A geometric and algebraic view of MHC-peptide complexes and their binding properties

ABSTRACT

The algebraic and geometric view of amino acid sequences provides a theoretical framework to study the function of proteins when there is enough variation in this sequence to account for the variation in their function, as it is the case with MHC molecules in regard to their ability to present peptides.

INTRODUCTION

Background The cellular immune response depends on antigen presentation to T cells by major histocompatibility complex (MHC) molecules, the antigens presented being small peptides with amino acid sequence limitations determined by the MHC alleles carried by a particular individual. In the study of the immune response it is of critical importance to discover the laws governing which peptides can be presented by which MHC alleles. As peptide receptors, major histocompatibility complex (MHC) molecules are capable of binding different peptides; but each MHC molecule–each allele–exhibits a predilection for a set of peptides with distinct sequence characteristics. From crystallographic studies of MHC-peptide complexes, as well as from sequence analysis of peptides eluted from these complexes, we learn that MHC genetic polymorphism is responsible for the differences in peptide binding between MHC molecules. We claim that peptide selection by a given MHC allele can be presented as a function of the MHC sequence, instead of the current approach to this problem which consists in a function of the allele in question. By setting the amino acids that occupy the different positions in the MHC molecule as the arguments of this function instead of a mere reference to an allele, the door is open to make generalizations from available elution data to include those cases in which no empirical data is yet available. Over the past decade data on peptide binding to the MHC molecule have been accumulating. Do these data provide by themselves–that is, independently of other data such as crystallographic and chemical-sufficient information to characterise the requirements imposed by the various MHC alleles on what kind of peptide can be presented to T cells? The problem at hand consists in finding the proper theoretical framework to look at and analyse the sequences of these MHC-peptide complexes. We claim geometry provides this theoretical framework. Amino acid sequences can be represented as vectors in a metric space. It is the essence of this metric space that two sequences are at a distance of each other. Our analytical tools are the transformations we can impose on this metric space to manipulate the distances between sequences, that is, between MHC-peptide complexes. With the duality principle in geometry and the concept of dual spaces we move from a space of distances between amino acid sequences populated by MHC-peptide complexes to a space populated by the sequence properties of these complexes. In this new space we can measure the distance between the properties of the MHC-peptide complexes rather than the distances between the complexes themselves. Looking at the data in this light we can see how the sequence of the MHC molecule affects the sequence restrictions for the peptide allowed to bind to the molecule. This is revealed in the form of proximities between certain sequence properties of the MHC molecule and those of the bound peptide. In this study we concentrate on the class I human leucocyte antigen (HLA) molecules and nonameric peptides bound to them. The algebraic and geometric structure of MHC-peptide complexes Let S be the space of MHC-peptide complexes.

In this space each MHC-peptide complex is a point, represented by the vector of their amino acid sequence properties, the values the point takes in the co-ordinates of the space. The co-ordinates (dimensions) of this space correspond to these sequence properties, both peptide sequence properties and MHC sequence properties (only polymorphic positions in the MHC $\alpha 1$ and $\alpha 2$ domains are included). Peptide binding data allows the population of this space with × points. These × points are represented by matrix M of size × n, where n is the dimensionality of the space, that is, the total number of sequence properties, and each row is the vector representing each point. (In this study peptide binding data included 2535 nonameric peptides known to bind alleles HLA-A1, HLA-A11, HLA-A3, HLA-A2, HLA-A24, HLA-A25, HLA-A33, HLA-A31, HLA-A30, HLA-A29, HLA-B45, HLA-B44, HLA-B35, HLA-B53, HLA-B51, HLA-B14, HLA-B8, HLA-B38, HLA-B13, HLA-B27, HLA-B7, HLA-B42, HLA-854, HLA-CW2, and HLA-CW4, obtained from the MHCPEP database.) Now, let S' be the x-dimensional space of the sequence properties of the MHC-peptide complexes. Each point in this space corresponds to a property of the type 'MHC-peptide complex has amino acid x, at position pi', where pi can be either in the peptide or in the MHC molecule. Each dimension in this space corresponds to a particular MHC-peptide complex. We say S' is the transposed space of S. S is a space of particulars (molecules) and S' is a space of universals (sequence properties). The n points in S' are represented by matrix M' of size n ×, where each row is the vector representing each point. M' is the transposed matrix of M. (See Fig. 1) The points in S become co-ordinate axes in S' and the co-ordinate axes in S become points in S'. To give an example, let us define the sequence properties x, y and z (coordinates or dimensions in space S) as: x ≡ 'having> amino acid D in peptide position 3', y ≡ 'having amino acid E in peptide position 3', and z ≡ 'having amino acid Y in peptide position 9'. Let us also define the MHC-peptide complexes a, b, c and d (points in space S) as; a ≡ [HLA-A*0101-lADMGHLKY], b ≡ [HLA-A*0101-STEPVNILY], c ≡ [HLA-A*2402-TYSAGIVQI], and d ≡ [HLA-A*0207-LLDVPTAAV]. Then, ignoring all other points and dimensions, space S of MHC-peptide complexes is defined as: S = {a(1,0,1), b(0,1,1), c(0,0,0), d(1,0,0)}, where each point is defined by its coordinates and the values of these coordinates indicate that a property holds true ('1') or that it does not ('0'). Space S can then be represented by the matrix: where each row represents a point (an MHC-peptide complex) in the space with its three coordinates. And the transposed space S' of sequence properties is defined as S' = {x (1,0,0,1), y(0,1,0,0,), z(1,1,0,0,)}, and it can be represented by the matrix: where each row represents a point (a sequence property) in the space with its four coordinates. It can be seen that the matrix of space S' is the transposed matrix of space S. For this reason we call S' the 'transposed space' of S. This conversion is similar to the conversion of the sentence 'peptide i has amino acid Y at position 2' to the sentence 'having amino acid Y at position 2 is a characteristic of peptide i'. The subject and the predicate in the first sentence become the predicate and the subject in the second sentence respectively. There is a duality principle at work here. As Ramsey indicates, there is no sense in the distinction between 'individual' and 'quality', or between 'particular' and 'universal', the two types being in every way symmetrically related. The conversion of S into S' by transposing M allows the measurement of the distance between the points in S', that is between the sequence properties of MHC-peptide complexes. Particularly, we are interested in the distance between sequence properties specific of the peptide and sequence properties specific of the MHC molecule. The distance between the sequence characteristics of the MHC molecule and those of the peptide has the potential of revealing how peptides bind to MHC molecules and how sequence requirements are

imposed by the different MHC alleles on peptides as binding candidates. Our data unit is the MHC-peptide complex. We define two types of categorical variables of the MHC-peptide complex. First, variables of the following type: amino acid Paa is present at position Pp in the peptide-these variables we call 'p' and the set of them all 'P'. Second, variables of the type: amino acid Haa is present at position Hp in the MHC molecule-these variables we call 'h' and the set of them all 'H'. In other words, a partition of space S' (we call 'K') is made to separate the class of peptide sequence properties (P) from the class of MHC sequence properties (H), so that K = {P, H}, S' = P $\cup$ H and P $\cap$ H = Θ. The variables or properties we are talking about- 'having amino acid threonine at position 9 in the peptide', for instance-are bipolar: a peptide either has that amino acid at that position or it has not. In the case of MHC sequence properties, only polymorphic positions in the alpha 1 and 2 domains are considered. Therefore, nonamers are characterised by 180 sequence properties, one for each of the 20 amino acids at each of the nine positions; and MHC alleles are characterised by 207 sequence properties for each of the possible amino acids at each one of the 73 polymorphic positions in the alpha-1 and alpha-2 domains. Only some amino acids are found at each polymorphic site; for instance, at position 65 the amino acids G, Q and R can be found in various alleles; at position 90, only A and D; etc. A function D of the Cartesian product H × P into R (the line of real numbers) gives the distance between every pair (h, p) of sequence properties of MHC molecules and peptides bound to them. That is, D: H × P $\rightarrow$ R where D = {(h, p, D(h, p))| h $\in$ H, p $\in$ P}. (See Fig. 2) In other words, this function takes two arguments: one is an HLA sequence property (e.g. 'having amino acid E at position 46 in the HLA molecule'), and the other argument is a peptide sequence property (e.g. 'having amino acid P at position 2 in the peptide'). The value returned by this function is a distance measure. In measuring the distance between two such properties-points or vectors in S'-there are various alternatives. We use the Ochiai similarity index defined as: where a is the number of MHC-peptide complexes where both h and p hold, b the number of complexes where h holds but p dos not, c the number of complexes where h does not hold but p does, and d the number of cases where neither h or p hold. In other words, a, b, c and d are the cells in a 2 × 2 contingency table. This similarity index ranges from 0 to 1. In this study the values returned by function D were such that only 1% were greater than 0.4, 0.1% were greater than 0.6, and 0.03% were greater than 0.8. Measures of similarity and measures of dissimilarity-distance-can be linked to each other by means of 'similarity functions'. In this study we may talk about similarity measures, but conceptually it is distances we are dealing with. To give a concrete example, let us consider the following sequence properties: h1 ≡ 'having amino acid Q at MHC position 65', h2 ≡ 'having amino acid N at MHC position 66', p1 ≡ 'having amino acid P at peptide position 2', p2 ≡ 'having amino acid G at peptide position 5', p3 ≡ 'having amino acid I at peptide position 6'. The values returned by function D for the corresponding members of H × P are as follows: That is, for the first row, there are 57 MHC-peptide complexes in our database where at the MHC position 65 there is a Q and at peptide position 2 there is a P; there are 270 cases where at the MHC position 65 there is a Q but at peptide position 2 there is not a P; there are 2 cases where at the MHC position 65 there is not a Q and at peptide position 2 there is a P; and there are 524 cases where at the MHC position 65 there is not a Q and at peptide position 2 there is not a P. The D value for these two sequence properties is 0.33. Etc. A relation L(d1, d2) is defined on the function D so that two elements d1(h1, p1) and d1(h1, p2) are related if h1 and h2 are sequence properties characteristic of the same MHC allele, where h1 and h2 are members of H, p1 and p2 are members of P and d1

and d2 are members of D. This relation is not an equivalence relation because although it is reflexive and symmetric, it is not transitive. Therefore the relation L(d1, d2) does not lead to a partition of D, but to a collection Q of overlapping subsets. Each element of Q-subset of D defined by L-represents itself an MHC allele. (See Fig. 3) On each element qi of Q-subset of D corresponding to a particular MHC allele-we define relation N so that two elements of D in qi are related if the distance returned by function D corresponds to the same peptide sequence property $p \in P$. Relation N is an equivalence relation that defines a partition on qi. Let Ui(N) be the partition defined by N on qi, this partition will have at most 180 equivalence classes, one corresponding to each peptide sequence property $p \in P$. In saying that N is defined on qi rather than on Q we are actually considering the product or intersection of two relations: $L \cap N$. Since L is not an equivalence relation, nor is $L \cap N$, and it does not lead to a partition of Q. Let function V of Q into R180 (180-dimensional real-number space) return a vector for each MHC allele qi., member of Q, with the maximum value d = D(h, p) in each equivalence class of partition Ui(N). This R180 space we call Y and is the metric space of MHC alleles where each axis corresponds to a peptide sequence property (20 amino acids and 9 peptide positions), and each point in this space is an MHC allele. (See Fig. 4) Function V returns an 180-dimension vector. The co-ordinate value each MHC allele-each point in Y-assumes in each axis is a measure of predilection of the peptides to be bound to that allele for that corresponding peptide sequence property of that axis. If the property is, for instance, having glutamic acid at position 2 in the peptide, alleles that allow peptides with glutamic acid at position 2 will have a high value as their co-ordinate in that axis, and those alleles that do not will have a low value. The vector representing each allele in Y is in fact equivalent to the 20 × 9 'matrix' or table of 'average relative frequencies' or 'binding affinities' that characterise the sequence of peptides that bind to an allele. By giving these data an algebraic and geometric structure we can answer such questions as: 'Which amino acids and in which positions in the MHC molecule are involved in defining peptide binding requirements?', and also 'How different are amino acids in peptides in regard to meeting the binding constraints imposed by the MHC?'

CONCLUSION

Conclusions By looking at the problem of peptide binding to the MHC molecule from an algebraic and geometric perspective, molecules and their properties are seen as vectors in a metric space in which distances can be measured. A range of analytical tools then become available to study these distances, from which important conclusions can be drawn. The positions in the MHC molecule that determine peptide binding requirements are revealed. How different peptide amino acids are in meetings those requirements becomes clear. And we can define allelic peptide-binding profiles in terms of the amino acid sequence of the MHC allele, allowing the prediction of peptide binding for alleles for which there is no binding data as long as they have sequence similarities with alleles for which there is. At this time these predictions are limited by the fact that the correlation of allele vectors in spaces Y and Z is not perfect; but being able to make predictions at all marks the way to use current peptide data to make inferences for alleles for which there is no such data. We should emphasise that all the results presented here were derived automatically, without any input of previous knowledge of the biological question at hand. Our findings come directly and strictly from an automatic data analysis of the MHCPEP data. By means of the analytical methods introduced here knowledge accumulated by many researchers over many years has been

reproduced. In addition, these methods have brought to light new facts about peptide binding. A major limitation in this study is that variables (sequence properties) are assumed to be independent of each other in their effect on peptide binding, when there is good reason to believe they are not. Dropping the assumption of independence, however, comes at a very high cost by increasing the computational complexity of the problem to such a magnitude that makes it apparently intractable. Future progress in elucidating how peptides are selected for their presentation to T cells depends on the development of algorithms to analyse peptide binding data in a combinatorial way that would account for the possibility that the effect of MHC sequence variables is interrelated. Here we have shown that a propositional calculus can be developed to represent any such combination of sequence variables, fI-type propositional functions for MHC sequence variables and fII-type propositional functions for peptide sequence variables. We have also shown that the study of all the possible combinations of variables amounts to evaluating how these two types of functions- fI and fII-hold together in the space S' of empirical data. Geometry, as the study of abstract spaces, results from the distinction between 'set' and 'space'. A space differs from the mere set of the elements that populate the space by possessing a structure that places the elements of the space in a certain relation with each other. The elements in the space-points, vectors-are close or far from each other, there is a distance between them. By defining a distance measure between the elements of a set we confer a geometric structure to that set converting it into a space. If we can tell how different MHC alleles are from each other–how distant they are–in reference to a particular function, let us say, their ability to present peptides of a specific amino acid sequence, we are actually converting the set of MHC alleles into a metric space where a distance measure has been defined. Now, we can similarly define a distance measure between MHC alleles in terms of their sequence differences. In fact we can define many such distance measures by changing the 'scale' of the co-ordinates, that is, by changing the 'weight' a variable (co-ordinate) has in contributing to the distance measure-each new measure being the result of a transformation imposed on that space. If we find which transformation in the sequence space of MHC molecules results in a distance measure that best parallels-correlates more closely with-the distance measure in the functional space of MHC molecules, we have a geometric model of the function of MHC molecules in terms of their amino acid sequences. In this paper we talk about the 'transposed space' as the result of transposing the matrix of vectors in the sequence space of MHC-peptide complexes. This matrix operation creates a dual space where the n dimensions in the original space become the points-vectors-in the new space, and the × vectors in the original space become the dimensions-co-ordinates-in the new space. By doing so we convert the distance measure between MHC-peptide complexes into a distance measure between their sequence properties. This last concept-the distance between the sequence properties of MHC-peptide complexes-is a key that opens the door to elucidate the peptide binding requirements imposed by the MHC amino acid sequence. We conclude that algebra and geometry provide a convenient theoretical framework to study the function of proteins as amino acid sequences when there is enough variation in this sequence to account for the variation in their function, as we have seen to be the case with MHC molecules in regard to their ability to present peptides. The algebraic and geometric concepts presented here are the foundation for the design of an information model to create a database and the algorithms to manipulate it. They are not presented for the sake of advancing a unique and peculiar theory, but with an entirely practical intent. Although databases and computer programmes are typically presented as

implementations, it is preferable to present them formally in mathematical terms so that their true nature comes to light. The theoretical concepts presented here allowed us to manipulate MHC-peptide-binding data in a successful manner. Although in this paper we have centred our attention on the conceptualisation of the problem and on the methodological aspects of data analysis, we are aware that the results presented here depend on the quality of empirical data used in the analysis. A critical review of the data sets currently available indicate that careful auditing of the data, as well as continuous compilation of new empirical data are necessary.