DOI 10.1002/pmic.200800203 *Proteomics* 2008, *8*, 2832–2839

RESEARCH ARTICLE

A whole-genome bioinformatics approach to selection of antigens for systematic antibody generation

Lisa Berglund, Erik Björling*, Kalle Jonasson, Johan Rockberg, Linn Fagerberg, Cristina Al-Khalili Szigyarto, Åsa Sivertsson and Mathias Uhlén*

School of Biotechnology, Royal Institute of Technology (KTH), AlbaNova University Center, Stockholm, Sweden

Here, we present an antigen selection strategy based on a whole-genome bioinformatics approach, which is facilitated by an interactive visualization tool displaying protein features from both public resources and in-house generated data. The web-based bioinformatics platform has been designed for selection of multiple, non-overlapping recombinant protein epitope signature tags by display of predicted information relevant for antigens, including domain- and epitope sized sequence similarities to other proteins, transmembrane regions and signal peptides. The visualization tool also displays shared and exclusive protein regions for genes with multiple splice variants. A genome-wide analysis demonstrates that antigens for approximately 80% of the human protein-coding genes can be selected with this strategy.

Received: March 2, 2008
Revised: April 18, 2008
Accepted: April 28, 2008

Keywords:
Antibody generation / Antigen selection / Bioinformatics / Epitope / Sequence similarity

1 Introduction

One of the great challenges in bioscience today is the need for well-validated affinity reagents to explore the human proteome [1]. Such protein specific probes could be used to explore the corresponding proteins both *in vivo* and *in vitro* to allow a comprehensive characterization regarding expression levels, subcellular localization, interaction networks and differences between normal and disease tissues. However, there are several concerns for such an effort, including the choice of antigen, the choice of affinity reagent and the quality assurance of the generated affinity reagents regarding specificity and cross-reactivity [2]. The huge dynamic range of proteins in biological systems, as exemplified by the 10^{10}-fold difference in concentration between abundant and low abundant serum proteins in human blood [3], makes potential cross-reactivity difficult to predict and enforces the need for careful selection of antigens for generation of affinity reagents suitable for several application platforms.

Most bioinformatics tools for prediction of epitopes [4–11] are based on the prediction of surface accessibility parameters using various amino acid propensity scales [12]. However, such methods do not predict the possibility for cross-reactivity based on sequence similarity to other proteins from the same species as the target protein. The fact that the human genome sequence is known [13, 14] and that the coding parts of the genome can be predicted and assembled into a list of potential proteins [15, 16] opens up the possibility to develop prediction methods to exclude regions within a target protein with high sequence identity to other human proteins. We have earlier described such an alternative strategy for selection of antigens, called Protein Epitope Signature Tags (PrESTs), based on a sliding window algorithm to predict the sequence similarity of the various parts of a particular human protein to all the other protein sequences of the human proteome [17]. This information has been used to select unique PrESTs with a size between 50 and 150 amino acids. A Grid-based method was used to compute the

Correspondence: Professor Mathias Uhlén, School of Biotechnology, Royal Institute of Technology (KTH), AlbaNova University Center, SE-106 91 Stockholm, Sweden
E-mail: mathias@biotech.kth.se
Fax: +46-8-5537-8482

Abbreviations: API, application-programming interface; **LIMS,** laboratory information management system; **PrEST,** protein epitope signature tag

* Joint first authors.

sequence identity of overlapping 8, 10, 12 [18] or 50 [19] amino-acid windows of every human protein. The results show that more than 8 out of 10 amino-acid residues identical between the target protein and other proteins is rare, and should be possible to avoid in the PrEST selection process [18].

Here, we present a bioinformatics strategy for the selection of recombinant protein fragments (PrESTs) to be used for systematic generation of antibodies on a whole-proteome level (Fig. 1). The PrEST selection is facilitated by a web-based tool – PRESTIGE - that allows for *in silico* selection of antigens using the most current protein information available. The strategy is based on the design of PCR primer pairs for *de novo* cloning of PrESTs from RNA extracted from various human tissues [20]. The DNA fragments are inserted into an *Escherichia coli* expression vector [21] and the recombinant PrESTs are validated by mass spectrometry and used for immunization. The polyclonal antisera are finally purified using the antigen as affinity ligand [22] and the obtained mono-specific antibodies are validated by protein arrays and Western blots. The approved antibodies are then used for a systematic analysis to determine protein profiles in human normal and disease cells and tissues using tissue microarrays [2] and confocal microscopy [23]. Here, we discuss the various criteria for selection of antigens, including the strategy to find lowest possible sequence identity of the selected fragment to proteins from other genes, to avoid cross-reactivity of the generated antibodies.

2 Materials and methods

2.1 PRESTIGE

PRESTIGE is a web-based tool for antigen selection and visualization, developed with PHP: hypertext preprocessor using the GD graphics library extension. All the interactivity is handled by JavaScript. The Ensembl application-programming interface (API) [24] is utilized to retrieve gene and protein data from a local instance of the Ensembl [25] core MySQL database. Protein sequence data are retrieved from Ensembl for prediction of membrane protein features by pre-processing with the TMHMM 2.0 software package [26]. Sliding window sequence identity information is pre-processed from Ensembl protein sequence data by a Grid-based procedure, as described elsewhere [19]. Shared and exclusive regions between splice variants are determined *via* a pre-process involving a Java application comparing coding exons (Rockberg, J., unpublished data). All pre-processed data are stored in a local MySQL database. The restriction sites for selected restriction enzymes are found "on-the-fly" based on the protein-coding parts of the transcript sequence. PRESTIGE is connected to a laboratory information management system (LIMS), where the status and sequence of previously selected antigens are stored. Any antigen already selected on a target protein will be displayed in PRESTIGE together with the status of that antigen. A previously selected antigen has to be at least 80% identical to the target protein sequence to be assigned to that protein and any sequence identity <100% is indicated in the protein view. Selected antigens are loaded into the LIMS MySQL database after design of PrEST-specific primers.

2.2 Whole-genome analysis

The data used for PRESTIGE, based on Ensembl version 48.36 (22 997 protein-coding genes), were analyzed to find PrESTs of length 50-amino-acid residues with less than 60% sequence identity of the PrEST to other proteins. Also, any PrEST containing a 10-amino-acid window with more than 8/10 amino acids identical to another protein was discarded.

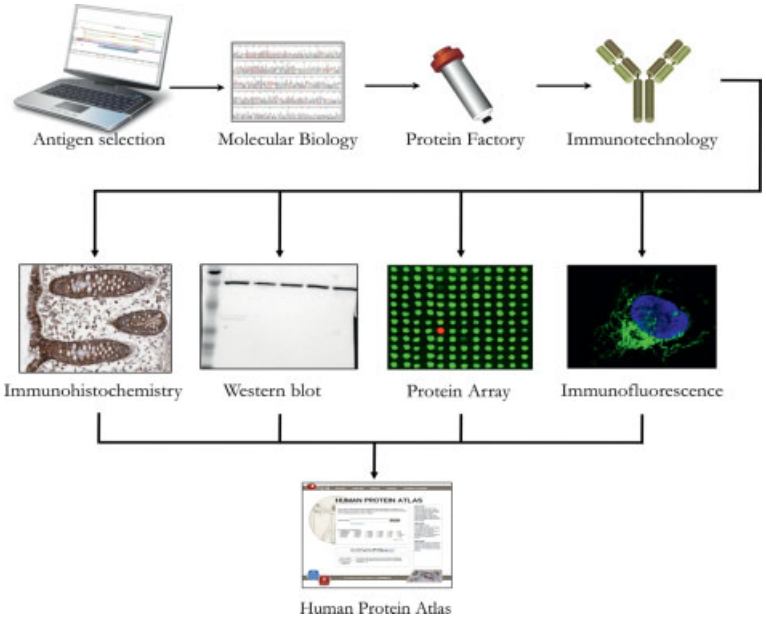

Figure 1. The workflow for generation of mono-specific antibodies with PrESTs as antigens. PrESTs are selected *in silico* and PrEST-specific primers are used for RT-PCR amplification of the selected fragments from total RNA. The amplified PrEST cDNA is sequence-verified and inserted into a vector for subsequent recombinant protein expression in *Escherichia coli*. The PrEST protein is used for immunization and the retrieved serum is affinity-purified with the PrEST as ligand. The purified antibodies are tested for specificity by protein arrays and Western blotting, and the expression pattern of the protein is studied by immunohistochemical staining of numerous human tissue and cell samples.

The PrESTs were not allowed to overlap with predicted transmembrane regions, or with a signal peptide, as predicted by SignalP [27]. The search for PrESTs was done from the N to the C terminus of the protein, counting every possible non-overlapping PrESTs.

3 Results

3.1 The antigen selection strategy

In Table 1, the criteria used for PrEST selection, to enable generation of protein-specific antibodies, are summarized. First, uniqueness of the selected fragment compared to proteins from other genes is ensured using a sliding window algorithm [18, 19] to predict the sequence identity of windows of 50 and 10 amino acids, respectively, throughout the target protein. The 50-amino-acid window is primarily used to avoid selection of domain-sized regions with high sequence identity between the analyzed protein and another human protein, which could indicate structural similarity between the proteins that in turn might lead to cross-reactive conformational epitopes. The smaller window (ten amino acids) is used to avoid local high sequence identity to another protein, which might cause cross-reactivity involving linear epitopes. The transmembrane regions are avoided in the selection of antigen due to the inaccessibility for the antibodies to the membrane-spanning region of the target protein. In addition, the production of recombinant proteins containing hydrophobic regions is problematic with low yields and success rates [28]. Any predicted signal peptide is avoided in the selection process, as it is cleaved from the native target protein. Antigens are preferably selected on parts of the protein shared by all splice variants of the target gene, to cover all proteins encoded by the target gene with the generated antibodies. Alternatively, antigens are selected on exclusive parts of the protein to enable studies of this particular splice variant. As an additional criterion, cleavage sites for particular restriction enzymes, for example to be used in subsequent cloning procedures, can be excluded from the PrEST.

3.2 The protein visualization and antigen selection tool

A web-based tool has been developed to allow for visualization of protein features and interactive selection of suitable antigen regions. The tool, called PRESTIGE, is divided into various sections (Fig. 2), including a sequence window displaying the protein sequence and a protein model with information about the protein using various prediction algorithms. The information behind the features displayed is obtained both by in-house computational analysis of sequence data and directly from external sources (Fig. 3). The protein- and gene identifiers and all sequences displayed in PRESTIGE are retrieved from the latest version of the Ensembl database [25], as well as data for predicted signal peptide and InterPro regions [29]. Predicted transmembrane regions and inside/outside location of loops for membrane proteins are visualized. The sequence identity of the different parts of the protein is retrieved *via* a sliding window sequence similarity search using multiple CPUs [19]. The target protein is analyzed for shared regions between all splice variants of the corresponding gene, as well as any regions exclusive to this particular splice variant (Rockberg, J., unpublished data). All the above given data are pre-computed and stored in a local database. PRESTIGE is tightly connected to a LIMS (Fig. 3), which allows for direct submission of selected PrESTs to the antibody generation workflow. In addition, this enables visual tracking of the status of selected PrESTs in the experimental workflow.

Table 1. Features for antigen selection

Feature	Comment	Selection strategy
Sequence identity of protein domains	Sliding window sequence comparison with 50 amino acids window	Avoid regions with >60% sequence identity
Sequence identity of linear epitopes	Sliding window sequence comparison with 10 amino acids window	Avoid regions with more than 8 out of 10 amino acids identical
Transmembrane regions	Prediction of transmembrane regions	Avoid transmembrane regions
Signal peptide	Prediction of signal peptide	Avoid signal peptide
Splice variants	Analysis of shared and exclusive protein regions between splice variants	Select antigens primarily on shared regions
Restriction enzyme recognition sites	Find sites in the protein-coding parts of the transcript sequence	Avoid restriction sites used in cloning

Figure 2. PRESTIGE – a tool for antigen selection. The antigen selection tool is divided into three different sections: A sequence window (A) displaying the protein sequence and the protein-coding part of the transcript sequence for the target protein, a button pane (B) with user interactivity such as zooming, choice of positions, primer design *etc.*, and a protein model (C-L). The protein model shows the scale and length of the protein (C), restriction sites used for subsequent cloning (D), InterPro regions (E), low complexity regions (in the amino acid sequence) (F), regions shared between splice variants as well as regions exclusive to this particular variant (G), signal peptide (H), transmembrane regions and inside/outside location of loops (I), sequence identity to proteins from other genes based on a 50 amino acids sliding window (J) or a 10 amino acids sliding window (K), and positions of selected PrESTs (L). Each pixel in the sequence identity features (J, K) represents the center of the sliding window and the color of the pixels represents the sequence identity of the current window to proteins from other genes.

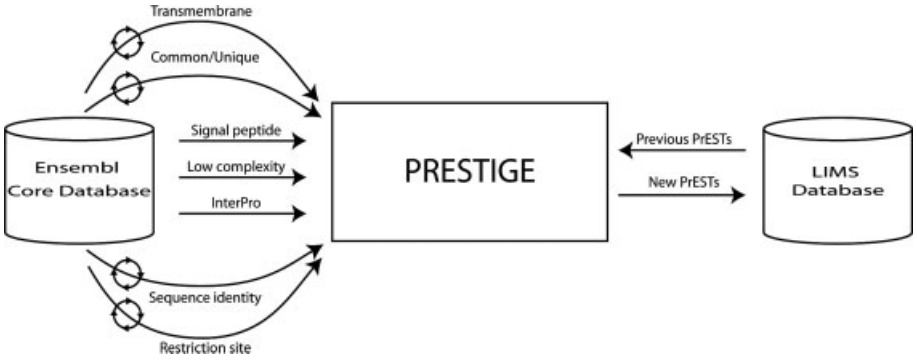

Figure 3. Schematic overview of the PRESTIGE framework. Transcript- and protein sequence data, InterPro regions, low complexity regions, and predicted signal peptides are retrieved from a local version of the latest Ensembl core database, *via* the Ensembl application programming interface (API), for visualization in PRESTIGE. The sequences used for sliding window protein sequence similarity searches, membrane protein predictions and restriction site mapping are also fetched *via* the API. Shared and exclusive regions for splice variants of the same gene are analyzed based on data from Ensembl. Previously selected PrESTs and their status in the experimental workflow are retrieved from a LIMS. Selected PrESTs are submitted to the LIMS after PrEST-specific oligonucleotide primer selection.

3.3 Selection of antigens

The design of suitable antigens for antibody generation is exemplified in Fig. 4. A PrEST length around 75 amino acids is preferred, which is a length proven successful in the experimental pipeline (Berglund, L., unpublished data), while at the same time, in most cases, allowing for the design of several PrESTs on the same protein. For our antibody-based proteomics effort [2], we have aimed to select up to four PrESTs *per* protein to increase the rate of successfully generated antibodies and to have antibodies directed to different parts of the protein. Below, three examples are given in which the design features and criteria described in Table 1 have been used.

The first example involves the ATP-binding cassette, subfamily A (ABC1), member 3 (ABCA3) gene product, which has been suggested to play an important role in the formation of pulmonary surfactant, probably by transporting lipids such as cholesterol [30]. One of the proteins encoded by ABCA3 is membrane bound with 11 predicted transmembrane regions (Fig. 4A). In the N terminus of the protein, a signal peptide is predicted. The ABC-transporter domains are represented by the InterPro feature and include regions with high sequence identity to other proteins, which is also shown on the protein identity scales. The protein has long regions with low sequence identity to proteins from other genes, both on the domain- and the epitope scale. In the protein-coding part of the transcript corresponding to this

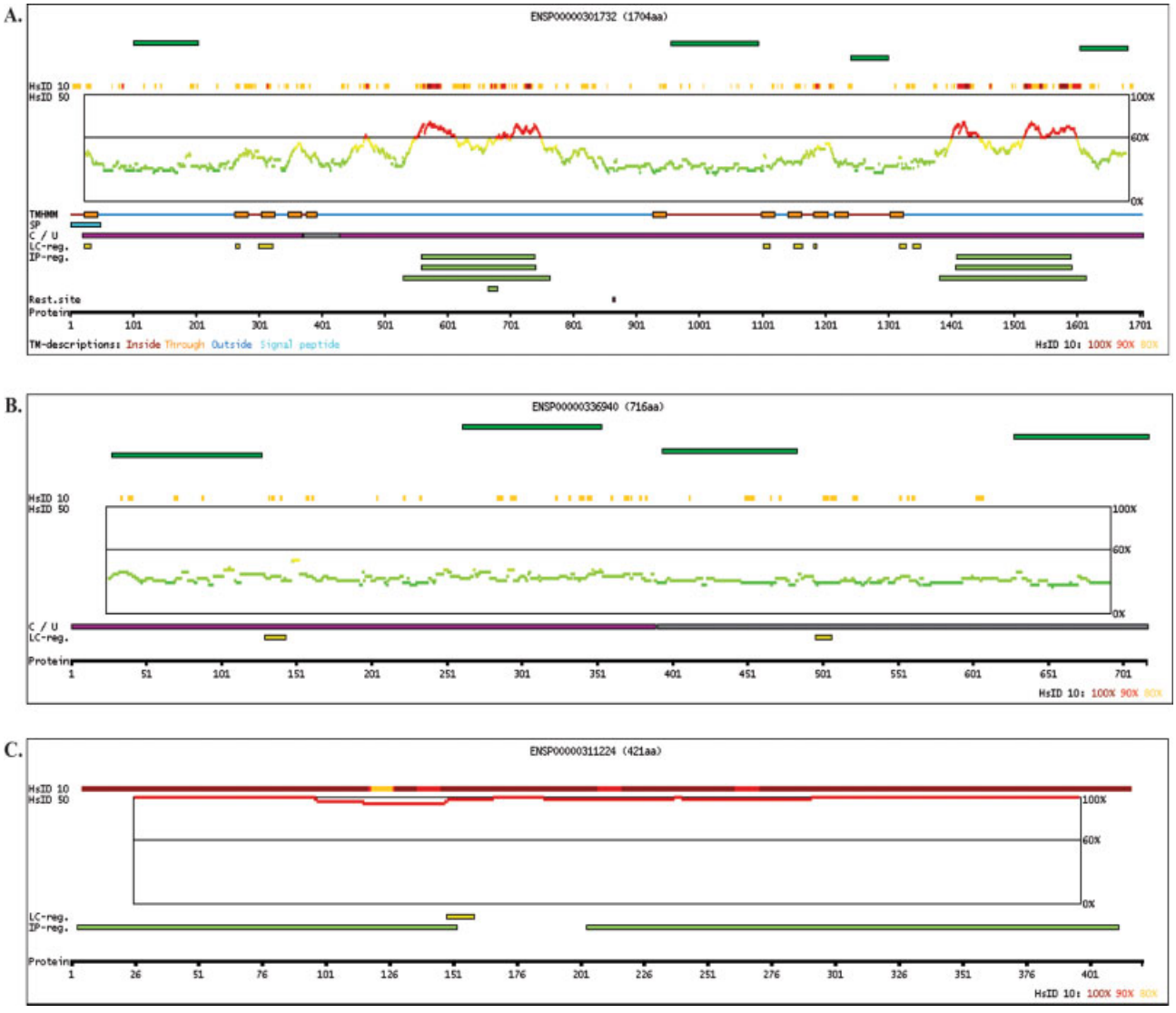

Figure 4. Antigen selection. Suggested antigens for A, the membrane-bound protein encoded by ABCA3, and B, a splice variant of the KIAA1958 gene. (C) No specific antigens can be selected for the gene ACOT1.

protein, there is a cleavage site for the restriction enzyme AscI, which is used in the subsequent cloning of the selected fragment. The restriction site is represented at amino acid position 863 in the protein. The majority of the protein sequence is shared between the splice variants of the ABCA3 gene, but there is a 58 amino acid long exclusive region spanning amino acid positions 371–428. The strategy to select antigens with low sequence identity to other proteins, while at the same time avoiding transmembrane regions, signal peptides and chosen restriction sites, allows for four non-overlapping PrESTs on this protein (Fig. 4A). The exclusive region of the protein is not used for antigen selection due to the overlap with a transmembrane region.

The second example involves the uncharacterized KIAA1958 gene (Fig. 4B). As this protein does not have a high similarity to any other protein and no InterPro regions, the function is difficult to predict from the protein sequence alone and hence it is valuable to create specific antibodies to study the tissue expression and function of the protein. The gene has, according to Ensembl, two alternative splice variants (Fig. 4B), of which one, ENSP00000336940, has a long region shared with the other variant, in addition to an exclusive 327-amino-acid region. It is possible to select PrESTs both on the shared and on the exclusive region and generation of antibodies towards this protein will possibly give some clues about the characteristics of KIAA1958.

The third example is the acyl-CoA thioesterase 1 (ACOT1) gene belonging to the C/M/P thioester hydrolase family and catalyzing the hydrolysis of acyl-CoA to free fatty acid and coenzyme A (CoASH). This protein has the potential to regulate intracellular levels of acyl-CoA, free fatty acids and CoASH [31]. As shown by our visualization software (Fig. 4C), all regions of the ACOT1 have high sequence identity to other human proteins. It is therefore not possible to select PrESTs with the Table 1 criteria (<60% sequence identity on the 50-amino-acid window level). A more thorough analysis reveals that the protein encoded by the ACOT1 gene is very sequence similar to other members in the C/M/P thioester hydrolase family.

3.4 Whole-proteome analysis

An important question is the possibility for complete coverage of the genome, selecting antigens to one representative protein from each human gene. A whole-genome analysis was therefore performed to investigate the number of genes, out of the 22 997 human protein-coding genes in Ensembl (version 48.36), for which it is possible to find at least one PrEST for the encoded protein(s) (Fig. 5). All possible 50-amino acid PrESTs, with full-length sequence identity less than 60% and no 10-amino-acid window with more than eight out of ten amino acids identical to other proteins, were located for each human protein. Additionally, no overlap with transmembrane regions or signal peptide was allowed in the selected region. The results suggest that it is possible to generate unique antigens, using the above given criteria,

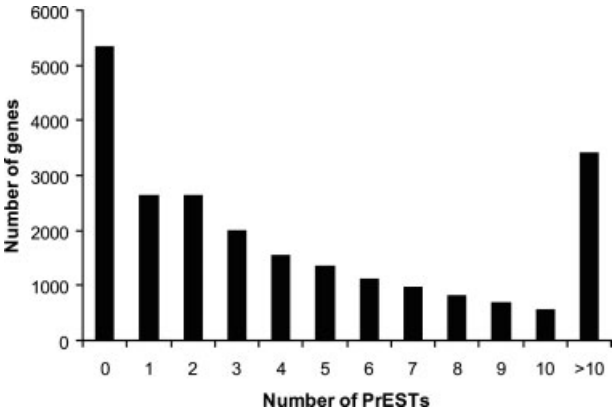

Figure 5. Whole-genome analysis. Number of predicted PrESTs *per* gene. All non-overlapping PrESTs with a length of 50 amino acid residues were counted for each protein and the protein with the highest number of PrESTs was selected as the representative for the target gene.

for the vast majority of the human genes (77%), and that at least two non-overlapping PrESTs can be selected for 65% of the genes. Interestingly, more than 10 separate PrESTs can be designed for more than 15% of the human genes. An analysis of the failed genes (23%) showed, in most cases, high sequence identity of the target protein to another human protein. Other features that prevents PrEST selection is too short protein sequence or too short loops between transmembrane regions.

4 Discussion

Here, we show that a web-based visualization tool can be used to facilitate the design and selection of antigens for generation of protein-specific antibodies to the human proteome. Due to the availability of the complete human genome sequence, the concept is based on the generation of protein or peptide fragments with low sequence identity to all other human proteins. Thus, selection of antigens can be combined with efficient methods to generate recombinant protein fragments, making the strategy suitable for high-throughput proteomics efforts. It is also possible to use the same strategy for the selection of shorter antigens, including the non-recombinant route in which peptides are chemically synthesized. In this case, the prediction using the 10-amino-acid window is of course more important than the longer 50-amino-acid window. Although conceptually attractive, the alternatives to generate full-length proteins or to isolate proteins from native sources has the disadvantage that careful mapping of the binding epitope for the resulting affinity reagent is needed, to ensure that unspecific regions have not been used.

Other methods for selection of antigens focus on the immunogenicity of the protein regions. Unfortunately, an

evaluation of these methods shows that they perform only marginally better than random [32] and therefore will only be added to the tool if improved.

The strategy described here is based on the prediction of domain-sized regions with high sequence identity between the analyzed protein and other human proteins, as well as shorter high sequence identity linear sequences. Many antibody-based applications, such as Western blots, immunohistochemistry and fluorescent-based confocal microscopy, involve protein targets that are wholly or partially unfolded and thus exclusion of linear highly sequence identical short sequences (linear epitopes) is important. Other applications, such as serum screening or therapeutic applications, involve proteins in their native form, where native (conformational) epitopes are most likely preferred. For these applications, domain-sized regions with high sequence identity between the analyzed protein and other human proteins, implicating structural similarity, are essential to avoid in the antigen.

The results of the whole-genome analysis showed that for the vast majority (77%) of the human proteins, it is possible to select a 50-amino-acid region with less than 60% sequence identity to other proteins, no local sequence identity higher than 8 out of 10 amino acids, and no predicted transmembrane or signal peptide region. The most common reason for not being able to select a PrEST for a gene is the existence of a related gene with similar or identical (duplicated) protein sequence. By adding information about such highly sequence similar proteins to the antigen selection tool, it might in the future be possible to select antigens that are common to all the sequence similar proteins, but still dissimilar to all other proteins. In this manner, it will be possible to generate antibodies that will recognize a defined group of proteins (pluri-specific antibodies) instead of one single protein (mono-specific antibodies). Alternatively, if the sequence identity of the protein to other proteins is less than 100%, shorter antigens (peptides) can be generated where the local sequence identity of the target protein to other protein is low.

Currently, only human protein data is displayed in the antigen selection tool, but it would be rather easy to shift to any other species, especially if that species is part of the Ensembl collection. It is therefore possible to develop the bioinformatics tool further to enable generation of inter-species reactive antibodies by adding data from other species. Selection of antigens for such studies would include identification of the most similar sequence region between human and the species of interest, while at the same time avoiding regions similar within the species to ensure generation of protein-specific antibodies.

In summary, we here describe a new bioinformatics tool, PRESTIGE, for selection of antigens suitable for antibody generation. The strategy outlined has, this far, been used for the selection of antigens corresponding to 13 600 human genes and it is aimed to generate antibodies for the Human Protein Atlas portal (http://www.proteinatlas.org), including protein expression data from 67 normal cell types, 216 patients representing all major cancers and 48 human cell lines. At present, the Human Protein Atlas (version 3.1) contains data from 3014 antibodies and more than 2.9 million high-resolution immunohistochemistry images, all manually annotated by a certified pathologist [2]. In addition, several thousand confocal microscopy images showing subcellular localization in human cell lines are present in the atlas [23]. The epitope visualization using the PRESTIGE software will, in the near future, be publicly available through the portal for all "in-house" generated antibodies. In this way, the exact sequence of the antigen in relationship with other relevant features of the protein target is obtained. This is important for international efforts [33] to generate paired antibodies to the human proteome [34], in which the paired antibodies should be produced towards two separate and non-overlapping epitopes of the same protein target. The ultimate goal of such efforts is to generate a toolbox of publicly available, renewable antibodies with defined epitope sequences to provide a resource for basic biomedical research as well as various clinical applications.

This work was supported by grants from the Knut and Alice Wallenberg Foundation.

The authors have declared no conflict of interest.

5 References

[1] Blow, N., Antibodies: The generation game. *Nature* 2007, *447*, 741–744.

[2] Uhlén, M., Björling, E., Agaton, C., Szigyarto, C. A. *et al.*, A human protein atlas for normal and cancer tissues based on antibody proteomics. *Mol. Cell. Proteomics* 2005, *4*, 1920–1932.

[3] Anderson, N. L., Anderson, N. G., The human plasma proteome: history, character, and diagnostic prospects. *Mol. Cell. Proteomics* 2002, *1*, 845–867.

[4] Menendez-Arias, L. and Rodriguez, R., A BASIC microcomputer program for prediction of B and T cell epitopes in proteins. *Comput. Appl. Biosci.* 1990, *6*, 101–105.

[5] Hopp, T. P., Woods, K. R., Prediction of protein antigenic determinants from amino acid sequences. *Proc. Natl. Acad. Sci. USA* 1981, *78*, 3824–3828.

[6] Greenbaum, J. A., Andersen, P. H., Blythe, M., Bui, H. H. *et al.*, Towards a consensus on datasets and evaluation metrics for developing B-cell epitope prediction tools. *J. Mol. Recognit.* 2007, *20*, 75–82.

[7] Jameson, B. A., Wolf, H., The antigenic index: a novel algorithm for predicting antigenic determinants. *Comput. Appl. Biosci.* 1988, *4*, 181–186.

[8] Alix, A. J., Predictive estimation of protein linear epitopes by using the program PEOPLE. *Vaccine* 1999, *18*, 311–314.

[9] Maksyutov, A. Z., Zagrebelnaya, E. S., ADEPT: a computer program for prediction of protein antigenic determinants.*Comput. Appl. Biosci.* 1993, *9*, 291–297.

[10] Pellequer, J. L., Westhof, E., Van Regenmortel, M. H., Correlation between the location of antigenic sites and the prediction of turns in proteins. *Immunol. Lett.* 1993, *36*, 83–99.

[11] Parker, J. M., Guo, D., Hodges, R. S., New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and X-ray-derived accessible sites. *Biochemistry* 1986, *25*, 5425–5432.

[12] Haste Andersen, P., Nielsen, M., Lund, O., Prediction of residues in discontinuous B-cell epitopes using protein 3D structures. *Protein Sci.* 2006, *15*, 2558–2567.

[13] Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C. *et al.*, Initial sequencing and analysis of the human genome. *Nature* 2001, *409*, 860–921.

[14] Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W. *et al.*, The sequence of the human genome. *Science* 2001, *291*, 1304–1351.

[15] Curwen, V., Eyras, E., Andrews, T. D., Clarke, L. *et al.*, The Ensembl automatic gene annotation system.*Genome Res.* 2004, *14*, 942–950.

[16] Clamp, M., Fry, B., Kamal, M., Xie, X. *et al.*, Distinguishing protein-coding and noncoding genes in the human genome. *Proc. Natl. Acad. Sci. USA* 2007, *104*, 19428–19433.

[17] Lindskog, M., Rockberg, J., Uhlén, M., Sterky, F., Selection of protein epitopes for antibody production. *Biotechniques* 2005, *38*, 723–727.

[18] Berglund, L., Andrade, J., Odeberg, J., Uhlén, M., The epitope space of the human proteome. *Protein Sci.* 2008, *17*, 606–613.

[19] Andrade, J., Berglund, L., Uhlén, M., Odeberg, J., Using Grid technology for computationally intensive applied bioinformatics analyses. *In Silico Biol.* 2006, *6*, 495–504.

[20] Agaton, C., Galli, J., Höiden Guthenberg, I., Janzon, L. *et al.*, Affinity proteomics for systematic protein profiling of chromosome 21 gene products in human tissues.*Mol. Cell. Proteomics* 2003, *2*, 405–414.

[21] Larsson, M., Gräslund, S., Yuan, L., Brundell, E. *et al.*, High-throughput protein expression of cDNA products as a tool in functional genomics. *J. Biotechnol.* 2000, *80*, 143–157.

[22] Nilsson, P., Paavilainen, L., Larsson, K., Ödling, J. *et al.*, Towards a human proteome atlas: high-throughput generation of mono-specific antibodies for tissue profiling. *Proteomics* 2005, *5*, 4327–4337.

[23] Barbe, L., Lundberg, E., Oksvold, P., Stenius, A. *et al.*, Towards a confocal subcellular atlas of the human proteome. *Mol. Cell. Proteomics* 2008, *7*, 499–508.

[24] Stabenau, A., McVicker, G., Melsopp, C., Proctor, G. *et al.*, The Ensembl core software libraries. *Genome Res.* 2004, *14*, 929–933.

[25] Hubbard, T. J., Aken, B. L., Beal, K., Ballester, B. *et al.*, Ensembl 2007. *Nucleic Acids Res.* 2007, *35*, D610–617.

[26] Krogh, A., Larsson, B., von Heijne, G., Sonnhammer, E. L., Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* 2001, *305*, 567–580.

[27] Bendtsen, J D., Nielsen, H., von Heijne, G., Brunak, S., Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.* 2004, *340*, 783–795.

[28] Cunningham, F., Deber, C. M., Optimizing synthesis and expression of transmembrane peptides and proteins. *Methods* 2007, *41*, 370–380.

[29] Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A. *et al.*, The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.* 2001, *29*, 37–40.

[30] Yoshida, I., Ban, N., Inagaki, N., Expression of ABCA3, a causative gene for fatal surfactant deficiency, is up-regulated by glucocorticoids in lung alveolar type II cells. *Biochem. Biophys. Res. Commun.* 2004, *323*, 547–555.

[31] Hunt, M. C., Alexson, S. E., The role Acyl-CoA thioesterases play in mediating intracellular lipid metabolism. *Prog. Lipid Res.* 2002, *41*, 99–130.

[32] Blythe, M. J., Flower, D. R., Benchmarking B cell epitope prediction: underperformance of existing methods. *Protein Sci.* 2005, *14*, 246–248.

[33] Taussig, M. J., Stoevesandt, O., Borrebaeck, C. A., Bradbury, A. R. *et al.*, ProteomeBinders: planning a European resource of affinity reagents for analysis of the human proteome. *Nat. Methods* 2007, *4*, 13–17.

[34] Uhlén, M., Mapping the human proteome using antibodies. *Mol. Cell. Proteomics* 2007, *6*, 1455–1456.

