Molecular Immunology 48 (2011) 577–585

Contents lists available at ScienceDirect

Molecular Immunology

journal homepage: www.elsevier.com/locate/molimm

A generic approach to evaluate how B-cell epitopes are surface-exposed on protein structures

Virginie Lollier*, Sandra Denery-Papini, Colette Larré, Dominique Tessier

UR1268 BIA (Biopolymers, Interactions, Assemblies), INRA, 44300 Nantes, France

ARTICLE INFO

Article history:
Received 21 September 2010
Accepted 24 October 2010
Available online 15 December 2010

Keywords:
Accessibility
Bioinformatics
Epitope
Prediction
Protein surface
Protrusion index

ABSTRACT

Methods that predict antibody epitopes could help to promote the development of diagnostic tools, vaccines or immunotherapies by affecting the epitope binding of antibodies during an immunological response to antigens.

It is generally assumed that there is a direct relationship between antibody accessibility to antigens and accessible surface of proteins. Based on this assumption, prediction systems often includes solvent accessibility values calculated from the primary sequence of proteins or from their three dimensional structures as a predictive criterion. However, the current prediction systems seem weakly efficient in view of benchmark tests.

We were interested in evaluating how amino acids that have been experimentally identified as epitopic elements could differ from the rest of the antigenic molecule at the level of surface exposure, hence we assessed the average accessibility of epitopes. The approach used here utilises published epitopes deduced from numerous identification techniques, including sequence scanning and structure visualisation after crystallography, and it involves many types of antigens from toxins to allergens. Our results show that epitopic residues are not distributed among any specific Relative Surface Accessibility and Protrusion Index values and that, in some cases, epitopes cover the entire antigenic sequence.

These results led to the conclusion that the classification of known epitopes with respect to the experimental conditions used to identify them should be introduced before attempting to characterise epitopic areas in a generic way.

1. Introduction

Immunological responses are characterised by the neutralisation of exogenous proteins by specific antibodies. Exogenous proteins, also called antigens, can be derived from viral particles, toxins or allergens. Recognition and binding of an antigen by the antibody products of B-cells are driven by epitopes, which are portions of the antigen molecule. Numerous studies are being conducted to identify B-cell epitopes because the localisation of these specific areas is of clinical interest for the development of diagnostic tools, vaccines and immunotherapies. Such knowledge could also lead to a better understanding of immunological mechanisms, particularly in the field of allergic response, by probing risk factors and advancing preventive strategies.

The B-cell epitopes are classified into two categories: Continuous epitopes are composed of contiguous amino acids along the primary sequence, whereas discontinuous epitopes combine several shorter segments scattered along the sequence and brought together in spatial proximity when the protein is folded. This classification is tightly related to the different experimental techniques developed to locate epitopes on antigens. The ability of synthetic overlapping peptides to bind antibodies leads to the identification of continuous epitopes. Usually, such epitopes consist of peptides of around ten amino acids, but their minimum size remains poorly defined (Van Regenmortel, 2006). In the context of allergic response, the identification of critical amino acids for antibody-binding may be useful in the development of immunotherapy strategies by modifying B-cell epitopes to prevent IgE binding while preserving T cell epitopes (e.g., Cocco et al., 2003; Robotham et al., 2002).

Discontinuous epitopes can be identified by directed mutagenesis of the antigen followed by the measurement of the antibody's residual affinity (Swoboda et al., 2002). The creation of a random peptide library displayed at the surface of phages or yeasts followed by immunopanning, whereby antibodies bind to surface-exposed peptides that mimic the epitopes, is an alternative method to identify discontinuous epitopes known as the mimotopes technique (Levy et al., 2007; Prabakaran et al., 2006; Riemer et al., 2004). Nevertheless, the fine identification of discontinuous epitopes is generally achieved by microscopic techniques that resolve

* Corresponding author. Tel.: +33 240 675073; fax: +33 240 675025.
E-mail address: Virginie.Lollier@nantes.inra.fr (V. Lollier).

0161-5890/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.molimm.2010.10.011

antigen–antibody complexes, such as the electron-microscopy density map (e.g., Belnap et al., 2003), the NMR (e.g., Morgan et al., 2004; Naik et al., 2008) and X-ray crystallography (e.g., Padavattan et al., 2007). Epitopes are then deduced from these observations by identifying the points of contact between the two molecules at the atomic level, which usually encompasses between 15 and 22 amino acids (Rubinstein et al., 2008; Van Regenmortel, 1996).

The human body can produce around 10^8 to 10^9 different antibodies, due to the high number of potential variable domains encoded following genetic recombination. Consequently, the antibody repertoire contains an almost infinite capacity to bind various peptides with different affinities. Accurate prediction tools could therefore help by reducing the number of potential epitopes. Further, the development of experimental techniques increases our understanding of these epitopes and their binding characteristics. B-cell epitope-specific databases centralise this information and enable statistical studies to describe B-cell epitopes in a predictive manner. AntiJen (Toseland et al., 2005) mainly contains continuous epitopes whereas Epitome (Schlessinger et al., 2006) is composed of epitopes identified from the structural analysis of antigen–antibody complexes. The Immune Epitope DataBase (Vita et al., 2010) gathers data on both B and T cell epitopes that have been curated from several publicly available databases, the scientific literature and those directly submitted by authors. Epitopes collected in such databases are described by their amino acid sequences plus the residue positions within the antigen sequences, particularly for discontinuous epitopes.

The most cited characteristic of an epitope is its presence on the protein surface. It is generally calculated from the three dimensional structures by determining the solvent-accessible surface (SAS) of the residues or by examination of their protrusion index, an alternative to the influence of atomic radii parameter on surface computations (Pintar et al., 2002). The importance of surface exposure for antibody binding has been observed for the first time by Novotný et al. (1986) and Thornton et al. (1986) using lysozyme, myoglobin, cytochrome c and myohemerythrin. This surface exposure characteristic has even been integrated with experimental work as a confirmation of *in vitro* epitope identification methods (Dolimbek et al., 2008; López-Torrejón et al., 2007; Robotham et al., 2009) and as an element of the epitope localisation by the mimotope technique (Pacios et al., 2008; Tordesillas et al., 2009). The basic assumption that epitopes will be found in regions of proteins that have a high degree of exposure to the solvent was also an element taken into account for both continuous and discontinuous epitope predictions. Epitope predictions are based on either protein sequence or three dimensional structure. In this last case, SAS visualisation of the known or homology-modelled antigen structure is often associated with changes in the electrostatic potential and these two features enable to select candidate sequences for immunoassays or for directed mutagenesis to build less reactive molecules (García-Casado et al., 2003; Sordet et al., 2009). Moreover, the accessible characteristic of epitopic elements is one of those integrated into prediction servers based on 3D structures, such as Discotope (Haste Andersen et al., 2006) and Ellipro (Ponomarenko et al., 2008).

Methods for sequence-based predictions originally used a combination of different amino acid propensity scales, such as HPLC-derived hydrophobicity, solvent accessibility, flexibility, hydropathy and the propensity for adopting a certain secondary structure. However, a complete benchmark of the available scales from the AAindex database, including the surface accessibility propensity (Emini et al., 1985), has been achieved by Blythe and Flower (2005) and displays a weak predictive performance. As an illustration of these results, a recent work comparing experimental data with epitope predictions does not show a correlation between the continuous epitopes identified and the computed surface accessibility for the dengue virus 3 protein sequence (da Silva et al., 2009). Aiming to improve the accuracy of continuous B-cell epitope predictions, some other tools using learning approaches have since been developed to classify sequences from two types of datasets: one from epitope fragments and one from non-epitope fragments. Indeed, these methods slightly improve performance predictions compared to amino acid scales; however, one drawback is that they work in a black box without giving any consideration to the underlying molecular mechanisms. Furthermore, their non-epitope dataset construction collects randomly selected fragments from SWISS-PROT database or randomly constructed peptides (El-Manzalawy et al., 2008; Saha and Raghava, 2006). These true negative datasets cannot be validated but their content may influence the prediction accuracy. The incorporation of structural predictions about secondary structures and solvent accessibility into the learning techniques of support vector machines seems to improve slightly the outcome, particularly improving specificity at the expense of sensitivity (Sweredoski and Baldi, 2009). Unfortunately, this work requires entire protein sequences and as a consequence it has been done on a very limited dataset (14 well known antigens and HIV antigenic proteins).

The prediction approaches that have been applied to date to epitopes rely either on the individual visualisation of the three-dimensional structure of an antigen, which has a limited scope, or on the comparisons made between the epitope sequences from several antigens and those of randomly built non-epitopes with a weak accuracy. To circumvent the limits of these approaches, the present work describes the surface accessible features of continuous and discontinuous epitopes as compared to the remaining molecule based on an extended structural dataset without regard to the antigen type. This generic approach aims to assess the average exposure of epitopes and to determine whether this could act as a discriminating factor for prediction methods.

2. Material and methods

2.1. Data collection

Both continuous and discontinuous epitope sequences from natural peptides, identified by positive B-cell assays involving known antibody isotypes, have been extracted from the IEDB v 2.0 WEB site (August 2009) along with their associated PubMed identifiers. Continuous epitopes greater than 20 amino acids in length have been removed.

The antigen sequences have been retrieved by their protein identifiers from the entrez protein database (http://www.ncbi.nlm.nih.gov/protein). Their 3D structures were downloaded from PDB site (http://www.pdb.org) using UniprotKB (http://www.uniprot.org) cross-references. Only structures representing a single polypeptide chain, i.e., without any complexed agent masking the epitope sites, were conserved. When multiple structures were available for a given antigen, the longest one or one resolved by X-ray crystallography was preferred.

In some cases it was deemed necessary to examine articles in detail and check the location of the epitopes.

2.2. Surface calculations

Since the epitopes are described by their amino acid sequences in databases the antigen surfaces were characterised by the residue accessibility (Lee and Richards, 1971) and the residue protrusion index. The residue accessibility is represented by the Relative Solvent Accessibility (RSA), which is defined for each residue as the ratio of the sum of the Accessible Surface Areas (ASA) of its atoms to the maximal ASA that the residue would have in an unfolded and