Molecular Immunology 51 (2012) 304–309

Contents lists available at SciVerse ScienceDirect

Molecular Immunology

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

An analysis of B-cell epitope discontinuity

Ganesh N. Sivalingam[a,b], Adrian J. Shepherd[b,*]

[a] Research Department of Structural and Molecular Biology, Institute of Structural and Molecular Biology, University College London, Gower Street, London WC1E 6BT, UK
[b] Department of Biological Sciences, Institute of Structural and Molecular Biology, Birkbeck, University of London, Malet Street, London WC1E 7HX, UK

ARTICLE INFO

Article history:
Received 29 February 2012
Accepted 25 March 2012
Available online 20 April 2012

Keywords:
Antigen
Antibody
Epitope
Structural epitope
Functional epitope

ABSTRACT

Although it is widely acknowledged that most B-cell epitopes are discontinuous, the degree of discontinuity is poorly understood. For example, given that an antigen having a single epitope that has been chopped into peptides of a specific length, what is the likelihood that one of the peptides will span all the residues belonging to that epitope? Or, alternatively, what is the largest proportion of the epitope's residues that any peptide is likely to contain? These and similar questions are of direct relevance both to computational methods that aim to predict the location of epitopes from sequence (linear B-cell epitope prediction methods) and window-based experimental methods that aim to locate epitopes by assessing the strength of antibody binding to synthetic peptides on a chip.

In this paper we present an analysis of the degree of B-cell epitope discontinuity, both in terms of the structural epitopes defined by a set of antigen–antibody complexes in the Protein Data Bank, and with respect to the distribution of key residues that form functional epitopes. We show that, taking a strict definition of discontinuity, all the epitopes in our data set are discontinuous. More significantly, we provide explicit guidance about the choice of peptide length when using window-based B-cell epitope prediction and mapping techniques based on a detailed analysis of the likely effectiveness of different lengths.

1. Introduction

1.1. B-cell epitope identification

It is widely recognized that knowledge about B-cell epitopes is important for the identification or design of therapeutic antibodies, and for gaining insights into vaccine effectiveness (Irving et al., 2001). Various methods may be used to determine – with varying degrees of accuracy – the location of B-cell epitopes, ranging from purely computational methods to X-ray crystallography. (For a useful survey of methods, see Ladner (2007).)

Whereas approaches such as X-ray crystallography and site-directed mutagenesis are capable of determining the location of B-cell epitopes with a high degree of accuracy, the efficacy of the methods we focus on here – computational methods for predicting the location of linear B-cell epitopes and short peptide mapping techniques – is somewhat uncertain. Nevertheless, these methods have an enduring appeal, as they are comparatively cheap and can be used as the basis for high-throughput screening, properties that more accurate methods do not possess.

A range of *computational methods* have been developed for predicting which of an antigen's residues are likely to form part of an epitope (El-Manzalawy and Honavar, 2010). In the absence of useful structural information about the antigen, predictions must be made using the primary amino-acid sequence alone. Typically a fixed-length profile is generated from a set of known examples and applied to a given antigen using a sliding window.

Such methods are primarily suited to find linear B-cell epitopes, i.e. epitopes that consist of a single more-or-less continuous segment from the primary sequence. But this begs the questions: How strict does the definition of "continuous" have to be? And what proportion of epitopes meet these requirements in practice?

Short peptide mapping involves the synthesis of relatively short overlapping peptides from the antigen of interest and measuring the extent to which they bind to a given antibody. The peptide may be in linear conformation, or constrained in some way to mimic, to some degree, the 3-dimensional conformation of that peptide in its natural (*in vivo*) structural context (Timmerman et al., 2009). Given an antigen of interest, it is up to the researcher to decide how to split it into individual peptides. In practice, experimentalists typically choose a fixed window size (peptide length) and shift that window by a fixed amount along the full length of the antigen sequence (maintaining a consistent degree of overlap). However, the window size and degree of shift can vary significantly between different experiments. For example, Geysen et al. (1984) chose a window of size six and shifted the window by a single position (hence an

Abbreviations: ASA, accessible surface area; PDB, Protein Data Bank.

* Corresponding author. Tel.: +44 20 7631 6886; fax: +44 20 7631 6803.
E-mail address: a.shepherd@mail.cryst.bbk.ac.uk (A.J. Shepherd).

0161-5890
http://dx.doi.org/10.1016/j.molimm.2012.03.030

overlap of five), whereas Behan et al. (1998) used a window of size 17 shifted by five residues (hence an overlap of 12). Peptides of up to 32 residues were used by Timmerman et al. (2007), but such large window sizes are exceptional.

Note that in this paper we deliberately exclude from consideration variations on these peptide-mapping approaches that model discontinuous epitopes by combining non-adjacent segments from a protein sequence. To be effective, such approaches generally require significant prior knowledge about the location of epitope residues – see, for example, the analysis of CD20 antibodies in Niederfellner et al. (2011).

Before considering whether these epitope prediction and small peptide mapping approaches have inherent limitations, it is essential to consider what is known about the properties of B-cell epitopes.

1.2. Properties of B-cell epitopes

There are various ways of defining what an "epitope" is (see Ladner, 2007), but probably the most widely used definition is that of a *structural epitope*. A structural epitope consists of the set of the antigen's amino-acid residues that are in direct contact with residues belonging to an antibody (the paratope).

Several fundamental properties of structural epitopes have been quantified in an analysis of 53 antigen–antibody complexes from the Protein Data Bank (PDB) (Berman et al., 2000) undertaken by Rubinstein et al. (2008). For example, the study concluded that approximately 75% of epitopes consist of 15–25 residues with a surface area of 600–1000 $Å^2$. They also partially quantified the degree to which B-cell epitopes are discontinuous. No epitopes in their data set were found to be strictly linear, i.e. composed of a single, continuous segment of the antigen's amino-acid sequence having all residues in direct physical contact with one or more antibody residues. Using a less strict criterion that allowed up to three non-contact residues to occur within a segment, the authors found that most epitopes consist between one and five segments, each containing one to six residues.

Whereas the definition of a structural epitope is widely used and easy to grasp, it is not necessarily the most relevant for the purpose of epitope mapping. On the one hand, some non-contact residues have been shown to induce conformational changes that affect antigen–antibody binding (Parry et al., 1990); on the other hand, it is widely recognized that, in general, only a subset of contact residues within an epitope make a significant contribution to the global binding energy (Novotny, 1991). These energetically important residues – which typically number between three and five, and which can be determined experimentally using site-directed mutagenesis (Benjamin and Perdue, 1996) – are commonly known as hot spot residues and collectively form a so-called *functional epitope*.

The properties of protein–protein interfaces in general have been widely characterized in the literature; a small number of hot-spot residues account for most of the binding energy (Bogan and Thorn, 1998) and are grouped in one or a few "hot regions" towards the centre of the interface (Keskin et al., 2004). However, whereas some authors assume there is nothing special about B-cell epitopes – indeed, the term epitope is sometimes used loosely to refer to any protein interface (see, for example, Ma et al., 2001) – this assumption may not be justified, as there are important differences between the binding characteristics of antigen–antibody complexes and those of other classes of complex. For example, Jackson found that serine protease–inhibitor complexes involve backbone interactions, whereas side-chain interactions dominate in antigen–antibody complexes (Jackson, 1999).

All things considered, we should expect antigen–antibody interfaces to be a special case. Whereas other protein–protein interfaces will typically have evolved cooperatively (with both partners making a complementary contribution), in antigen–antibody interfaces the antigen is either passive or actively evolves to resist the formation of the complex. Whereas other protein–protein interfaces are likely to be mature (established over significant periods of time), antigen–antibody interfaces are often transients (witness, for example, the short-lived effectiveness of most antibodies against the evolving influenza A virus (Wilson and Cox, 1990)).

1.3. The limitations of window-based methods

Some of the current limitations of window-based computational prediction methods and peptide mapping techniques have already been discussed elsewhere. Blythe and Flower (2005) demonstrated that simple sequence profiles based on a single propensity scale are little better than random at predicting the location of linear epitopes. For short peptide mapping techniques, the likely conformational differences between a given peptide and the corresponding region from the intact protein have been widely acknowledged (Van Regenmortel, 2006; Chen et al., 2009).

But arguably there remains an even more fundamental question: for methods that utilize relatively small windows onto the primary amino-acid sequence, how likely is it that a segment will be found that spans a significant number of epitope residues? This is one of the questions we address in this paper. More generally, we seek to extend the analysis already carried out by Rubinstein et al. (2008) into the properties of structural epitopes by quantifying the degree to which B-cell epitopes – both structural and functional – are discontinuous.

2. Methods

2.1. Protein data set

A dataset of X-ray crystallographic antigen–antibody structures was constructed based on an initial list derived from the Summary of Antibody Crystal Structures (SACS) database (Allcorn and Martin, 2002). Various criteria were imposed to filter out inappropriate structures, notably those with missing data or of low quality. Hence structures were removed that did not have:

- A resolution of $\leq$3 Å.
- An antibody component comprising at least part of both heavy and light chains.
- An antigen component containing at least 30 amino-acid residues.
- And a complete structural epitope (i.e. no missing information relating to the antigen's epitopic residues).

We also excluded structures where the epitope includes residues from multiple chains of the antigen. Such epitopes are relatively uncommon (accounting for only 2.6% of those selected according to the preceding criteria), although they are potentially important in specific contexts; for example, epitopes that span the HA1 and HA2 domains in influenza A haemagglutinin are at the centre of research seeking to identify the targets for vaccines with long-term effectiveness (Ekiert et al., 2009; Corti et al., 2011). However, such epitopes are hard or impossible targets for the window-based methods we are concerned with here, which consider only continuous peptides, and including them would have over-complicated our analyses.

Given the need to automate the selection of appropriate structures for our data set, various heuristics were implemented (based on Li and Wang (2009)) to exclude PDB entries for which we were