Molecular Immunology 68 (2015) 465–475

Contents lists available at ScienceDirect

Molecular Immunology

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

Peptide motif analysis predicts alphaviruses as triggers for rheumatoid arthritis

Charissa Hogeboom*

Independent Researcher, Albany, CA 94706, USA

ARTICLE INFO

Article history:
Received 18 August 2015
Received in revised form
11 September 2015
Accepted 17 September 2015
Available online 23 October 2015

Keywords:
Rheumatoid arthritis (RA)
HLA DRB1
Shared epitope
Alphavirus
STAT1 inhibition
Molecular mimicry

ABSTRACT

Rheumatoid arthritis (RA) develops in response to both genetic and environmental factors. The strongest genetic determinant is HLA-DR, where polymorphisms within the P4 and P6 binding pockets confer elevated risk. However, low disease concordance across monozygotic twin pairs underscores the importance of an environmental factor, probably infectious. The goal of this investigation was to predict the microorganism most likely to interact with HLA-DR to trigger RA under the molecular mimicry hypothesis. A set of 185 structural proteins from viruses or intracellular bacteria was scanned for regions of sequence homology with a collagen peptide that binds preferentially to DR4; candidates were then evaluated against a motif required for T cell cross-reactivity. The plausibility of the predicted agent was evaluated by comparison of microbial prevalence patterns to epidemiological characteristics of RA. Peptides from alphavirus capsid proteins provided the closest fit. Variations in the P6 position suggest that the HLA binding preference may vary by species, with Ross River virus, Chikungunya virus, and Mayaro virus peptides binding preferentially to DR4, and peptides from Sindbis/Ockelbo virus showing stronger affinity to DR1. The predicted HLA preference is supported by epidemiological studies of post-infection chronic arthralgia. Parallels between the cytokine profiles of RA and chronic alphavirus infection are discussed.

1. Introduction

Rheumatoid arthritis (RA) is a chronic immune-mediated disease that affects approximately 0.5–1% of the population worldwide. Genetically, RA is most strongly associated with HLA type, specifically a set of DR alleles containing what has been termed the "shared epitope" (SE) (Gregersen et al., 1987), based on polymorphisms that affect the electrostatic charge of the P4 binding pocket. While these polymorphisms clearly play a role, the reason why so few carriers of the SE go on to develop RA has not yet been elucidated. A rough estimate of the proportion of the world's population for whom at least one allele contains the SE is upwards of 30% (based on Solberg et al., 2008), suggesting that approximately 3% of SE carriers develop the disease. While genetic factors aside from HLA type undoubtedly contribute, even genetically identical twins are more likely than not to be discordant for the disease. Studies in the UK, Denmark, Finland, and Australia have estimated RA concordance across monozygotic twin pairs to range from 9% to 21% (Silman et al., 1993; Svendsen et al., 2013; Bellamy et al., 1992; Aho et al., 1986). This suggests involvement of an environmental factor that is somewhat uncommon. Epidemiologists have proposed that the environmental component may be an infectious agent, but studies investigating the link with microorganisms such as Epstein–Barr virus, parvovirus, *Proteus*, and *Mycoplasma* species have been unconvincing (Silman and Pearson, 2002) and the environmental trigger remains elusive.

Initiation of autoimmune disease by an infectious pathogen is hypothesized to involve immune recognition of self-antigens by way of "molecular mimicry" (Olson et al., 2001), in which structural similarities between a microbial peptide and a self-peptide trigger activation of autoreactive T cells. The way in which HLA type might play into this process is unclear. In the case of RA, complexity is increased by the number of haplotypes conferring risk and the degree of risk associated with each. The distribution of specific SE alleles varies geographically and by racial or ethnic group. Classically, among Caucasians, the highest RA risk has been observed among carriers of several DR4 alleles, specifically DRB1*0401, *0404, and *0405 (Buckner and Nepom, 2002). As association studies expanded to include non-European populations, additional high-risk haplotypes were identified. RA risk in the Mediterranean, the Middle East, and portions of Latin America is elevated among carriers of DRB1*0101 or *1001 (Debaz et al., 1998; Mourad and Monem, 2013; Citera et al., 2001). In East Asians, the DRB1*0901 allele is implicated (Lee et al., 2004; Kong et al.,

* Corresponding author.
E-mail address: Charissa.Hogeboom@outlook.com

http://dx.doi.org/10.1016/j.molimm.2015.09.010
0161-5890/

2002), while in Native Americans, risk is especially high for carriers of DRB1*1402 or *1406 (Ferucci et al., 2004). All of these alleles carry the SE pattern, with residues QRA, QKA, or RRA in DRβ positions 70, 71, and 74, respectively. In addition, polymorphisms in the DRβ11 and 13 positions, affecting the P6 binding pocket, have been found to intensify RA risk in both European and Asian populations (Raychaudhuri et al., 2012; Okada et al., 2014). It seems reasonable to assume that molecular mimicry would be related to the degree to which a microbial peptide preferentially interacts with these polymorphisms in the P4 and P6 pockets of HLA DR.

The aim of this investigation was to identify the microbial agent most likely to trigger autoimmune reactivity in RA under the molecular mimicry hypothesis. The self-antigens responsible for pathogenesis of RA are unknown, but collagen II (CII) is suspected to be involved since autoantibodies against this protein are found in joints of RA patients (Banerjee et al., 1988). The target chosen for mimicry in this analysis was CII 1168–1180, one of two collagen peptides predicted to selectively bind to the DR4 haplotypes associated with RA (Hammer et al., 1995), and for which X-ray crystallographic imaging of the peptide-DR4 complex has been published (Dessen et al., 1997). A set of proteins from microbial pathogens was scanned for regions of sequence similarity to CII 1168–1180, and homologous peptides were compared to the target on five scales representing characteristics predictive of protein binding and configuration. The highest scoring microbial peptides were evaluated for similarity to a binding motif based on characteristics that distinguish high-risk from protective DR alleles, particularly affecting the P4 and P6 pockets. The plausibility of the proposed trigger agent was then evaluated by comparing the geographic distribution of antibody seroprevalence to twin concordance rates and to RA prevalence rates after controlling for the frequency of high-risk DRB1 alleles. Finally, a mechanism by which infectious agents might initiate HLA-mediated RA is proposed.

2. Materials and methods

2.1. General

All computations were done with custom programs written in the R language (Hornik, 2014).

2.2. Viral proteins

A list of viruses and facultatively intracellular bacteria or parasites was generated from review of medical reference and microbiology texts. Pathogens typically causing severe disease or death (e.g., rabies virus, Ebola virus) or that infect only the skin (e.g., human papillomavirus) were excluded. Because low twin concordance implies an agent with low prevalence, some very common viruses (e.g., herpes simplex, influenza) were also excluded. One hundred eighty-five protein sequences from the viral capsid or envelope or from bacterial outer membranes were selected for testing. Protein sequences were obtained from the UniProt database.

2.3. Reference proteins

Viral homology scores were contrasted with scores from two other comparators. As a negative control for CII 1168–1180, an arbitrary 13-mer from serum albumin, ALB 151–163, was selected. As a negative control for the microbial proteins chosen for testing, a set of 185 control proteins was created by generating a chain of randomly selected amino acids whose frequency of occurrence was similar to the amino acid composition of proteins in general. The random proteins were matched for length with the microbial proteins.

2.4. Homology search

The homology search was conducted as described previously (Hogeboom, 2015). Specifically, microbial protein sequences were scanned for regions of high homology with each of 24 windows on the full length of CII 1168–1180. The windows chosen were those of length 4–13 that covered the region CII 1170–1173, the primary anchor region for peptide bound to DRB1*0401 (Dessen et al., 1997).

To allow comparison of homology scores across window locations and widths, scores were normalized to window length by averaging. Average homology scores were scaled to percent of maximum, defined as the homology score obtained when comparing the windowed segment of the CII immunogen with itself, using the value –1 as minimum of the range. The distribution of homology scores was compared across peptide source (microbial or random) using a Wilcoxon rank sum statistic at level 0.05. A Simes test was used to control for multiple comparisons. Since the windows were overlapping and nested, the statistics used are considered exploratory rather than indicating true statistical significance.

2.5. Peptide characteristics profiles

Microbial peptides were compared to CII 1168–1180 on five scales used to predict protein conformation or binding. These were surface accessibility, antigenicity, flexibility, hydrophobicity, and hydrophilicity (Janin et al., 1978; Welling et al., 1985; Karplus and Schulz, 1985; Eisenberg et al., 1984; Parker et al., 1986). For each scale, a profile for CII 1168–1180 was defined as the sequence of scale scores for the 13 residues. The CII profile was then compared with that of each candidate homologous microbial peptide by summing the positionwise squared deviations in scale scores. Dividing by the peptide length, each homologous microbial peptide was given a "mean squared error" (MSE) value that reflected the degree to which the profile of the microbial peptide deviated from that of CII on the given scale. The MSE values were converted to percents by dividing by the maximum possible deviation from CII that could have been obtained using that scale.

2.6. Motif for peptide binding

A motif was built to codify the criteria a peptide must meet to bind to the high-risk HLA alleles and present a similar face to the T cell receptor as CII 1168–1180. These criteria were based on X-ray crystallographic data and on characteristics that distinguish high-risk from low-risk DRβ alleles as provided by others and are summarized in the first two columns of Table 3.

High-risk and protective DRβ alleles were identified based on the work of Gregersen et al. (1987), Gibert et al. (2003), Raychaudhuri et al. (2012), and Okada et al. (2014). High-risk alleles differ from protective ones in DRβ chain polymorphisms affecting primarily the P1, P4, and P6 pockets.

Requirements for the P4 pocket are derived from Dessen et al. (1997) and Gregersen et al. (1987). Polymorphisms in DRβ positions 67, 70, and 71 distinguish the P4 pocket of high-risk from protective alleles. Substitution of QR or QK in place of DE at DRβ70–71 changes the electrostatic charge of the P4 pocket, making it positively charged. In addition, DRβ67 is occupied by leucine in the high-risk alleles, in contrast to the hydrophobic isoleucine or phenylalanine, which may accentuate the charged nature of this pocket. This set of commonalities among the high-risk alleles has been labeled the "shared epitope" (Gregersen et al., 1987). Preferred peptide residues in this position are D and E (Dessen et al., 1997).

Requirements for the P1 pocket are derived from Dessen et al. (1997), Hennecke and Wiley (2002), and Stern et al. (1994). Most of the high-risk alleles contain DRβ86G, in contrast to valine in this position in most of the protective alleles. The effect of this substi-