao SQ 2008 Peptide microarrays: next generation biochips for detection, diagnostics and high-throughput screening. *Curr Pharm Des* **14** 2428-2438.
van Alphen RJ, Wiemer EA, Burger H & Eskens FA 2009 The spliceosome as target for anticancer treatment. *Br J Cancer* **100** 228-232.
Wilce MCJ, Aguilar M & Hearn MTW 1995 Physicochemical Basis of Amino Acid Hydrophobicity Scales: Evaluation of Four New Scales of Amino Acid Hydrophobicity Coefficients Derived from RP-HPLC of Peptides. *Anal Chem* **67** 1210-1219.
Wright GW & Simon RM 2003 A random variance model for detection of differential gene expression in small microarray experiments. *Bioinformatics* **19** 2448-2455.
Zhu Y, Lin H, Li Z, Wang M & Luo J 2001 Modulation of expression of ribosomal protein L7a (rpL7a) by ethanol in human breast cancer cells. *Breast Cancer Res Treat* **69** 29-38.