# The epitope space of the human proteome

LISA BERGLUND,[1] JORGE ANDRADE,[1] JACOB ODEBERG, AND MATHIAS UHLÉN
School of Biotechnology, AlbaNova University Center, Royal Institute of Technology, SE-106 91 Stockholm, Sweden

(RECEIVED November 8, 2007; FINAL REVISION December 22, 2007; ACCEPTED December 27, 2007)

**Abstract**

In the post-genome era, there is a great need for protein-specific affinity reagents to explore the human proteome. Antibodies are suitable as reagents, but generation of antibodies with low cross-reactivity to other human proteins requires careful selection of antigens. Here we show the results from a proteome-wide effort to map linear epitopes based on uniqueness relative to the entire human proteome. The analysis was based on a sliding window sequence similarity search using short windows (8, 10, and 12 amino acid residues). A comparison of exact string matching (Hamming distance) and a heuristic method (BLAST) was performed, showing that the heuristic method combined with a grid strategy allows for whole proteome analysis with high accuracy and feasible run times. The analysis shows that it is possible to find unique antigens for a majority of the human proteins, with relatively strict rules involving low sequence identity of the possible linear epitopes. The implications for human antibody-based proteomics efforts are discussed.

**Keywords:** proteomics; antigen; epitope; sequence similarity; antibody; grid

A great need exists for the generation of protein-specific affinity reagents to explore the human proteome (Uhlén 2007). The systematic generation and use of antibodies to functionally explore the proteome has been called antibody-based proteomics (Uhlén and Pontén 2005). One of the main challenges for such efforts is the selection and generation of antigens, to ensure a good antibody response in combination with low cross-reactivity to other human proteins.

The B-cell antibody binding regions can be classified into continuous (linear) epitopes with a short consecutive stretch of amino acids, or discontinuous (conformational) epitopes consisting of segments that are distantly separated in the protein sequence and brought into proximity by the folding of the protein (Barlow et al. 1986). It has been noted that the definition is not absolute, since discontinuous epitopes often consist of stretches of several consecutive amino acids, and that these, in some cases, therefore could be considered as a series of linear epitopes (Van Regenmortel 2006). In addition, the linear epitopes need to adopt a specific conformation to be recognized by their cognate antibody.

The linear and conformational epitopes have been studied with binding assays using synthetic overlapping peptides (Rodda et al. 1986; Dunn et al. 1999; Fleury et al. 2000) and phage display (Fack et al. 1997). In a few cases, these studies have been complemented with structure-based analysis using three-dimensional structures of complexes between an antibody and its peptide epitope (Vyas et al. 2003). These analyses have shown that linear epitopes normally range from six to nine amino acids. As an example, Atassi (1975) showed already in 1975 that the native myoglobin contains five distinct linear epitopes, each consisting of six to seven polar amino acids in a consecutive sequence located on the surface of the molecule.

It has been estimated that most of the B-cell epitopes are conformational, but the exact ratio between linear and conformational epitopes is difficult to estimate. The adsorption of myoglobin antibodies using synthetic (linear) epitopes suggested that only 30%-40% of the antibodies were

[1]These authors contributed equally to this work.

Reprint requests to: Mathias Uhlén, School of Biotechnology, AlbaNova University Center, Royal Institute of Technology SE-106 91 Stockholm, Sweden; e-mail: mathias@biotech.kth.se; fax: 46-8 5537-8482.

Article and publication are at http://www.proteinscience.org/cgi/doi/10.1110/ps.073347208.

conformational (Lando et al. 1982), but other studies have shown a higher ratio of conformational epitopes (Van Regenmortel 1996). Haste Andersen et al. (2006) compiled a data set of 76 X-ray structures of antibodies in complex with antigens representing discontinuous epitopes, and their analysis showed that the total number of amino acid residues per epitope ranged from 9 to 22, with most epitopes consisting of 14–19 residues. Interestingly, these conformational epitopes often had stretches of several linear epitopes, with most epitopes having at least one linear stretch in the range of 4–7 residues. These findings suggest that most discontinuous epitopes are composed of short linear epitope sequences forming a binding region for the antibody.

It is important to point out that the aim of an antibody-based proteomics effort is to generate antibodies that in most cases should recognize wholly or partially unfolded proteins (Uhlén and Pontén 2005), since the most common applications for the generated antibodies involve various methods for protein denaturation, such as detergents (Western blots), formaldehyde (immunohistochemistry), or acetone (immunofluorescence). One could, therefore, speculate that linear epitopes might play a much larger role for such efforts compared to the more conventional efforts to try to create vaccines, antibodies for serum screening (recognizing native proteins), or therapeutic antibodies. The fact that most antibodies mapped so far have been shown to be predominantly conformational (Van Regenmortel 1996) might therefore be a consequence of bias toward the generation of antibodies with the aim to recognize the native protein target. It remains to be seen if the antibodies generated by the approach involving recombinant protein fragments with low probability of native folds instead generate B-cell epitopes of mainly linear character.

Many methods exist for the prediction of linear epitopes using the protein sequence as input (Hopp and Woods 1981; Alix 1999; Odorico and Pellequer 2003; Greenbaum et al. 2007). In general, these methods are based on prediction of hydrophilicity, flexibility, β-turns, and other surface accessibility determinants using a number of amino acid propensity scales (Haste Andersen et al. 2006). However, predicting linear epitopes is still not very reliable, and based on a data set of 50 linear epitopes, Blythe and Flower (2005) concluded that B-cell epitope predictions are only marginally better than random. All these methods also have the limitation that they 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 (International Human Genome Sequencing Consortium 2004) and that the coding parts of the genome can be predicted and assembled into a list of potential proteins (Hubbard et al. 2007) opens up the possibility to perform such studies. As a consequence for a human antibody-based proteomics effort, the antigen selection could therefore be focused on prediction methods to exclude regions within a target protein with high sequence identity to other human proteins.

We have earlier described a strategy for selection of protein epitope signature tags (PrESTs) (Lindskog et al. 2005), usually between 50 and 150 residues, where the selection is facilitated by a visualization software displaying regions of low sequence identity to other proteins. A grid-based blastp (Altschul et al. 1990, 1997) method was used to assemble the sequence identity of overlapping 50 amino acid windows of every human protein to the human proteome (Andrade et al. 2006). The result from that analysis was implemented in a visualization tool that has now been used to design more than 27,000 human PrESTs (L. Berglund, E. Björling, K. Jonasson, J. Rockberg, L. Fagerberg, C. Al-Khalili Szigyarto, Å. Sivertsson, and M. Uhlén, unpubl.). The PrESTs are used for generation of validated antibodies for the creation of the Human Protein Atlas (www.proteinatlas.org) portal with tissue profiles of more than 2600 human proteins in normal and cancer tissue (Uhlén et al. 2005).

However, in these earlier attempts to predict and exclude protein epitopes (Andrade et al. 2006), a window of 50 amino acid residues was used for the analysis. Since the analysis is based on a rather long sequence window, it will not show local sequence identity in the size corresponding to linear epitopes. A need therefore exists to complement the sequence similarity searches using a 50 amino acid window with an identity analysis using windows corresponding to the length of typical linear epitopes. Here, we have therefore extended these earlier studies with analysis of windows ranging from eight to 12 amino acid residues, including a comparative analysis of a heuristic method and a non-heuristic string comparison method. The results show that it is possible to find unique epitopes to a majority of the human proteins with relatively strict rules involving local sequence identity based on short peptide sequences.

## Results

### *Analysis of all possible epitopes using a sliding window approach*

A sliding window approach (Fig. 1) was used to get the maximum sequence identity of all possible fragments, or windows, of length eight, 10, and 12 amino acid residues in a protein to the rest of the human proteome. In Ensembl (Hubbard et al. 2007) version 43.36, there are 43,738 proteins. When a window size of 10 amino acid residues was used, each window had to be compared to ~21,700,000 other windows to cover the complete Ensembl protein set. The average protein length is 500

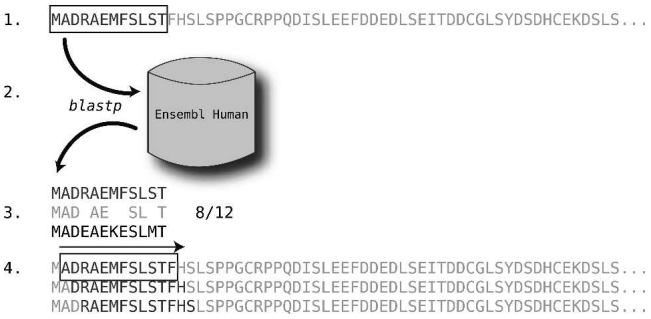


**Figure 1.** Principle of the sliding window method. Here, a window size of 12 amino acids is used. The first 12 amino acids of the protein sequence (1) is compared to all proteins in the human Ensembl database (2), here by using the blastp program. The best hit (highest number of identical amino acids between the hit and the query sequence) is retrieved from the results of the sequence similarity search. In this example, the best hit has eight amino acids identical to the query sequence, or 67% sequence identity (3). The window is moved one amino acid toward the C terminus of the protein (4). Steps 1–4 are repeated until the full length of the protein has been covered.

amino acids, which, for the 10 amino acid window, means on average 491 string comparisons per protein, resulting in a total of ~10,654,700,000 string comparisons per protein. The search can be terminated if a 100% match is found, but this will in most cases not happen. One way of making the string comparison is to use the Hamming distance method (Hamming 1950), where the distance between two strings of equal length is measured as the number of positions for which the corresponding symbols are different. The sequence identity will therefore be the Hamming distance subtracted from the length of the window. This method will return exact results, but is time consuming. An alternative, faster way of retrieving sequence similarity values is via the heuristic blastp program of the Basic Local Alignment Search Tool (BLAST) package (Altschul et al. 1990; 1997). To evaluate the sacrifice in accuracy inherent in a heuristic method like blastp, we performed a pilot study with 24 proteins for an initial comparison between the Hamming distance method and blastp using a 10 amino acid window. A 98% accuracy could be found for the heuristic method, with ~380 times shorter run time compared to the non-heuristic method (Table 1), suggesting that the heuristic method can be used for sequence similarity searches with short sliding windows.

### *Protein profile depending on method and window size*

As an example, the results for the human estrogen receptor (*ESR*) are shown in Fig. 2 using a string comparison method (Fig. 2A) and a heuristic method (Fig. 2B–D). The estrogen receptor has a length of 595 amino acid residues, giving 586 possible windows when using a 10 amino acid window. A comparison between the exact Hamming distance method (Fig. 2A) and the heuristic method with the same 10 residue windows (Fig. 2B) shows 14 differences. Both methods clearly show a region of the estrogen receptor between residues 184 and 252 with high (80%–100%) sequence identity to at least one other protein in the human proteome. This region should obviously be avoided when selecting antigens for this protein. In Figure 2B–D, a comparison of window size between eight and 12 residues is shown using the heuristic method. The analysis shows similar results, suggesting that any of these window sizes could be used to predict linear epitopes in the human proteome. The region between 184 and 252 residues with high sequence identity to another human protein is identified independently of window size.

### *A proteome-wide analysis of the epitope space*

The high accuracy and speed of the heuristic method allowed for a proteome-wide analysis of all possible linear epitopes involving all proteins of the human proteome (Fig. 3; Hubbard et al. 2007). The whole proteome analysis using a window size of eight amino acids showed that all human proteins have at least one window with five or more amino acids identical to other proteins within a continuous stretch of more than 25 amino acids (Fig. 3A). For 3900 proteins, there is not a single window with less than six out of eight identical amino acids to other proteins. For the 10 amino acid window analysis (Fig. 3B), few proteins (264 proteins) could be found that did not have at least one window with six or more amino acids identical to another window in the proteome, this within a stretch of more than 25 amino acids. Two thousand seven hundred sixty-seven proteins did not have a single window with less than six out of 10 amino acids identical to other proteins. For the 12 amino acid window (Fig. 3C), the critical limit is at seven out of 12 amino acids identical to other proteins, and stretches longer than 25 amino acids containing unique windows can be found for most proteins if allowing eight out of 12 amino acids identical to other proteins. If taking full-length proteins into consideration (Table 2), approximately one quarter of

**Table 1.** *Comparison of non-heuristic and heuristic algorithms for sliding window sequence identity searches on 24 human proteins*

| | Hamming distance | blastp |
|---|---|---|
| Algorithm type | Exact | Heuristic |
| Window size | 10 | 10 |
| Total number of windows | 18,196 | 18,196 |
| Average run time per protein | 29 h | 4.5 min |
| Number of windows with erroneous results | 0 | 356 |
| Accuracy | 100% | 98% |

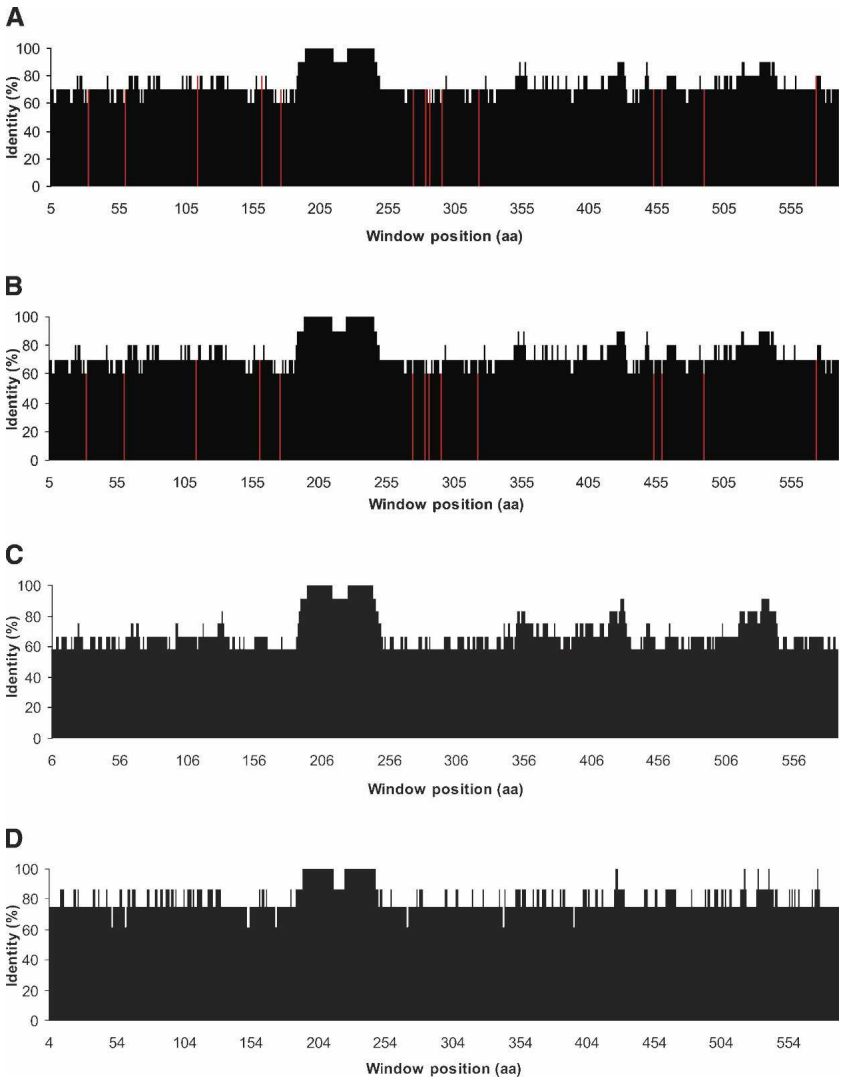


**Figure 2.** Protein profile for the estrogen receptor protein (*ESR*) using the Hamming distance method with 10 amino acids window (*A*) and blastp with 10 amino acids window (*B*), 12 amino acids window (*C*), and eight amino acids window (*D*). The bars marked in red in *A* and *B* indicate positions where the results from the exact method (*A*) and the heuristic method (*B*) are deviating. The *middle* position of each window on the protein is given on the *X*-axis.

all protein-coding genes have a product where no window has more than nine out of 12 amino acids identical to another protein. On the other hand, there are only a few proteins (637 proteins) without a single window with more than eight out of 12 amino acids identical to another protein, and no protein exists without a window with at least seven out of 12 amino acids identical.

### *The epitope space of proteins associated with well-characterized antibodies*

The 12 amino acid window using the heuristic method seems to give useful information about linear epitope predictions ranging from eight to 12 identical residues. In Figure 4, two epitope profiles for proteins used as a target in clinical diagnosis are displayed using the 12 amino acid window. For antigen selection on the protein encoded by *ERBB2*, a region around residue 250 should be avoided (11/12) as well as a large region between 700 and 975 with several subregions with 100% sequence identity (12/12). The epitope region for the therapeutically important trastuzumab (Cho et al. 2003), used to treat human breast cancer patients, is marked in Figure 4A, and the analysis shows that the antibodies recognize a region with low sequence identity to other human proteins. For the prostate specific antigen (*KLK3*) shown in Figure 4B, many regions with high (12/12) sequence identity are observed. The selection of antigens for generation of specific antibodies

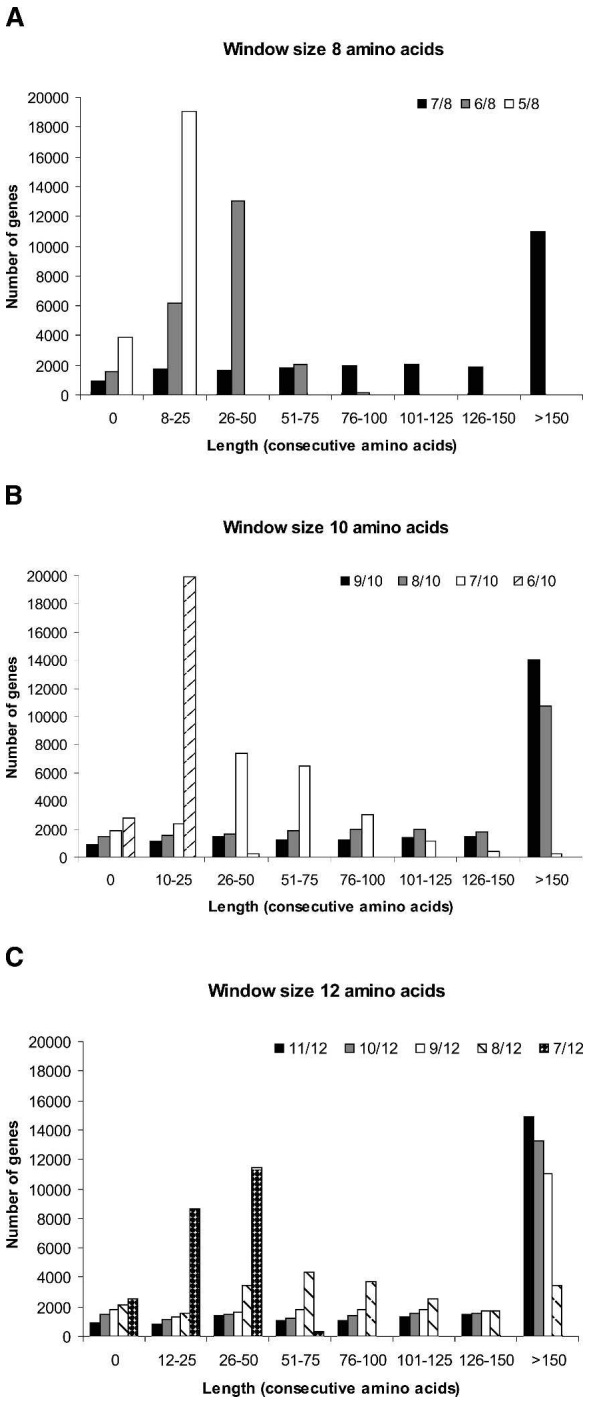


**Figure 3.** The linear epitopes of the human proteome space. For each window size, the longest consecutive amino acid stretch with all windows under a threshold value (e.g., no more than six out of eight amino acid residues identical to a protein from another gene), was determined for each of the 22,983 human genes in Ensembl. The maximum consecutive length found for the proteins encoded by each gene was selected as representative for that gene. The number of human genes (*Y*-axis) for each category of maximum consecutive length (*X*-axis) is presented for window sizes of *A*. Eight amino acids (threshold values 5, 6, and 7 identical amino acids). (*B*) Ten amino acids (threshold values 7, 8, and 9 identical amino acids). (*C*) Twelve amino acids (threshold values 7, 8, 9, 10, and 11 identical amino acids).

**Table 2.** *Number of genes with no window over threshold*

| | Number of identical residues | | | | | | |
|---|---|---|---|---|---|---|---|
| Window size | 11 | 10 | 9 | 8 | 7 | 6 | 5 |
| ***12*** | 11,896 | 9307 | 5786 | 637 | 0 | 0 | 0 |
| ***10*** | na | na | 10,208 | 4683 | 107 | 0 | 0 |
| ***8*** | na | na | na | na | 5786 | 12 | 0 |

Number of human genes coding for at least one protein where not a single window in the protein has more than the specified number of amino acid residues identical to proteins from other genes. A total of 43,738 proteins have been analyzed, corresponding to 22,983 genes. "na" = not applicable.

to the protein encoded by *KLK3* must therefore be done in a precise manner to avoid these regions. The epitope for commercial antibodies specifically recognizing free prostate specific antigen (Piironen et al. 1998), marked in Figure 4B, is in a short region with relatively low sequence identity to other proteins in the human proteome.

### *Local versus global sequence identity*

It is interesting to compare the sequence identity analysis using a longer (global) 50 amino acid window compared to a shorter (local) 12 amino acid window. Figure 5 shows the epitope profile for the leukocyte common antigen protein (*PTPRC*). Regions with high sequence identity are detected independently of window size. However, for regions with lower global similarity, the higher resolution achieved using a 12 amino acid window will allow for detection of local regions with high sequence identity to other proteins, which should be avoided in antigen selection for generation of protein-specific affinity reagents.

## Discussion

We have used a heuristic method to predict the local sequence identity of all human proteins based on a comparison using eight, 10, and 12 amino acid sliding windows for each protein against all other proteins of the human proteome. The results from the 12 amino acid window analysis show that for many of the human proteins, it is possible to select regions larger than 150 amino acids with no local sequence identity higher than nine out of 12 amino acids (Fig. 3C). The analysis also shows that with this criterion, at least one specific epitope can be found for 90% of the human proteins. For the majority of the proteins (88%), it is possible to find continuous stretches of more than 50 amino acids, to be used as antigens for generation of protein-specific antibodies.

The results presented are encouraging for efforts to try to generate antibodies without cross-reactivity toward all the proteins encoded by the human genome. However, cross-reactivity is a complex phenomenon related to the ability of a particular antibody to adapt several

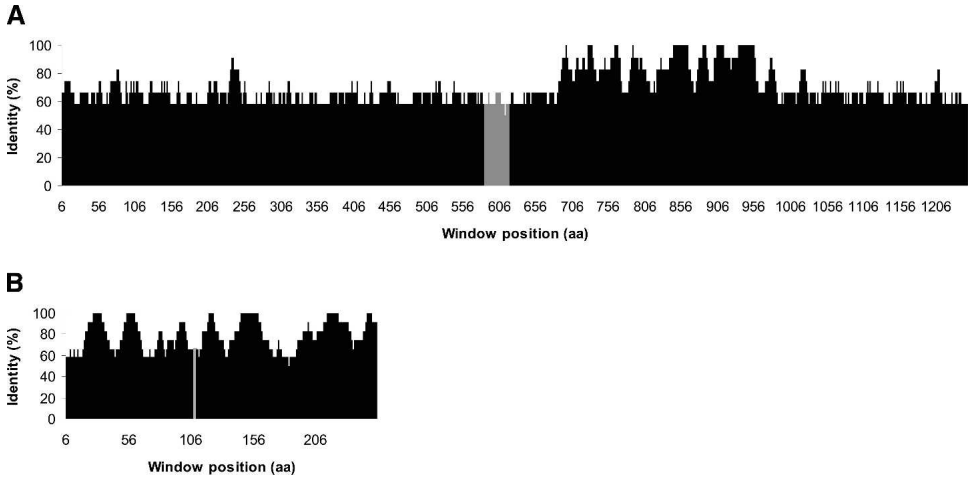


**Figure 4.** Examples of sequence profiles (12 amino acids window) for proteins with well-characterized, protein-specific antibodies used in clinical diagnostics. The window position given is the *middle* position for the window on the protein. The gray lines indicate where the protein-specific antibody binds to the protein. (*A*) Receptor tyrosine-protein kinase erbB-2 protein (*ERBB2*). The binding site of the antibody has been structurally determined (Cho et al. 2003). (*B*) Prostate-specific antigen protein (*KLK3*). The antibodies recognizing this epitope are known to have specific binding to free protein (Piironen et al. 1998).

conformations (James et al. 2003) as well as to the intrinsic conformational polyspecificity of antibodies (Bonnycastle et al. 1996), in particular, of antibodies produced against synthetic peptides or linear epitopes on proteins. In fact, peptide-binding antibodies have been shown to be cross-reactive with many members of phage-displayed libraries (Bonnycastle et al. 1996). These results emphasize the need to perform whole proteome analysis, as presented here, to avoid regions of high sequence identity to other proteins when selecting epitopes for antibody generation, although the level of cross-reactivity for a particular antibody must be subsequently validated in an application-specific manner. In this context, it is important to point out that the protein targets might have different folding states in various applications. For example, proteins are often partly or fully denaturated in Western blots, immunohistochemistry, or confocal microscopy, and in this way exposing different epitopes in the assay. Most validation assays are also semi-quantitative and context-dependent and the results will be influenced by choice of tissue and sample

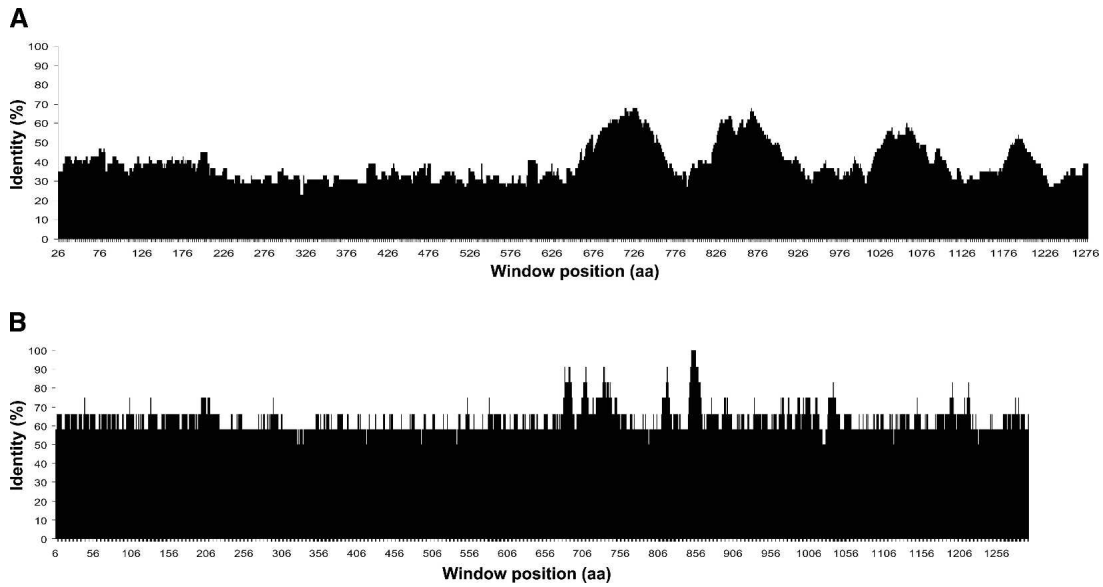


**Figure 5.** Comparison of a global sequence profile (window size 50 amino acids) (*A*) and a local sequence profile (window size 12 amino acids) (*B*) for the leukocyte common antigen protein (*PTPRC*).

preparation methods. The incredible dynamic range of at least $10^{10}$ of proteins in human body fluids and tissues (Anderson and Anderson 2002) makes interpretations even more difficult. This problem enforces the need to define standards for affinity reagent validation, and points to the need for standard operating procedures to carry out such validation.

Linear epitopes have shown to be typically six to nine residues (Rodda et al. 1986; Dunn et al. 1999; Fleury et al. 2000), and it is, of course, relevant to avoid stretches of amino acids in this length bracket with identity to other human proteins. However, our analysis shows that no antigen over 25 residues can be found if no stretch of eight residues in the antigen is allowed to contain more than five identical residues (Fig. 3A). This is, of course, due to the size of the human proteome, suggesting that most possible pentameric peptides are present. We have therefore decided to aim for a more realistic objective, such as to avoid 8 mers with a sequence identity >7 or 10 mers with an identity >8 or 12 mers with an identity >9.

Exact string comparison is very time consuming when searching large databases. Algorithms from the BLAST (Altschul et al. 1990; 1997) family use a much faster, heuristic (finding the shortest path) approach. The method scans the database for ''words'' of a predetermined length (a hit) with some minimum threshold parameter T, then extends the hit until the score falls below the maximum score yet attained minus some value X (Altschul et al. 1997). The setting of a higher value of T yields greater speed, but also an increased probability of missing weak similarities (Altschul et al. 1997). Thus, BLAST can make similarity searches very quickly because it takes shortcuts. However, the speed-versus-accuracy tradeoff could represent a constriction, especially in applications were high accuracy is required. We demonstrate that the results are deviating in 2% (356 windows) of the analyzed windows when comparing the results of the heuristic blastp to the non-heuristic Hamming distance (Hamming 1950) using a 10 amino acid sliding window on a test set of 24 proteins (18,196 windows) (Table 1). This is a consequence of the speed-versus-accuracy tradeoff inherent in the BLAST implementations. For the main purpose of this work, to estimate the epitope space of the human genome, this sacrifice has little influence on the results. However, when selecting antigens on a single protein level, the underlying similarity data and profiles along a protein will not be fully accurate, and potentially result in that a region with a similarity score above threshold is selected. An alternative would be to do the computationally intensive non-heuristic grid based similarity analysis of the human proteome once and for all, and construct a database of the results. As long as the Ensembl database is continuously updated, this is not a practical option.

The study presented here is focusing on the selection of suitable antigens for generating antibodies with low cross-reactivity to proteins from the same species. Since more and more genome sequences are being determined and released to the public (Kneale and Kennard 1984; Bilofsky et al. 1986; Tateno and Gojobori 1997), similar approaches can be used to generate antibodies toward proteins from other species than humans, such as mouse, rat, dog, or primates. For efforts to generate antibodies toward infectious agents, such as bacteria or viruses, it might also be relevant to use similar bioinformatics approaches to analyze possible antigens to avoid cross-reactivity to host proteins. It might be appropriate to collect all relevant pathogen proteome sequences in a comparative database to facilitate the selection of antigens to yield protein-specific antibodies.

It is wise to complement the information from the short window similarity searches with the results from a larger sliding window (50 amino acid residues) sequence similarity search to get a more global view of the proteins. With the global view, highly sequence-similar domains with high probability of structural resemblance and risk of cross-reactive conformational epitopes can be avoided, while the local view makes it possible to avoid regions with high risk of cross-reactive linear epitopes. We plan to release the results of the 10 amino acid sliding window analysis on the whole proteome as part of the next version of the Human Protein Atlas portal (www.proteinatlas.org). The results can thus be accessed for all human genes/proteins without password protection or other restrictions.

In conclusion, we here show the use of a whole proteome analysis method to select regions suitable for antibody-based proteomics efforts. A heuristic method was shown to allow for a grid-based analysis of the whole proteome, and a linear epitope prediction based on a sliding 8, 10, and 12 amino acid window was computed for all human proteins. The analysis shows that it should be possible to make specific antibodies to a large majority of all human proteins.

## Materials and Methods

### *Ensembl protein sequences*

The protein sequences were retrieved from the Homo_sapiens.NCBI36.43.pep.all.fa file in Ensembl (Hubbard et al. 2007), version 43.36. This file contains 43,738 protein sequences, corresponding to 22,983 genes.

### *Non-heuristic string comparison (Hamming distance)*

A symbolic base approach is applied when using the Hamming distance (Hamming 1950) method for protein similarity searches. The query protein segment of a given window size $N$ is considered as ''string1'' and the reference (the proteome

database that will be analyzed) is divided into all possible stretches of length *N*, each considered as ''string2.'' The comparison is performed between string1 and each possible string2, at the same string position, counting the number of matching symbols *X*. The similarity (%) is *X/N* *100.

### *Heuristic string comparison (blastp)*

The blastp program of the BLAST package (version 2.2.10) (Altschul et al. 1990, 1997) was run with the following parameters: -F F -g F -W 2 -e 10,000. In this way, the low complexity filter was turned off and all alignments were ungapped. The word size was set to two amino acids to increase the sensitivity of the search. For the same reason, the maximum allowed expectation value was set high (10,000).

### *The sliding window algorithm*

The sliding window algorithm was written in Perl programming language with the window size as input. Hits to proteins from the same gene as the query protein (splice variants) were excluded from the result. The sequence identity (%) was calculated as the number of identical amino acids between the query and the hit, divided by the window size, and finally multiplied by 100.

### *Grid*

The Nordugrid (Smirnova et al. 2003) infrastructure with about 600 processors was used for a grid-based implementation of blastp together with the sliding window algorithm (Andrade et al. 2006). The set up for running the Hamming distance method for sequence identity was identical, with the blastp Perl implementation replaced by a Hamming distance Perl implementation.

## Acknowledgments

We are grateful to Fredrik Pontén, Sophia Hober, Erik Björling, and Caroline Kampf for useful comments and advice. This work was supported by grants from the Knut and Alice Wallenberg Foundation.

## References

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

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

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

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

Atassi, M.Z. 1975. Antigenic structure of myoglobin: The complete immunochemical anatomy of a protein and conclusions relating to antigenic structures of proteins. *Immunochemistry* **12:** 423–438.

Barlow, D.J., Edwards, M.S., and Thornton, J.M. 1986. Continuous and discontinuous protein antigenic determinants. *Nature* **322:** 747–748.

Bilofsky, H.S., Burks, C., Fickett, J.W., Goad, W.B., Lewitter, F.I., Rindone, W.P., Swindell, C.D., and Tung, C.S. 1986. The GenBank genetic sequence databank. *Nucleic Acids Res.* **14:** 1–4.

Blythe, M.J. and Flower, D.R. 2005. Benchmarking B-cell epitope prediction: Underperformance of existing methods. *Protein Sci.* **14:** 246–248.

Bonnycastle, L.L., Mehroke, J.S., Rashed, M., Gong, X., and Scott, J.K. 1996. Probing the basis of antibody reactivity with a panel of constrained peptide libraries displayed by filamentous phage. *J. Mol. Biol.* **258:** 747–762.

Cho, H.S., Mason, K., Ramyar, K.X., Stanley, A.M., Gabelli, S.B., Denney Jr., D.W., and Leahy, D.J. 2003. Structure of the extracellular region of HER2 alone and in complex with the Herceptin Fab. *Nature* **421:** 756–760.

Dunn, C., O'Dowd, A., and Randall, R.E. 1999. Fine mapping of the binding sites of monoclonal antibodies raised against the Pk tag. *J. Immunol. Methods* **224:** 141–150.

Fack, F., Hugle-Dorr, B., Song, D., Queitsch, I., Petersen, G., and Bautz, E.K. 1997. Epitope mapping by phage display: Random versus gene-fragment libraries. *J. Immunol. Methods* **206:** 43–52.

Fleury, D., Daniels, R.S., Skehel, J.J., Knossow, M., and Bizebard, T. 2000. Structural evidence for recognition of a single epitope by two distinct antibodies. *Proteins* **40:** 572–578.

Greenbaum, J.A., Andersen, P.H., Blythe, M., Bui, H.H., Cachau, R.E., Crowe, J., Davies, M., Kolaskar, A.S., Lund, O., Morrison, S., et al. 2007. Towards a consensus on data sets and evaluation metrics for developing B-cell epitope prediction tools. *J. Mol. Recognit.* **20:** 75–82.

Hamming, R. 1950. Error detecting and error correcting codes. *Bell Syst. Tech. J.* **26:** 147–160.

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

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

Hubbard, T.J., Aken, B.L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T., et al. 2007. Ensembl 2007. *Nucleic Acids Res.* **35:** D610–D617. doi: 10.1093/nar/gkl996.

International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. *Nature* **431:** 931–945.

James, L.C., Roversi, P., and Tawfik, D.S. 2003. Antibody multispecificity mediated by conformational diversity. *Science* **299:** 1362–1367.

Kneale, G.G. and Kennard, O. 1984. The EMBL nucleotide sequence data library. *Biochem. Soc. Trans.* **12:** 1011–1014.

Lando, G., Berzofsky, J.A., and Reichlin, M. 1982. Antigenic structure of sperm whale myoglobin. I. Partition of specificities between antibodies reactive with peptides and native protein. *J. Immunol.* **129:** 206–211.

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

Odorico, M. and Pellequer, J.L. 2003. BEPITOPE: Predicting the location of continuous epitopes and patterns in proteins. *J. Mol. Recognit.* **16:** 20–22.

Piironen, T., Villoutreix, B.O., Becker, C., Hollingsworth, K., Vihinen, M., Bridon, D., Qiu, X., Rapp, J., Dowell, B., Lövgren, T., et al. 1998. Determination and analysis of antigenic epitopes of prostate specific antigen (PSA) and human glandular kallikrein 2 (hK2) using synthetic peptides and computer modeling. *Protein Sci.* **7:** 259–269.

Rodda, S.J., Geysen, H.M., Mason, T.J., and Schoofs, P.G. 1986. The antibody response to myoglobin–I. Systematic synthesis of myoglobin peptides reveals location and substructure of species-dependent continuous antigenic determinants. *Mol. Immunol.* **23:** 603–610.

Smirnova, O., Eerola, P., Ekelöf, T., Ellert, M., Hansen, J.R., Konstantinov, A., Kónya, B., Nielsen, J.L., Ould-Saada, F., and Wäänänen, A. 2003. The NorduGrid architecture and middleware for scientific applications. In *Computational Science - ICCS 2003: International Conference, Melbourne, Australia and St. Petersburg, Russia, June2–4, 2003. Proceedings, Part 1* (eds. P.M.A. Sloot et al.), pp. 264–273. Springer, NY.

Tateno, Y. and Gojobori, T. 1997. DNA data bank of Japan in the age of information biology. *Nucleic Acids Res.* **25:** 14–17.

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

Uhlén, M. and Pontén, F. 2005. Antibody-based proteomics for human tissue profiling. *Mol. Cell. Proteomics* **4:** 384–393.

Uhlén, M., Björling, E., Agaton, C., Szigyarto, C.A., Amini, B., Andersen, E., Andersson, A.C., Angelidou, P., Asplund, A., Asplund, C., et al. 2005. A human protein atlas for normal and cancer tissues based on antibody proteomics. *Mol. Cell. Proteomics* **4:** 1920–1932.

Van Regenmortel, M.H.V. 1996. Mapping epitope structure and activity: From one-dimensional prediction to four-dimensional description of antigenic specificity. *Methods* **9:** 465–472.

Van Regenmortel, M.H. 2006. Immunoinformatics may lead to a reappraisal of the nature of B-cell epitopes and of the feasibility of synthetic peptide vaccines. *J. Mol. Recognit.* **19:** 183–187.

Vyas, N.K., Vyas, M.N., Chervenak, M.C., Bundle, D.R., Pinto, B.M., and Quiocho, F.A. 2003. Structural basis of peptide-carbohydrate mimicry in an antibody-combining site. *Proc. Natl. Acad. Sci.* **100:** 15023–15028.