Bioinformatics, 31(11), 2015, 1854–1856
doi: 10.1093/bioinformatics/btv068
Advance Access Publication Date: 1 February 2015
Applications Note

OXFORD

Gene expression

Expitope: a web server for epitope expression

Kerstin Haase[1], Silke Raffegerst[2,3], Dolores J. Schendel[2,3] and Dmitrij Frishman[1,4,5,*]

[1]Department of Genome Oriented Bioinformatics, Wissenschaftszentrum Weihenstephan, Technische Universität München, 85354 Freising, Germany, [2]Immune Monitoring Group, Helmholtz Zentrum Munich, Institute of Molecular Immunology, 81377 München, Germany, [3]Medigene Immunotherapies GmbH a Subsidiary of Medigene AG, 82152 Planegg, Germany, [4]Helmholtz Zentrum Munich, Institute of Bioinformatics and Systems Biology, 85764 Neuherberg, Germany and [5]St Petersburg State Polytechnical University, St Petersburg, 195251, Russia

*To whom correspondence should be addressed.

Associate Editor: Janet Kelso

Received on November 27, 2014; revised on January 21, 2015; accepted on January 27, 2015

Abstract

Motivation: Adoptive T cell therapies based on introduction of new T cell receptors (TCRs) into patient recipient T cells is a promising new treatment for various kinds of cancers. A major challenge, however, is the choice of target antigens. If an engineered TCR can cross-react with self-antigens in healthy tissue, the side-effects can be devastating. We present the first web server for assessing epitope sharing when designing new potential lead targets. We enable the users to find all known proteins containing their peptide of interest. The web server returns not only exact matches, but also approximate ones, allowing a number of mismatches of the users choice. For the identified candidate proteins the expression values in various healthy tissues, representing all vital human organs, are extracted from RNA Sequencing (RNA-Seq) data as well as from some cancer tissues as control. All results are returned to the user sorted by a score, which is calculated using well-established methods and tools for immunological predictions. It depends on the probability that the epitope is created by proteasomal cleavage and its affinities to the transporter associated with antigen processing and the major histocompatibility complex class I alleles. With this framework, we hope to provide a helpful tool to exclude potential cross-reactivity in the early stage of TCR selection for use in design of adoptive T cell immunotherapy.

Availability and implementation: The Expitope web server can be accessed via http://webclu.bio.wzw.tum.de/expitope.

Contact: d.frishman@wzw.tum.de

1 Introduction

In adoptive immunotherapy, engineered T cell receptors (TCRs) are introduced into natural patient cytotoxic T lymphocytes. After this treatment, the T cells recognize a specific tumor antigen and will thus start to target cancer cells. It is vital for the success of the therapy that the antigen is only expressed in cancer cells or non-vital tissue, otherwise the effects can be devastating for the patients.

Recent studies showed that not only the expression of the direct target has to be examined across all vital tissues, but also approximate sequences have to be considered, as TCRs are not perfectly exact in their epitope choice. In one study, Morgan *et al.* (2013) reported cross-recognition of a MAGE-A3 TCR with a MAGE-A12 epitope that was later found to be expressed in the brain. The MAGE-A12 epitope had one mismatch when compared with the initial target of the study, but was apparently recognized by the TCR and the treatment was fatal for some patients (Morgan *et al.*, 2013). In another case, Linette *et al.* (2013) used a different MAGE-A3-specific TCR that was found to show cross-recognition of an epitope present in titin, a protein expressed in the heart, although the titin-associated epitope had four mismatches compared with the original

MAGE-A3 epitope. Nevertheless, titin was targeted by the engineered T cells and the patients suffered cardiac arrest (Linette *et al.*, 2013).

In order to see potential off-target recognition when designing new lead targets, until now one needed to search protein databases for approximate hits and than evaluate each hit for its potential to be an epitope. Our Expitope web server combines all these searches and evaluation in one place and even reports the expression of the associated transcripts in all vital human tissues to facilitate TCR selection.

2 Methods

2.1 RNA-seq database

As the basis for the epitope expression analysis we set up a database containing RNA-seq results from multiple different healthy tissues. A very comprehensive set can be found in the Illumina Human Body Map (GEO identifier: GSE30611), which provides 16 normal tissues from unrelated donors. To provide a positive control for most of the cancer antigens, we also included expression values for three cancerous cell lines from the ENCODE project (GEO identifier: GSM758575, GSM981253 and GSM958749) (ENCODE Project Consortium, 2011). To obtain expression values for all annotated transcripts, we used GenCodeV19 (Harrow et al., 2006) and counted the reads per every exon, so we could sum up the coverage over all alternative transcripts. As raw read counts are not easily comparable between different samples due to different library sizes, we normalized the counts to fragments per kilobase of exon per million fragments mapped values with the bamutils tool count (Breese and Liu, 2013).

As the brain constitutes one of the most vital organs, for which cross reaction has to be excluded very vigorously, we integrated additional brain isoform expression data published by Wang *et al.* (2008). They analysed the transcriptomes of 15 different human tissues, among them six individual brain samples, and provide the RPKM (reads assigned per kilobase of transcript per million mapped reads) values for 23 115 Ensembl gene identifier. As the integration of RNA-seq data into the database is fully automated, it is easy to add additional tissues or cell types on demand.

2.2 Epitope lookup

Our server requires an epitope (a string of amino acids in one letter code) and a number of allowed mismatches (integer value) as input.

A search for all occurrences of the given epitope is implemented against the entire protein sequence database of the National Center for Biotechnology Information, including all annotated isoforms. All matches with zero up to the defined number of mismatches are reported and the corresponding protein identifiers stored. All obtained protein identifiers from entries of interest are mapped to Ensembl transcript identifiers via a lookup file downloaded from biomart (Smedley *et al.*, 2009).

The set of transcript IDs is then used to query the database of expression values in all tissues, as described earlier. These results are presented to the user in form of a table, which additionally contains the exact epitope found in a certain protein and its sequence position.

2.3 Output ranking

2.3.1 Combined score

To sort the potentially long list of results with regard to their real potential to function as an epitope, we apply a scoring function as proposed by Keşmir *et al.* (2002). It combines the probability that a given peptide is cleaved from its original sequence, transported to the endoplasmic reticulum and bound by major histocompatibility complex (MHC) class I proteins. The resulting score Q is defined as

$$Q = \frac{P}{A_{\mathrm{TAP}} \times A_{\mathrm{MHC}}}$$

where P is the proteasomal cleavage probability, and the A-terms are affinities in IC_{50} values (dose of peptide which displaces 50% of a competitive ligand) to the transporter associated with antigen processing (TAP) and MHC complex.

2.3.2 Proteasomal cleavage prediction

To calculate the proteasomal cleavage probability, we used the program NetChop 3.1 (Keşmir *et al.*, 2002; Nielsen *et al.*, 2005). We ran the program on all current RefSeq protein entries and obtained a cleavage probability for every position. These values are stored in an additional database table to avoid executing NetChop for every web server query. We are using the prediction method 'C-term 3.0' which is a neural network trained on a database containing 1260 publicly available MHC class I ligands.

2.3.3 TAP affinity prediction

Peters *et al.* (2003) have established a 9×20 matrix, $\mathrm{mat}_{i,j}$, that contains for each amino acid at every possible epitope position (of length nine) a $\log(IC_{50})$ value which can be summed up to obtain an IC_{50} value for the complete peptide. When evaluating their method, the authors observed that the best concordance to experimental values is achieved, when taking precursor peptides into account, i.e. instead of the initial nonamer they calculated the affinity for an N-terminal elongated sequence. In order to use this approach with epitopes of fixed length provided by the users, we modified the established formula to work without precursor sequences. Hence, only the IC_{50} values for the C-terminal residue as well as a weighted sum of the three N-terminal amino acids are used for the scoring.

2.3.4 MHC binding prediction

For the affinity prediction of the epitopes to the MHC for a large range of human leukocyte antigen (HLA) alleles, we integrated NetMHC 3.0 (Nielsen *et al.*, 2003; Lundegaard *et al.*, 2008a,b) into our web server. It offers artificial neural networks trained on 55 different MHC alleles and returns the affinity of a given peptide to the specified alleles in nM IC_{50} values. The Expitope server reports the exact IC_{50} values predicted by NetMHC for every MHC allele that was selected in the query, but only the best (lowest) is used in the calculation of the combined score Q.

3 Conclusion

To test the capability of Expitope, we investigated a previous TCR gene therapy in which unanticipated cross-recognition of healthy tissues led to patient deaths. We used the target that Linette *et al.* (2013) had engineered in their study and allowed for up to four mismatches. Cross-recognition of titin was identified by our web server and although the sequence has four mismatches, the predicted affinity to MHC allele A0101 was even higher for the titin antigen than that for the original MAGE-A3 peptide. Although we would like to remind all users that the predictions are only to be used as a first instance of TCR selection and need to be validated experimentally before used in therapy, we expect our Expitope web server to be a useful tool for recognizing potential cross-reactivity in the early stage of TCR selection and designing adoptive T cell immunotherapies.

Funding

K.H. was supported by a Research Scholarship under the Bavarian Elite Aid Act (BayEFG).

Conflict of Interest: none declared.

References

Breese,M.R. and Liu,Y. (2013) NGSUtils: a software suite for analyzing and manipulating next-generation sequencing datasets. *Bioinformatics*, **29**, 494–496.

Harrow,J. *et al.* (2006) GENCODE: producing a reference annotation for ENCODE. *Genome Biol.*, **7** (Suppl. 1), 1–9.

Keşmir,C. *et al.* (2002) Prediction of proteasome cleavage motifs by neural networks. *Protein Eng.*, **15**, 287–296.

Linette,G.P. *et al.* (2013) Cardiovascular toxicity and titin cross-reactivity of affinity-enhanced T cells in myeloma and melanoma. *Blood*, **122**, 863–871.

Lundegaard,C. *et al.* (2008a) NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8–11. *Nucleic Acids Res.*, **36** (Suppl. 2), W509–W512.

Lundegaard,C. *et al.* (2008b) Accurate approximation method for prediction of class I MHC affinities for peptides of length 8, 10 and 11 using prediction tools trained on 9mers. *Bioinformatics*, **24**, 1397–1398.

Morgan,R.A. *et al.* (2013) Cancer regression and neurological toxicity following anti-MAGE-A3 TCR gene therapy. *J. Immunother.*, **36**, 133–151.

Nielsen,M. *et al.* (2003) Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Sci.*, **12**, 1007–1017.

Nielsen,M. *et al.* (2005) The role of the proteasome in generating cytotoxic T-cell epitopes: insights obtained from improved predictions of proteasomal cleavage. *Immunogenetics*, **57**, 33–41.

Peters,B. *et al.* (2003) Identifying MHC class I epitopes by predicting the tap transport efficiency of epitope precursors. *J. Immunol.*, **171**, 1741–1749.

Smedley,D. *et al.* (2009) BioMart - biological queries made easy. *BMC Genomics*, **10**, 22.

The ENCODE Project Consortium. (2011) A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.*, **9**, e1001046.

Wang,E.T. *et al.* (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**, 470–476.