8. discoveryBASE.Virtual screening for disease associated proteins

Ronald Rapberger, Martin Haiduk, Paul Perco, Arno Lukas, Christian Siehs and Bernd Mayer
[1] Emergentec biodevelopment Rathausstrasse 5/3
A-1010 Vienna Austria
[2] Institute for Biomolecular Structural Chemistry, University of Vienna Dr. Bohrgasse 6 A-1030 Vienna, Austria

discoveryBASE is a knowledge driven, virtual screening workflow for fast and reliable identification of (disease-specific) proteins with strong antigenicity profiles (B-cell epitopes). The process combines identification of candidate proteins, identification of accessible, linear B-cell epitopes on the candidates, and experimental verification of identified epitopes on the disease-specific proteins. **Keywords: genomics, proteomics, machine learning, knowledge driven**

Knowledge driven drug discovery is an emerging field in computational biology and currently enjoying great popularity.
Identification of target proteins and their respective target sites contributing to certain disease progression constitutes the basis towards development of new therapeutics.
We have realized a knowledge driven, virtual screening workflow, discoveryBASE, for identification of candidate proteins exhibiting the following properties:
The proteins are disease specific, they are relevant in the cellular context, and they exhibit distinct solvent accessible domains, which show strong antigenicity profiles (B-cell epitopes).

The process combines identification of candidate proteins (Targeting), identification of accessible, linear B-cell epitopes on the candidates (Profiling), and experimental verification of identified epitopes on the disease-specific proteins (Validation).

The screening platform (Targeting) starts with raw genome / proteome data, either representing the full proteome of e.g. a pathogen, the full, presently annotated human proteome, or refined data from differential gene expression / proteomics analyses. Public domain, as well as proprietary information are used at this candidate protein selection stage.
A range of bioinformatics is applied to elucidate the regulatory context of the proteins under study: promoter analysis using our in-house developed program GAnalyze identifies transcription factor binding motifs common to a set of proteins thus indicating their transcriptional co-regulation, regulatory protein networks reveal the metabolic context of proteins (i.e. the pathways these proteins contribute to / are part of), molecular function, biological process and cellular component are derived from functional characterizations (gene ontologies) as well as cellular location predictors.

Further explorative and statistical testing is applied to refine candidate protein selection. The result of this targeting step is a list of candidate proteins which are associated with the disease area of interest, are relevant in the cellular context, and are located at sites amenable to a humoral immune response i.e. either located on the cell surface or secreted.

In the next step the list of candidate proteins is analyzed on the level of structure, solvent accessibility, and antigenicity, focusing on the identification of B-cell epitopes. These antigenic determinants on proteins play a major role in particular in the field of cancer, autoimmune disorders, transplantation, and infectious diseases. A variety of therapeutic applications as vaccines and monoclonal antibodies, as well as major diagnostic technologies fundamentally depend on the presence of antigenic determinants.

Identification of the latter is therefore of prior importance for the rational design of new diagnostics and therapeutics respectively.

The central element of protein antigenicity profiling with discoveryBASE is our B-cell epitope prediction routine E-Score. This algorithm represents a mathematical model derived from statistical and bio-computational analysis of a set of over 23.000 experimentally verified antigenic determinants as reference, and is trained using exhaustive peptide property computation in a neural network classification setting. Besides classical physicochemical properties, novel linguistic descriptors based on reduction alphabets of the common amino acids have been derived to develop E-score. Exhaustive in silico validation experiments confirmed that the prediction quality of our E-Score classifier alone exceeds by far presently available routines used for B-cell epitope scoring. The reasons for these accurate results are the excellence of the experimental data used as reference and the performance of selected and combined epitope descriptors implying classical physicochemical as well as linguistic descriptor sets.

discoveryBASE unfolds its full predictive power by combining E-Score with additionally computed protein characteristics as solvent accessibility, localization of membrane spanning domains, post-translational modification signals and 2D/3D structure models. In the absence of experimentally generated protein structures (by the use of NMR or crystallography) and reliable ab initio threading results, the 2D structure and the solvent accessibility prediction form the basis for epitope selection. Thus for our approach the amino acid sequence information of the protein of interest is the only prerequisite.

The result of this profiling step is a list of candidate epitopes which are tested experimentally in the validation step via either ELISA-based screening, or via generation and reverse testing of antibodies. Antibodies reactive to derived epitopes are further tested for their diagnostic and therapeutic potential. To date discoveryBASE derived epitopes of pathogenic microbial (S.aureus, S. pneumoniae, P. aeruginosa, E. faecalis, S. epidermidis, C. albicans), viral (HIV, Influenza, SARS) and tumor-associated proteins (Melanoma, Ovarian) are validated experimentally.

In the case of melanoma-associated proteins we recently identified epitopes on melanoma associated endogenous retrovirus proteins using discoveryBASE. Experimental verification of these candidate sequences screening sera from melanoma patients identified peptide epitopes reactive to antibodies prevalent in melanoma sera. An immunodominant peptide was further analyzed utilizing samples from stage I – stage IV melanoma patients and sera from healthy subjects. Statistically significant differences between ELISA signal distributions comparing early stage (I, II) and late stage (III, IV) patient sera with ELISA signals derived from sera of healthy subjects were identified.
Sensitivity of 90 % at a specificity of 70 % indicated the potential of this novel marker. In particular the reactivity found in sera from stage I and II melanoma patients gives hope to utilize this candidate peptide to support diagnostics already at early stages of the disease.

Further experimental screening results of ongoing validation projects are on the way and expected in the near future.

Recapitulating it could be shown that identification of antigenic (B-cell) determinants on proteins applying our bioinformatics suite discoveryBASE is a fast, efficient, and validated approach.

Successful application of discoveryBASE was confirmed previously by identification of epitopes on melanoma-associated proteins with one predicted epitope as novel biomarker for early stage disease detection. Further proof of concept studies are on the way.