BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 9 2014, pages 1205–1213
doi:10.1093/bioinformatics/btu009

Sequence analysis

Advance Access publication January 9, 2014

Identification of short terminal motifs enriched by antibodies using peptide mass fingerprinting

Hannes Planatscher[1,*], Frederik Weiß[1], David Eisen[1], B.H.J. van den Berg[1], Andreas Zell[2], Thomas Joos[1] and Oliver Poetz[1]

[1]Natural and Medical Sciences Institute (NMI) at the University of Tübingen, Markwiesenstr. 55, D-72770 Reutlingen and [2]Center for Bioinformatics, University of Tübingen, Sand 1, D-72076 Tuebingen, Germany

Associate Editor: Dr John Hancock

ABSTRACT

Motivation: Mass spectrometry-based protein profiling has become a key technology in biomedical research and biomarker discovery. Sample preparation strategies that reduce the complexity of tryptic digests by immunoaffinity substantially increase throughput and sensitivity in proteomic mass spectrometry. The scarce availability of peptide-specific capture antibodies limits these approaches. Recently antibodies directed against short terminal motifs were found to enrich subsets of peptides with identical terminal sequences. This approach holds the promise of a significant gain in efficiency. TXP (Triple X Proteomics) and context-independent motif specific/global proteome survey binders are variants of this concept. Principally the binding motifs of such antibodies have to be elucidated after generating these antibodies. This entails a substantial effort in the lab, as it requires synthetic peptide libraries and numerous mass spectrometry experiments.

Results: We present an algorithm for predicting the antibody-binding motif in a mass spectrum obtained from a tryptic digest of a common cell line after immunoprecipitation. The epitope prediction, based on peptide mass fingerprinting, reveals the most enriched terminal epitopes. The tool provides a *P*-value for each potential epitope, estimated by sampling random spectra from a peptide database. The second algorithm combines the predicted sequences to more complex binding motifs. A comparison with library screenings shows that the predictions made by the novel methods are reliable and reproducible indicators of the binding properties of an antibody.

Availability: Mass spectrum data, predictions, sampling tables, consensus peptide databases and the applied protocols are available as Supplementary Material. TXP-Terminus Enrichment Analysis (TEA) and MATERICS (Mass-spectrometric Analysis of Terminal Epitope Enrichment in Complex Samples) are available as web services at http://webservices.nmi.de/materics.

Contact: hannes.planatscher@nmi.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on July 30, 2013; revised on December 14, 2013; accepted on January 6, 2014

*To whom correspondence should be addressed.

1 INTRODUCTION

Immunoaffinity enrichment, combined with mass spectrometry, increases sample throughput and sensitivity in proteomic research (Anderson *et al*., 2004). However, the generation and detailed characterization of appropriate capture reagents act as a bottleneck and involve substantial lab resources. This has led various groups to develop techniques enabling the enrichment of multiple targets using generic binders. Rush *et al*. (2005) showed that an antibody directed against phosphotyrosine can enrich tyrosin-phosphorylated proteins. The Triple X Proteomics (TXP) technique (Poetz *et al*., 2009) and the Global Proteome Survey (GPS) methodology have extended this concept and are rational approaches aimed to achieving motif-specific peptide enrichment. In the TXP approach, antibodies bind to short linear epitopes present in multiple peptides of complex samples after protein fragmentation by trypsin (see Fig. 1). *In silico* selection of antigens reduces to minimum the set of TXP antibodies required to cover a pre-defined protein target list (Planatscher *et al*., 2010).

Search space restriction gained from the revealed binding epitope could improve protein identification from MS and MSMS data. However the enrichment of proteotypic peptides with TXP antibodies leads to new challenges in the analysis of mass spectrometric datasets. Results from immunoaffinity experiments using TXP-antibodies revealed the enrichment of peptides containing the targeted epitope. Some identified peptides also matched sequence variants (Hoeppe *et al*., 2011). In that study it came clear that the terms ‘specific’ and ‘unspecific’ must be considered inappropriate in charcterizing these binders. ‘Specificity’ generally refers to binding of one protein or peptide to an antibody. Other binding events are deemed to be unspecific, off-target or cross-reactive. In the TXP strategy, the binding of the antibody towards multiple peptides is inherent. Therefore, novel concepts for epitope identification are needed, which would enable properties of the antibody-binding domain to be distinguished from interactions occurring elsewhere.

There are three main reasons why unexpected peptides can be detected in the immunoprecipitates. First, this could be due to epitope variations in the polyclonal antibody. Second, off-target binding can occur due to sequence similarity, and unlike the immunized antigen, this involves the apparent epitope or binding motif of the antibody mixture. Third, carry-over peptides may remain detectable in the sample, solely due to their persisting massive presence, despite multiple washing steps and meticulous

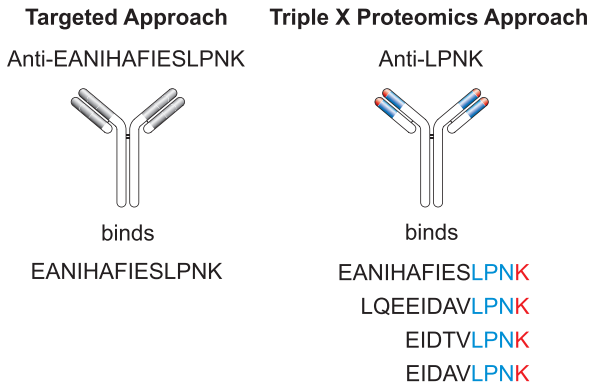


Fig. 1. Immunoaffinity MS-Approaches: peptide-specific antibodies are used to enrich tryptic fragments derived from analytes prior MS-quantification (left). In the TXP-approach antibodies for short terminal epitopes are employed for peptide-group-specific enrichment (right)

care on the part of the lab operator. While carry-over is merely noise, it is worth quantifying the variation caused by polyclonality and the apparent epitope.

Using recombinant motif-specific antibodies (GPS/context-independent motif specific-binders) which also enrich classes of peptides (Olsson *et al*., 2012a) observed that these binders were markedly promiscuous. They compared sequence data obtained from tandem mass spectrometry (LC-ESI-LTQ Orbitrap) with structural information. Their findings showed that the antibodies not only captured peptides containing the targeted epitope but also variations of it. Thus, the enriched peptides revealed binding motifs. Different amino acid side chains are known to interact with the antibody-binding site at individual positions. However such in-depth characterizations of epitope motifs analysis call for extensive experiments.

A common approach to elucidate a detailed linear epitope is to conduct an interaction analysis between an antibody and peptide libraries. These libraries comprise synthetic peptides which have at least one modified position (Houghten *et al*., 1991). For instance, if an antibody is raised against the C-terminal sequence LGYR, the peptide library for an epitope would consist of XGYR, LXYR, LGXR and LGYX-peptides, with the X representing all 20 amino acids. Instead of adding a single amino acid at step n, all 20 are added in equal amounts. This leads to a library of 80 different peptides for a four-amino-acid epitope. The synthesis and subsequent measurement of these peptides are significant cost drivers in this phase of antibody development. Moreover, the quality control for such a binder would be more expensive than its generation. Our study aimed at devising a simpler, cost-effective method entailing less effort than that required in the labour-intensive peptide-library approach. We present an algorithm for calculating the detailed epitope using MS data from an immunoaffinity-MS experiment.

Peptide mass fingerprinting is a well-established method in proteomics (Henzel *et al*., 1993; James *et al*., 1993; Mann *et al*., 1993; Pappin *et al*., 1993; Yates *et al*., 1993), on which we based the TXP-TEA (Terminal sequence Enrichment Analysis) algorithm. The original technique involves identifying an isolated and subsequently digested protein by the characteristic pattern in the peptide mass spectrum. TXP-TEA performs searches for patterns related to sets of peptides sharing a specific terminal amino acid sequence, instead of spectral patterns associated with a specific protein. The method compares the masses observed in the measurement with a theoretical spectrum from a database. The MATERICS algorithm we describe here merges the results of TXP-TEA to the antibody binding motif.

2 MS TERMINAL SEQUENCE ENRICHMENT

When an antibody enriches peptides with a common sequence, it can be anticipated that the mass spectrometer will detect more matching signals. The TXP-TEA algorithm scores signals from an observed peak list, based on a scoring table which indicates the probability of observing the same signal in a random peak under the same conditions. The less likely such a signal is, the more probable it is that the observed peak list is not random. The software reports every sequence and score that is found.

A peak list is a set of mass-to-charge/intensity pairs (mz_1, i_1), .., (mz_n, i_n), obtained from a mass spectrum by further signal processing (noise filtering, peak-picking, normalization, etc.). The dominant ion species produced in MALDI mass spectrometry has the charge $+1[M+H]^+$. The mz-values are $\frac{m+M_h}{1}$ and thus equal in value, but not in dimension, to the molecule plus a proton mass. By subtracting the proton mass $m_i = mz_i - M_H$, the mz-values $mz_1, \ldots, mz_n$ are transformed into mass values $m_1, \ldots, m_n$.

The sequence database contains information about peptides that can be expected in samples of a given species. Protein sequence databases, such as Uniprot (Wu *et al*., 2006) and proteotypic peptides databases such as the Global Proteome Machine (Fenyö *et al*., 2010) or the Peptide Atlas (Deutsch *et al*., 2008) are well-established data repositories. Proteotypic peptide databases are particularly valuable because each sequence had already been observed in mass spectrometry-based experiments. In our algorithm, a database D is a set of peptide sequences $\{p_1, \ldots, p_j\}$.

Using database D, TXP-TEA first makes a comprehensive list of peptides with theoretical masses, whose observed mass in M_{obs} matches a mass in the peak list by a predefined error threshold. Common MALDI mass spectrometers operate at a medium resolution and mass errors below 30 parts per million are normal. Due to the limited resolution and isobaric peptides, each signal in a spectrum can originate from different peptides. The set of selected peptides in range is:

$$S(D, M_{\text{obs}}, \epsilon_{\text{tol}}) = \{p_i | \frac{M_{\text{obs}} - M_{p_i}}{M_{p_i}} 10^6 | \leq \epsilon_{\text{tol}}\} \quad (1)$$

The result is a set of peptides

$$D_M = \bigcup_{M_{\text{obs}} \in M} S(D, M_{\text{obs}}, \epsilon_{\text{tol}}) \quad (2)$$

and finally a list of possible epitope candidates is generated. Each candidate epitope can explain signals in the spectrum, provided that the peptides matching that specific terminal sequence have been enriched. The number of matching peptides found in the

Table 1. Sampling table S for the consensus peptide database, number of sampled random spectra 25 000, number of peaks 100, mass tolerance 30 ppm

	number of peptides observed in spectrum				
Occurrence	1	2	3	4	5
1	3 205 722	18 373	85	0	0
2	2 111 151	22 101	182	0	0
3	1 445 403	21 227	235	0	0
4	1 189 578	23 363	350	5	0
5	1 074 187	25 731	457	6	0
6	983 959	27 333	536	5	0
7	900 073	28 868	704	14	0
8	1 009 570	37 708	1004	13	1
9	879 334	36 178	1009	21	1
10	858 291	38 574	1187	34	1
11	918 326	46 414	1495	48	1
12	931 052	50 009	1830	49	2
13	829 150	49 125	2086	69	0
14	762 225	47 894	2035	69	2
15	790 782	53 216	2445	75	2
16	888 119	64 125	3099	109	1
17	858 695	66 359	3463	127	7
18	692 556	55 687	2949	128	2
19	887 858	77 471	4556	198	2
20	796 262	72 015	4481	211	9

full database search is also relevant, because it defines the background probability.

For example, if five masses in a 73-peak spectrum match peptides sharing the c-terminal sequence LGYR and 65 peptides in the full database terminate in LGYR, this event must be rated by estimating the probability of finding five (or more) out of 65 (or less) peptides which share the same C-terminal sequence of length 4 in a random 73-peak spectrum.

We define an enrichment event $E_\Phi(i,j)$ as: i matching signals out of j masses from the same epitope class, by applying the parameters $\Phi = (D, \epsilon_{\text{tol}}, t, l, k)$ in a peak list of k masses. TXPTEA estimates the likelihood of such enrichment events by sampling from random spectra. The parameters Φ of the sampling are: the terminus (C- or N-terminal) t, sequence length l, number of peaks k, mass error tolerance ϵ_{tol} and peptide database D. The number of masses in the peak list k is also a sampling parameter. Each setting of these parameters requires a dedicated sampling table. Sampling generates many random peak lists by randomly selecting theoretical masses from the peptide database.

The random peak lists are then processed as described above. Algorithm 1 counts the number of repeated enrichment events and generates a sampling table. A sampling table is a $n \times m$-matrix S_Φ. Here m is the size of the largest epitope class in the database and n is the largest number of peaks attributable to one epitope-class, observed in a random spectrum, during the sampling process. $S_\Phi(i,j)$ is the frequency of the event that i of j expected masses match, depending on the parameters Φ.

Table 1 is the result of a sampling run of 25 000 iterations for mass spectra of 100 peaks and a mass tolerance of 30 ppm in a consensus data peptide database ($\Phi = (\text{Consensus}DB, \epsilon_{\text{tol}} = 30, t = C, l = 4, k = 100)$). The event of observing four out of 17 peptides with a common terminus occurred 127 times in 25 000 random spectra. The sampling table is not a perfect lower triangular matrix because of the occurrence of overlapping peaks. If the mass of a peptide with a unique terminal sequence is in within close range of two masses in a random spectrum, the event-counter for $E_\Phi(2,1)$ increases (i.e. observed 18 373 times in sampling Table 1).

It is better to use standard settings to limit the available choices relating to the assumed error tolerance, background database, terminus and sequence length. This limits the computational effort to a required minimum.

However the parameter k, number of peaks, varies from spectrum to spectrum. The sampling Algorithm 1 solves this difficulty by incrementally updating the sampling table. It creates the sampling table for 73-peak spectra by adding a random peak to all 72-peak spectra, from the 72-peak sampling table in the previous step. This is considerably faster than analyzing 73 peaks from scratch.

Algorithm 1 The sampling algorithms result is a table S necessary to estimate the distribution of epitope detection events in random spectra of up to k_{max} peaks

Input: parameters $\Phi = (D, \epsilon_{tol}, t, l, k_{max})$, iterations n
Output: sampling table S
foreach $seq \in D$ **do**
  $term = getTerminalEpitope(seq, l, t)$;
  $tcount[term] + +$;
end
$\Phi' = (D, \epsilon_{tol}, t, l, 0)$;
$S_\Phi = emptymatrix$;
$k = 1$;
while $k \leq k_{max}$ **do**
  $\Phi' = (D, \epsilon_{tol}, t, l, k)$;
  $S_{\Phi'} = S_\Phi$;
  $i = 1$;
  while $i \leq n$ **do**
    $randompep = D\,[randomInteger(|D|)]$;
    $P = getPeptides(D, mass(randompep), \epsilon_{tol})$;
    foreach $seq \in P$ **do**
      $term = getTerminalEpitope(seq, l, t)$;
      $count[i][term] + +$;
      $T = T \cup term$;
    end
    $i = i + 1$;
  end
  foreach $term \in T$ **do**
    if $count[i][term] > 1$ **then**
      $S_{\Phi'}[count[i][term] - 1, tcount[term]] - -$;
    end
    $S_{\Phi'}[count[i][term], tcount[term]] + +$;
  end
  $\Phi = \Phi'$;
  $k = k + 1$;
end
return S

TXP-TEA estimates the *P*-value $\hat{p}(E_\Phi(i,j))$ by dividing the count of the same and more extreme events by the total number of events:

$$\hat{p}(E_\Phi(i,j)) = \frac{\sum_{k\geq i}\sum_{j\geq l} S_\Phi(k,l)}{\sum_{k}\sum_{l} S_\Phi(k,l)}. \quad (3)$$

The *P*-value is the probability of observing a specific or more extreme event, given that the null hypothesis is true. If a rare event occurs, the estimated *P*-value is sufficiently small. The null hypothesis (Pharoah, 2007) can be rejected, because observing such data under this assumption is improbable. The null hypothesis in TXP-TEA is: 'The binder does not enrich any single terminal sequence in the sample.' A simple alternative hypothesis is: 'The binder enriches the terminal epitope LGYR in the sample.'

Figure 2 visualizes *P*-values obtained from sampling tables generated by 25 000 random spectra from a background peptide database for *Homo sapiens* (merged GPM, PeptideAtlas, *in silico* tryptic digest Uniprot).

Mass accuracy and the number of peaks determine the number of different candidate epitopes to be analyzed by TXP-TEA in a search. Each candidate represents an alternative hypothesis. As these represent different hypotheses, it is important to correct for multiple testing. By assuming a significant level of $\alpha = 0.05$ and 3000 different epitopes in a single search. By chance $3000 \times 0.05 = 150$ epitopes will have a significant *P*-value, provided that the *P*-values follow a uniform distribution. Bonferroni correction adapts the significance level to $\alpha' = \frac{\alpha}{n}$, dividing by the number of candidate sequence. Finally TXP-TEA reports enrichment of a terminal epitope sequence if $\hat{p}(E_\Phi(i,j)) \leq \alpha'$.

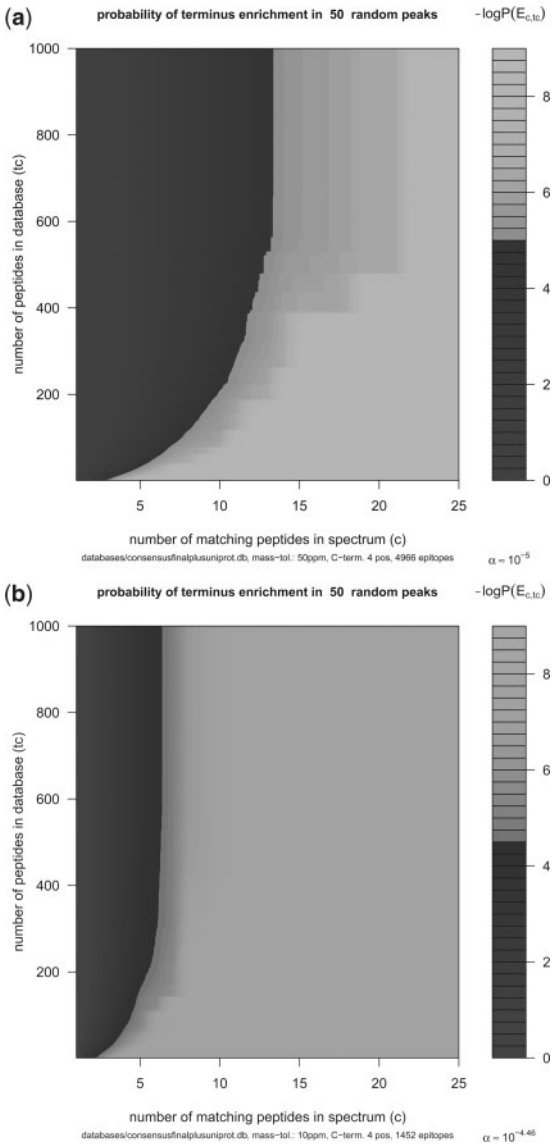


Fig. 2. Visualization of scoring tables $\hat{p}(i,j)$ with different mass tolerances and peak numbers. The lighter areas mark events that are considered to be statistically significant. The $\alpha = 0.05$ was Bonferroni-corrected by the average number of events which occurred during sampling: 50 peaks/50 ppm 4966 events, 50 peaks/10 ppm 1452 events. If mass accuracy is high, the number of peptides matching a mass will reduce. This also reduces matches of peptides with the same epitope to different peaks in the spectrum. This explains why higher mass accuracy allows TXP-TEA to detect significant enrichment events with low numbers of matching masses

3 FROM SEQUENCES TO COMPLEX EPITOPES

MS/MS experiments revealed that the enriched epitopes are more complex than the immunization antigen sequence. The antibody binds to variations of a main sequence, therefore with less affinity (Olsson *et al.*, 2012a, b). Variation often occurs mainly in one or two positions, while the other positions remain constant. These findings form the central idea of the MATERICS (mass-spectrometric analysis of terminal epitope enrichment in complex samples) algorithm, a novel approach to ascertain the motif rapidly and automatically.

Residue variation at key positions is a well-known concept which is applied by computational immunologists to MHC molecules, cell surface proteins similar to antibodies. The binding specificity of MHC-molecules is often characterized by peptide motifs (Falk and Rötzschke, 1993; Stern, 2007). A MHC class I peptide motif defines one or two internal anchor positions and an additional fixed position at the C-terminus. Each MHC allele has a characteristic motif (Sherman, 2006).

Wildcards such as LG[AYL]R describe such sequence motifs. This expression matches LGAR, LGYR and LGLR. This form gives no information about which of the three different amino acids is more probable at the variable position. Another common way to define a peptide motif is to use a position weight matrix (PWM). These matrices assign a probability p_{ij} to each amino acid A_i for a specific sequence position s_j in a sequence s.

Each motif represents a trade-off between sensitivity, specificity and model complexity. A PWM, which assigns equal probability to all amino acids at all positions, obviously matches all peaks in the observed set. Such a motif lacks any useful information, whereas a motif which assigns a probability of 1.0 to one specific amino acid at every position as well as matches a peptide for all peaks in the spectrum is a remarkable finding.

Information content, derived from Shannon Entropy, is a complexity measure for PWMs (Lund *et al.*, 2005). $H(X)$ denotes a measure of uncertainty

$$H(X) = \sum_{j}^{n} p(x_i) \cdot \log(p_{(x_i)}) \tag{4}$$

to a discrete random variable X with n different outcomes. $H(X)$ is minimal if all events are equally probable and the uncertainty is thus maximal. If one event occurs, the uncertainty is minimal, and the entropy term will maximize and be equal to 0. The complexity measure for a PWM P

$$\mathrm{IC}(P) = -\sum_{i}^{L}\sum_{j}^{N} p_{ij} \cdot \log(p_{ij}) \tag{5}$$

follows, by applying the entropy score to each position and calculating the sum.

The space of possible PWM epitopes is

$$E = \{x \in [\mathbb{R}^{20}_{[0,1]}| \sum x_i = 1.0]^l\}. \tag{6}$$

While that set is not countable, the set of wildcards $W = \mathcal{P}(A)^l$ is enumerable. A denominates the set of amino acids. With $|\mathcal{P}(A)| = 2^{20} = 1.048.576$ the number of possible wildcards of length 4 is $1.048.576^4 \approx 1209 \times 10^{24}$. MATERICS can appropriately limit a search from start, or abort it in a timely manner during the process.

In the first step of MATERICS, TXP-TEA generates the ranking of enriched sequences. MATERICS scores all motifs with one unspecific position (?XXX,X?XX,XX?X,XXX?) by combining the P-values of all the matching terminal sequences using Fisher's Method. This method is applied in meta-analysis statistics to accumulate evidence from different studies. The sum ρ of logarithms of P-values

$$\rho = -2\sum_{i=1}^{k} \log_e(p_i) \tag{7}$$

from independent tests follows a χ^2 distribution. The complementary χ^2 cumulative distribution function with $2k$ degrees of freedom is

$$p\mathrm{val}_c(\rho, k) = 1 - \frac{\gamma(k, \frac{\rho}{2})}{\Gamma(k)} \tag{8}$$

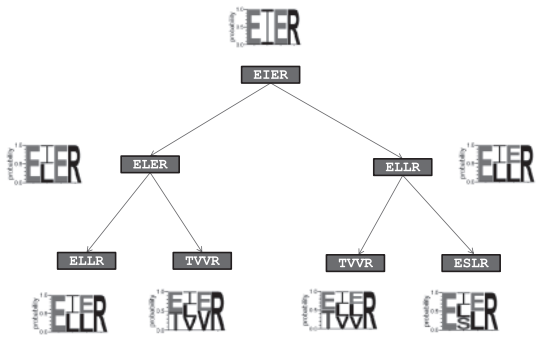


Fig. 3. Tree representation of the recursive enumeration

which gives the combined P-value. $\gamma(k, \rho)$ is the lower incomplete gamma function. The list L will include epitopes for further processing, if the combined P-values < 0.1, divided by the total number of motifs scored. This step removes candidate epitopes from the search which, even in the context of similar epitopes, will not contribute to a relevant motif.

In the next step, MATERICS combines sequences using recursive enumeration (see Fig. 3) and calculates the complexity of each motif. The recursion stops if complexity exceeds a certain limit. During recursion, the method updates the coverage score using Fisher's Method. The recursive loop calculates the sum of logarithms incrementally.

MATERICS uses a suitable property of Fisher's method as a recursion bound. Assuming the method is applied to a vector of sorted P-values

$$(p_1, p_2, \ldots, p_j, \ldots, p_n)$$

and that the combined P-value $p\mathrm{val}_c(\rho_j, j)$ increases

$$p\mathrm{val}_c(\rho_{j-1}, (j-1)) \leq p\mathrm{val}_c(\rho_j, j)$$

at one point, it follows that by adding further tests (with P-values $> p_j$) there will be no improvement in the combined P-value. This means that $p\mathrm{val}_c$ has a well-defined global optimum. The algorithm can terminate recursion once the combined P-value starts to increase. Other termination criteria include a complexity measure exceeding a predefined limit and sequence P-values > 0.05. These bounds ensure reasonably fast processing. A motif prediction takes from few seconds up to a minute on a single AMD Phenom X6 core clocked at 3.3 Ghz.

Algorithm 2 Recursive enumeration in the MATERICS algorithm

```
enum(L,i_l,c_max, c', pvals,E, ρ,d)
if ((c' < c_max)) then
    for j ← i_l to |L| do
        E' ←addToEpitope(E, L_j) ;
        c' ← −IC(E');
        ρ ← ρ + log(pvals[L_j]) ;
        pval' ← 1 − γ(2d/2, −2sumlogpval/2) / Γ(2d/2)
        updateParetoFront(E',−c',score') ;
        if ((pval' < oldpval) ∧ (pvals[L_j] < 0.05)) then
            enum(L,j,c_max,c',pvals,E', pval',ρ,d + 1) ;
        end
    end
end
```

The algorithm will include a motif in the solution set M if—and only if—no motif with lower complexity and a higher score was found. It reports motifs representing good compromises of complexity and P-value in the final output. This concept is known as Pareto optimality in the field of multi-objective optimization. A motif M_1 with a complexity score $\mathrm{IC}(M_1)$ and a P-value $p\mathrm{val}(M_1)$ will dominate a second motif M_2—if and only if—$\mathrm{IC}(M_1) > \mathrm{IC}(M_2)$ and $p\mathrm{val}(M_1) < p\mathrm{val}(M_2)$. It reflects the process of model building by the expert, which also weighs model complexity, against how much the model can explain.

The user interface presents each solution in M, enabling the user to make an informed judgement. Whereas during

enumeration of the motif subspace equal probabilities were assumed at each position, the candidate PWMs are refined afterwards. The set of matching sequences of size denominate S and f_{aj} is the frequency of amino acid a at position j.

$$p_{aj} = \frac{f_{aj}}{|S|}. \quad (9)$$

Matthews correlation coefficient (Matthews, 1975) is a good indicator of the robustness of the identified models. It measures the suitability of a classification model by the number of true positive, true negative, false positive and false negative predictions. It is impossible to compress true/false positives/negatives in a single number without information loss. However the coefficient is well established, and is particularly useful in dealing with heavily imbalanced two-class classification problems.

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}. \quad (10)$$

MATERICS deals with the MCC as an alternative measure for robustness. This convention for model fitness evaluation was applied to PWM models for TXP-epitopes enrichment in mass spectra:

	Presence predicted	Absence predicted
Observed	TP	FN
Not observed	FP	TN

The number of true negatives (absence predicted, not observed) is estimated by the number of distinguishable peaks in the range from 800–2500 Da and the specified error tolerance.

Our work compares the novel algorithms to peptide libraries for motif elucidation. Experiments included measurements before and after immunoprecipitation (IP) took place. The difference between measured signal intensities is the effect of the antibody specificity at the given position.

The pre-IP measurements account for sequence-specific ionization characteristics in the mass spectrometer as well as differences in the outcome of the peptide synthesis. The signal intensities were used to normalize the data from the post-IP experiment. The normalization coefficient for a specific peptide found in the library is:

$$F(a) = \frac{A(m_a)}{\sum_{b \in D} A(m_b)} |D| \quad (11)$$

where m_a is the known mass of the library peptide with amino acid a at a variable position, $A(m_a)$ the peak area at the respective position, and D the set of peptides found in the library. This 'flight factor' reflects the ionization properties: thus even if all the synthetic peptides are equally abundant in the library, the measured signal intensities will differ by orders of magnitude. Peptide prevalence $P(A)$ can be calculated by using $F(A)$ and the post-IP data

$$P(a) = \frac{A(M_a)F(a)}{\sum_{b \in D} A(M_b)F(b)}. \quad (12)$$

For amino acid exchanges which are not distinguishable by the resolution of the mass spectrometer, the probability is shared in the inferred motif. Therefore leucine and isoleucine will always appear as equally probable in motifs constructed by the peptide library approach. MATERICS is able to discriminate isobaric exchanges, because predictions are based on sequence database. For example if MATERICS detects only enriched peptide sequences with leucine, but none with isoleucine at the respective position, this will be reflected in the motif.

4 EXPERIMENTS

Three different antibodies generated against the 4-mer peptides AMTR, LGYR and EIER were analyzed using the MATERICS workflow. Four peptide libraries (Intavis) per binder were used, one for each amino acid position. The immunoprecipitation was performed automatically in 96-well Robotic PCR Plates (AB-1300, Thermo Scientific) using a KingFisher 96 Magnetic Particle Processor (Thermo Scientific). Of each peptide, 10-pmol library was incubated with 1 μg antibody for 1 h in 100 μl phosphate buffered saline + 0.3% n-Octyl-β-D-glucopyranoside (PBSN). The antibody–peptide complexes were precipitated for 1 h by incubation with 5 μl protein G-coated magnetic beads (20% w/v slurry, Dynabeads, Thermo Scientific). Five washing steps then followed, by transferring the beads two times into 100 μl PBSN and three times into 100 μl 50 mM ammonium hydrogen carbonate (pH 7.4) + 0.3% n-Octyl-β-D-glucopyranoside. Finally, the peptides were eluted by transferring them to 20 μl 1% formic acid. Eight-fold spotting of a 1-μl eluate was deposited onto a MALDI target (Prespotted AnchorChip, PAC II 384, Bruker Daltonics). The spots were washed with cold 10 mM ammonium phosphate + 0.1% trifluoroacetic acid in a beaker glass prior to MS analysis.

In a second experiment cultured HEK293, HELA and A751 cells were lysed at 80% confluence with lysis buffer containing 0.50% triton, 0.01% sodium dodecyl sulfate, 0.15 M NaCl, 0.01 M NaH_2PO_4, 2 mM EDTA, and 1× protease inhibitor (cOmplete Protease Inhibitor Cocktail Tablets, Roche) (pH 7.2). For tryptic digestion, 1000 μg protein extract (2μg/μl) was reduced by adding 5 μl 0.5 M dithiothreitol to reach a final concentration of 5 mM. The samples were incubated for 60 min at 60°C and 650 rpm in an orbital shaker (Thermomixer, Eppendorf). They were then cooled down and alkylated by adding 25 μl 0.2 M iodacetamide and incubated at 25°C in the dark for 30 min. After the alkylation process, the sample was diluted 1:2 with 50 mM Tris(hydroxymethyl)aminomethane and 1 mM $CaCl_2$ (pH 8.5). Trypsin (Trypsin Gold, Promega) was added in an enzyme to substrate ratio of 1:40. The digestion was performed for 16 h at 37°C. Trypsin was heat-inactivated at 100°C for 5 min. Additionally, 5.8 μl phenylmethanesulfonyl fluoride in methanol (200 mM, Sigma-Aldrich) was added. Finally, the samples were centrifuged at 13 000 rpm for 5 min. In the immunoprecipitation step, 50 μg digested protein extract per cell line was applied. In our study, we used 5 μg antibody and, accordingly, 25 μl protein G-coated magnetic beads. Immunoprecipitations were carried out thrice per cell line.

The spots were analyzed using an Ultraflex III MALDI-TOF/TOF mass spectrometer (Bruker Daltonics) in positive ion reflectron mode. The deflector cutoff was set up to 500 Da.

Table 2. Comparison of the peptide-library results to the best (smallest combined *P*-value) results obtained by a run of MATERICS with 30-ppm error tolerance, maximum complexity 2.0 using the consensus peptide database

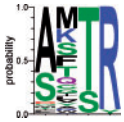
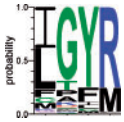
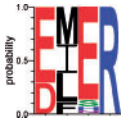
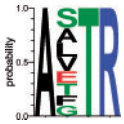
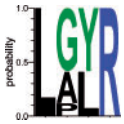
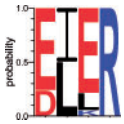
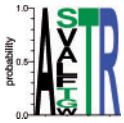
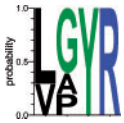
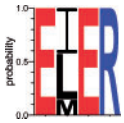
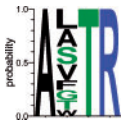
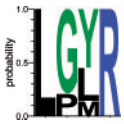
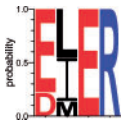
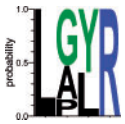
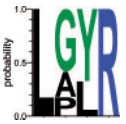
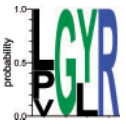
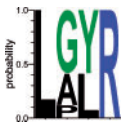
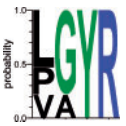
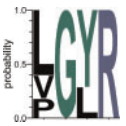
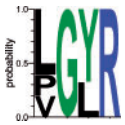
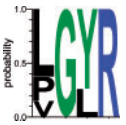
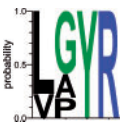
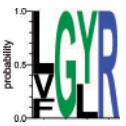
antigen	-AMTR	-LGYR	-EIER
library			
HELA			
HEK293			
A397			

Table 3. Results for all biological and technical replicates for anti-LGYR serum in combination with digested HELA-cell-lysates

	IP1	IP2	IP3
rep 1			
rep 2			
rep 3			
rep 4			

Mass calibration was performed by using pre-spotted calibrants on PAC II 384 plates. The detection mass range was set from 600 to 4000 Da. The laser power was adjusted manually. The signal intensities of 2000 shots were accumulated per spot. Peaks were annotated automatically with a signal-to-noise threshold of 3 and a mass range from 750 to 4000 Da using flexAnalysis 3.0 software (Bruker Daltonics).

Each replicate was analyzed with MATERICS by means of a unified peptide database containing the peptide sequences from GPM (Beavis, 2006), PeptideAtlas (Deutsch *et al.*, 2008), and peptide identification from the Human Plasma Proteome Project (Omenn *et al.*, 2006), with 30-ppm error tolerance, and maximum complexity 2.0.

In addition benchmark experiments were performed using artificial data to test the influence of noise on the performance of MATERICS. Methodology and results are discussed in depth in the Supplementary Material.

5 RESULTS

Table 2 summarizes the results for the positional peptide library experiments as well as those for the MATERICS algorithm in the cell line experiments. The library results for the 'AMTR' antibody indicate minor variation at the first and third positions and high variation at the second position. MATERICS predictions detected a high degree of variation at the second position. For the 'LGYR'-antibody, the library experiments revealed low variations at the first, third and fourth positions and high variations at the second position. Since isoleucine (I) and leucine (L) are isobaric and therefore indistinguishable by mass spectrometry, the library method assigns equal probability to both amino acids at the first position. MATERICS predictions suggest a variation at the second position in the enriched peptide samples precipitated from the digested HELA and A397 cell line. Our results do not indicate that isoleucine is bound at the first position, the prediction only showed leucine. When applied to HEK293 immunoprecipitate mass spectra, the algorithm suggests an alternative binding of valin (V) at the first position. The 'EIER' antibody showed a high preferential binding to its antigen sequence. Sequences which vary at the first and third position have low binding affinity. The prediction confirmed the low variability at the first position in the HELA and A397 immunoprecipitate. MATERICS detected sequence variability at the second position in all samples. It also predicted equal probabilities to leucine and isoleucine, and detected binding to methionine at the second position in the HEK293 and A397 cell line. The algorithm predicted variability at the third position of the motif in the HELA sample.

Table 3 compares the models obtained by the spectra of different immunoprecipitates and technical replicates of the same sample.

Table 4 displays the average performance of the top-ranked solutions measured by MCC. Note that the solutions ranked lower by their *P*-value can have better MCC scores (see Supplementary Material).

6 DISCUSSION

The predictions by MATERICS closely reproduced the binding motifs identified by the positional peptide-library experiments in

Table 4. Summary of the MCC scores observed on the highest ranked (*P*-value sorted) results for three binders/tissues in three IP replicates

		AMTR		LGYR		EIER
HELA	0.202 (4)	0.018	0.16 (4)	0.011	0.107 (3)	0.018
HELA	**0.216**[a] (4)	0.016	0.189 (4)	0.036	0.127 (4)	0.006
HELA	0.182 (3)	0.028	0.171 (4)	0.030	0.124 (2)	0.000
HEK	0.19 (4)	0.036	0.178 (4)	0.046	–(0)	
HEK	0.18 (4)	0.040	0.212 (4)	0.034	0.096 (1)	0.000
HEK	0.173 (4)	0.026	**0.222**[a] (4)	0.034	–(0)	
A357	0.208 (4)	0.025	0.192 (4)	0.035	**0.167**[a] (4)	0.018
A357	0.154 (4)	0.040	0.178 (4)	0.009	0.149 (4)	0.042
A357	0.19 (2)	0.018	0.193 (3)	0.043	0.137 (1)	0.000

First column is the average MCC and the number of technical replicates/spectra with a prediction result, the second column contains the standard MCC deviation. [a]Reports concluded that each cell line leads to the best MCC score for a given binder.

different cell lines for the three antibodies. However the comparison also revealed some discrepancies between the control experiment and the MATERICS results. The motif for the anti-AMTR antibody found by the library experiment is more complex than the motifs predicted by MATERICS. This could be a consequence of complexity restrictions by the algorithm. The complexity-scoring function penalizes an additional amino acid at the first position ([AS]?TR instead of A?TR) in terms of the degree of variability at the second position. The motif exceeds the complexity bound and is therefore not considered.

The binding property of the anti-EIER antibody was the most difficult to predict. MATERICS did not detect a significant enrichment in the pre-selection step for most spectra. Instead of predicting a wrong motif, it did not make any prediction at all and, instead, reported that no significant enrichment was detectable. If the input data passed the pre-selection step, the predicted motifs were observable close to the results of the peptide libraries. On only one occasion did MATERICS (see Supplementary Material, anti-AMTR, A357, biological replicate 2, technical replicate 2) make a wrong prediction: a comparison to the other technical replicates of the same immunoprecipitation ruled out this result. As shown in Table 4, experiments for the anti-EIER antibody resulted in bad quality spectra for the HEK cell lysate, whereas immunoprecitipations from HELA- and A357-digests enabled motif prediction in many technical replicates. On the whole, the HELA digest appears to be the most stable 'standard' sample for motif prediction pertaining to the three observed binders.

Although the results are obviously dependent on the observed epitope, it must be noted that using HELA cell lysate does not always produce the best prediction results. Some terminal epitopes are more abundant in some cell lines than others. Reports concluded that each cell line leads to the best MCC score for a given binder (results marked bold in Table 4). There was adequate reproducibility within different immunoprecipitates and the technical replicates, at least for the stronger binders. The results obtained from different cell lines show a reasonable level of agreement. Apart from the minor variation described, this novel approach appears to work independently of the chosen line. This strengthens flexibility as cell lines in stock at the lab can be used to perform MATERICS experiments.

Benchmark experiments using artificial data showed, that MATERICS can make reliable predictions even if only a minor fraction of the signals found in a peak list is related to the enriched epitope (see Supplementary Material).

7 CONCLUSION

Our experimental study shows that TXP-TEA and MATERICS are able to identify terminal binding motifs in immunoaffinity MS experiments. The motifs obtained closely resemble patterns found by using a peptide library approach. The described methods for motif elucidation lead to a substantial reduction in costs. In addition the novel method enables the weighting of isobaric variations in the binding motifs. These techniques might well lead to improved peptide-identification algorithms, which exploit the existing data on potentially enriched sequences during the search process. Our findings are relevant to other fields of biomedical research, such as in the identification of the binding properties of MHC molecules. Future versions of the algorithms will include options to identify internal epitopes and binding motifs as well as new ways to deal with post-translational modifications.

ACKNOWLEDGEMENT

Elise Ross helped revising the document.

Funding: This research was funded by the German Federal Ministry of Education and Research (BMBF), Foerderprogramm GO-Bio, Project "XIM - Cross Species Immunoassays fuer eine effizientere Medikamentenentwicklung" - FKZ: 031A142.

Conflict of Interest: none declared.

REFERENCES

Anderson,N.L. *et al.* (2004) Mass spectrometric quantitation of peptides and proteins using Stable Isotope Standards and Capture by Anti-Peptide Antibodies (SISCAPA). *J. Proteome Res.*, **3**, 235–244.

Beavis,R.C. (2006) Using the global proteome machine for protein identification. *Methods Mol. Biol. Clifton NJ*, **328**, 217–228.

Deutsch,E.W. *et al.* (2008) PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. *EMBO Reports*, **9**, 429–434.

Falk,K. and Rötzschke,O. (1993) Consensus motifs and peptide ligands of MHC class I molecules. *Semin. Immunol.*, **5**, 81–94.

Fenyö,D. *et al.* (2010) Computational Biology. *Methods*, **673**, 189–202.

Henzel,W.J. *et al.* (1993) Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. *Proc. Natl Acad. Sci. USA*, **90**, 5011–5015.

Hoeppe,S. *et al.* (2011) Targeting peptide termini - a novel immunoaffinity approach to reduce complexity in mass spectrometric protein identification. *Mol. Cell. Proteom. MCP*, **10**, M110.002857.

Houghten,R.A. *et al.* (1991) Generation and use of synthetic peptide combinatorial libraries for basic research and drug discovery. *Nature*, **354**, 84–86.

James,P. *et al.* (1993) Protein identification by mass profile fingerprinting. *Biochem. Biophys. Res. Commun.*, **195**, 58–64.

Lund,O. *et al.* (2005) *Immunological Bioinformatics*. The MIT Press, Cambridge, Massachusetts & London, England.

Mann,M. *et al.* (1993) Use of mass spectrometric molecular weight information to identify proteins in sequence databases. *Biol. Mass Spectrom.*, **22**, 338–345.
Matthews,B.W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et biophysica acta*, **405**, 442–451.
Olsson,N. *et al.* (2012a) Epitope-specificity of recombinant antibodies reveals promiscuous peptide-binding properties. *Protein Sci.*, **21**, 1897–1910.
Olsson,N. *et al.* (2012b) Quantitative proteomics targeting classes of motif-containing peptides using immunoaffinity-based mass spectrometry. *Mol. Cell. Proteom. MCP*, **11**, 342–354.
Omenn,G.S. *et al.* (2006) The HUPO plasma proteome project: a report from the Munich congress. *Proteomics*, **6**, 9–11.
Pappin,D.J. *et al.* (1993) Rapid identification of proteins by peptide-mass fingerprinting. *Curr. Biol.*, **3**, 327–332.
Pharoah,P. (2007) How not to interpret a *P* value? *J. Natl Cancer Inst.*, **99**, 332.
Planatscher,H. *et al.* (2010) Optimal selection of epitopes for TXP-immunoaffinity mass spectrometry. *Algorithms Mol. Biol. AMB*, **5**, 28.
Poetz,O. *et al.* (2009) Proteome wide screening using peptide affinity capture. *Proteomics*, **9**, 1518–1523.
Rush,J. *et al.* (2005) Immunoaffinity profiling of tyrosine phosphorylation in cancer cells. *Nat. Biotechnol.*, **23**, 94–101.
Sherman,L.A. (2006) To each (MHC molecule) its own (binding motif). *J. Immunol.*, **177**, 2739–2740.
Stern,L.J. (2007) Characterizing MHC-associated peptides by mass spectrometry. *J. Immunol. (Baltimore, Md.: 1950)*, **179**, 2667–2668.
Wu,C.H. *et al.* (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.*, **34**, D187–D191.
Yates,J.R.R. *et al.* (1993) Peptide mass maps: a highly informative approach to protein identification. *Anal. Biochem.*, **214**, 397–408.