

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
Programa de Pós-Graduação em Genética e Biologia Molecular

MatchTope: Ferramenta de busca de similaridades entre epítomos apresentados sobre o contexto MHC-I.

MARCUS FABIANO DE ALMEIDA MENDES

Tese submetida ao Programa de Pós-Graduação em Genética e Biologia Molecular da UFRGS como requisito parcial para a obtenção do grau de Doutor em Ciências (Genética e Biologia Molecular).

Orientador: Prof. Dr. Francisco Mauro Salzano
Co-orientador: Prof. Dr. Gustavo Fioravanti Vieira

PORTO ALEGRE

JUNHO DE 2018

Este trabalho foi realizado em parte no Núcleo de Bioinformática do Laboratório de Imunogenética do Departamento de Genética do Instituto de Biociências da Universidade Federal do Rio Grande do Sul e no Laboratório de *Molecular and Cellular Modeling*, do *Heidelberg Institute for Theoretical Studies*.

Apoio financeiro

CAPES - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior

Sumário

Lista de abreviaturas	4
Resumo	5
Abstract	6
Capítulo I	7
Introdução	8
1. Sistema Imune	8
2. Sistema Imune Inato	9
3. Sistema Imune Adaptativo	10
4. Rota de apresentação de antígeno	11
5. O MHC-I e o TCR: Suas estruturas e funções	13
6. A reatividade cruzada e sua utilização na prospecção de novas terapias	15
7. Bioinformática e Imunologia: O casamento perfeito	17
8. Dengue	19
9. Hepatite C	19
10. Zika	20
11. Câncer	20
Objetivos	22
12. Objetivo Geral	22
13. Objetivos Específicos	22
Capítulo II	23
Artigo publicado, Molecular Immunology	24
Capítulo III	33
Artigo a ser submetido à Scientific Reports	34
Capítulo IV	53
Discussão Geral	54
Referências Complementares (Capítulos I e IV)	68
Anexos	73

Lista de abreviaturas

BCR – Receptor de Células B (do inglês, *B Cell Receptor*).

CD4 – Agrupamento de Diferenciação 4 (do inglês, *Cluster of Differentiation 4*).

CD8 – Agrupamento de Diferenciação 8 (do inglês, *Cluster of Differentiation 8*).

CDR – Regiões Determinantes de Complementaridade (do inglês, *Complementarity Determining Region*).

CTL – Linfócito T Citotóxico (do inglês, *Cytotoxic T Lymphocyte*).

ERAP – Enzima aminopeptidase (do inglês, *Endoplasmic Reticulum Aminopeptidase*)

MHC-I – Complexo Principal de Histocompatibilidade de classe I (do inglês, *Major Histocompatibility Complex*).

PAMPs – Padrões moleculares associados a patógenos (do inglês, *Pathogen-associated molecular pattern*).

PDB – *Protein Data Bank*. Neste texto, refere-se ao formato do arquivo.

pMHC – Complexo peptídeo:MHC.

RGB – Sistema que utiliza três cores para compor uma imagem (do inglês, *Red, Green, Blue*).

TAP – Transportador Associado ao Processamento de Antígenos (do inglês, *Transporter associated with Antigen Processing*).

TCR – Receptor de Linfócitos T (do inglês, *T Cell Receptor*).

TCR:pMHC – Complexo TCR:peptídeo:MHC.

UHBD – Algoritmo que calcula o potencial eletrostático (do inglês, *University of Houston Brownian Dynamics*).

Resumo

O sistema imune é formado por uma série de conexões entre diversas células e moléculas, em uma rede de alta complexidade tendo como função primária a manutenção do equilíbrio homeostático do hospedeiro. Dentre os diversos tipos de respostas, uma das principais é a chamada resposta citotóxica, presente no conjunto classificado como sistema imune adaptativo. Os principais protagonistas são o linfócito T CD8+ e o complexo MHC de classe I. Através de uma rota de apresentação de antígeno, uma proteína, seja viral, tumoral ou própria, sofre um processo de degradação, sendo clivada para um tamanho de 8 a 13 aminoácidos, e migração até ser alocada dentro da fenda do MHC-I, sendo esta exposta na membrana celular para ser reconhecida pelo receptor da célula T citotóxica. Ao reconhecer o epítipo como não próprio, gera-se uma cascata de sinalização que leva à apoptose celular. Um mesmo receptor de célula T pode reconhecer mais de um tipo de epítipo, mesmo se já houve uma resposta prévia para um alvo específico, tendo este fenômeno o nome de reatividade cruzada. Este fenômeno tem uma grande atribuição no combate a doenças e tumores, possuindo um grau de importância no desenvolvimento de novas terapias e vacinas. Apesar da sua relevância, existem poucas ferramentas *in silico* disponíveis para a sua predição, gerando assim um grande nicho a ser explorado. Devido a isto, desenvolvemos o MatchTope, ferramenta que utiliza o complexo do pMHC em sua conformação tridimensional para a busca de similaridades, utilizando o campo eletrostático das estruturas como dado de entrada para o cálculo de similaridade. Com resultados promissores, temos em mãos uma ferramenta pronta que estará disponível para diversos pesquisadores que trabalham com prospecções de alvos para vacinas tanto para vírus quanto para novos alvos tumorais que podem ser utilizados em uma abordagem imunoterapêutica.

Abstract

The immune system is formed by a series of connections between several cells and molecules, in a network of high complexity having as primary function the maintenance of homeostatic equilibrium of the host. Among the several types of responses, one of the main is the so-called cytotoxic response, which is present in the set classified as adaptive immune system. The main protagonists are the CD8⁺ T lymphocyte and the MHC class I complex. Through an antigen presentation route, a viral, tumor or own protein undergoes a degradation process, cleaving it to a size of 8 to 13 amino acids, and migration until it is allocated into the MHC-I cleft, which is exposed in the cell membrane to be recognized by the cytotoxic T cell receptor. By recognizing the epitope as non-self, a signaling cascade that leads to cellular apoptosis is generated. A single T cell receptor can recognize more than one type of epitope, even if there has been a previous response to a specific target, and this phenomenon is called cross-reactivity. This phenomenon has a great attribution in the fight against diseases and tumors, having a degree of importance in the development of new therapies and vaccines. Despite their relevance, there are few *in silico* tools available for their prediction, thus generating a great niche to be explored. Due to this fact, we developed MatchTope, a tool that uses the pMHC complex in its three-dimensional conformation for the search of similarities, using the electrostatic field of the structures as input data for the calculation of similarity. With promising results, we have at hand a ready tool that will be available to several researchers who work with prospective vaccinations for both viruses and new tumor targets that can be used in an immunotherapeutic approach.

Capítulo I

Introdução e Objetivos

Introdução

1. Sistema Imune

Tentar descrever a complexidade do sistema imune em algumas poucas páginas é o mesmo que tentar descrever a vastidão do universo utilizando apenas alguns números. Um dos grandes charmes da Imunologia é o seu mistério devido à falta de explicações comprovadas em diversos pontos. Mas podemos pautar alguns conceitos importantes, simplificando alguns pontos e fazendo com que o leitor dessa humilde tese consiga compreender sua ideia geral, entendendo dessa maneira toda a lógica por trás deste trabalho, sem maiores dificuldades.

Dentre os sistemas, o imunológico é um dos mais complexos presentes nos organismos. Em um mundo “ideal”, todos os seres poderiam conviver em total harmonia, sem depender uns dos outros para obter recursos, todos juntos cantando *Imagine* do cantor John Lennon. Mas vivemos em um ambiente extremamente hostil e competitivo, onde todos os indivíduos são expostos a dezenas de milhares de patógenos, sendo necessário um sistema de defesa extremamente complexo e interconectado para conseguir manter a sobrevivência dos organismos. Se contarmos como partes do sistema imune as proteções intracelulares, como as enzimas de degradação, podemos então definir que desde bactérias até a superclasse gnatostomados existe um mecanismo imunitário de defesa.

O sistema imune é definido como um sistema de defesa do hospedeiro, composto por conjuntos de várias estruturas e processos biológicos, que são responsáveis por manter o organismo em um equilíbrio entre defender-se contra o que é não próprio e tolerar o que é próprio. As bactérias possuem um sistema enzimático rudimentar, principalmente para a proteção contra bacteriófagos (Stram & Kuzntzova, 2006). Outras formas básicas de proteção surgiram em eucariontes, como por exemplo fagócitos, sistema complemento, e defensinas (Beck & Habicht, 1996; Medzhitov, 2007). O clímax do sistema imune é encontrado a partir de vertebrados mandibulados (gnatostomados), com a ocorrência da forma mais complexa do sistema imune, graças ao surgimento de uma resposta adaptativa, possuindo este uma melhora na eficiência e no combate a patógenos no decorrer do tempo (Flajnik & Kasahara, 2010). Para compreendermos melhor como funcionam esses sistemas, duas categorias foram

geradas: sistema imune inato, onde possui uma primitiva memória imunológica, e adaptativo, o qual possui a geração de memória imunológica de forma mais complexa.

2. Sistema Imune Inato

O sistema imune inato é mais primitivo, comparado ao adaptativo, sendo encontrado desde bactérias até vertebrados (Medzhitov, 2007). É considerado a primeira barreira de defesa de um organismo. Apesar de algumas barreiras deles serem físicas e químicas, como a pele ou o pH do nosso estômago, algumas das respostas inatas podem ocorrer através da ação dos receptores de reconhecimento de padrões, capazes de reconhecer componentes que são conservados entre diversos microorganismos, chamados de PAMPs (*Pathogen-associated Molecular Patterns*) evidenciando uma resposta um pouco mais complexa comparada às outras supra citadas (Ausubel, 2005). Em geral o sistema imune inato não desenvolve uma resposta específica a um patógeno, porém possui uma rápida ativação e acaba sendo assim a linha de defesa imediata contra organismos invasores ou até mesmo contra células próprias que possuem alguma degeneração, como é o caso das células tumorais.

Entre os vários conjuntos de células e moléculas que compõem o sistema inato, podemos destacar como principais:

- Barreiras físico-químicas, que são compostas pela pele, enzimas digestivas, e mucos em vias respiratórias, para a contenção da proliferação de patógenos, entre outros (Boyton, 2002);
- Inflamação, sendo considerada esta a primeira resposta do sistema contra a infecção. É ela responsável pela sinalização e criação de uma barreira física, como por exemplo a vasoconstrição, que impede o avanço da infecção (Ferrero-Miliani et al., 2007);
- Sistema complemento, que é importante na resposta contra o patógeno, ligando-se à membrana e causando assim a citólise do invasor (Nonaka, 2014);
- Macrófagos, responsáveis por fagocitar patógenos, principalmente bactérias (Mills, 2012).

- Natural Killers, responsáveis por gerar resposta citotóxica contra células que não estão apresentando o complexo principal de histocompatibilidade (MHC, do inglês *major histocompatibility complex*) de classe I de forma correta. É uma das principais linhas de defesa contra vírus e tumores (Smyth et al., 2002; Vivier et al., 2011).

3. Sistema Imune Adaptativo

Evidências apontam que o surgimento do sistema imune adaptativo (SIA) ocorreu nos peixes mandibulados há aproximadamente 500 milhões de anos. Grande parte das células e moléculas associadas a este sistema de defesa já são encontradas em peixes cartilagosos (Flajnik & Kasahara, 2010). A evolução do sistema imune adaptativo como visto hoje em diversos vertebrados foi decorrente de uma duplicação gênica e uma transferência dentro do próprio genoma por transposição do gene RAG (do inglês recombination-activating genes) que está relacionado à imunoglobulina e receptores de célula T. Em peixes não mandibulados ocorreu o surgimento de um tipo diferenciado de sistema imune adaptativo, com poucas similaridades com o encontrado nos outros vertebrados (Boehm et al., 2018).

Diferentemente do sistema imune inato, o SIA pode possuir uma alta especificidade contra um determinado patógeno e ainda gerar uma memória contra este, sendo que em uma reinfecção ocorrerá uma resposta mais rápida e eficaz contra este invasor (Alder et al., 2005). O sistema imune adaptativo é constituído por dois tipos de células principais, chamadas de linfócitos T e linfócitos B. Estas são responsáveis pela resposta celular e humoral, respectivamente.

É denominada resposta humoral aquela mediada pelos anticorpos que é a principal via de combate ao patógeno em sua fase extracelular. Os linfócitos B são os responsáveis pela produção dos anticorpos. Gerados na medula óssea em mamíferos, os linfócitos B maduros geram os receptores de células B (BCR, do inglês *B-cell receptor*), sendo estes responsáveis pelo reconhecimento do antígeno (LeBien & Tedder, 2008). Simplificadamente, ao reconhecer um antígeno como não próprio, ocorrem sinalizações da célula T auxiliar (ou *Helper*, em inglês) para que a célula B sofra um processo de divisão e transformação, modificando-se assim em células de memória e plasmócitos (Yuseff et al., 2013). As células de memória irão ser geradas e armazenadas durante a primeira infecção, sendo acionadas

caso haja uma reinfecção pelo mesmo patógeno. Os plasmócitos vão produzir grande quantidade de anticorpos, os quais serão lançados na circulação sanguínea. Estes anticorpos possuem os mesmos receptores que estavam presentes na célula B original, sendo reativos a antígenos não próprios, ou em determinados casos a antígenos próprios, como ocorre nas doenças autoimunes (Yuseff et al., 2013). O enfoque desta tese é na resposta citotóxica, que será mais detalhada a partir de agora.

No que tange à resposta citotóxica, temos como cerne o linfócito T. Este também é gerado na medula óssea, porém diferentemente dos linfócitos B, antes de eles serem lançados na circulação passam por uma seleção no timo. No timo, o linfócito T virgem vai adquirir um fenótipo CD4 ou CD8 e passará por uma seleção positiva e negativa (Schwarz & Bhandoola, 2006). Esta seleção ocorre para que apenas células T que não geram resposta contra MHCs contendo peptídeos próprios sofram maturação e deixem o timo. As células que não apresentam resposta contra epítomos não próprios e que apresentam resposta contra epítomos próprios sofrem apoptose (Starr et al., 2003). As células T, após sofrerem a seleção, deixam o timo e migram para a corrente sanguínea.

Os linfócitos CD4⁺ são classificados como T auxiliares e os CD8⁺ são classificados como T citotóxicos. Além destes 2 grupos, existem várias subpopulações de células T, classificadas como células T de memória, reguladoras, e gama-delta entre outras (Vantourout & Hayday, 2013). Para uma melhor condução desta tese, iremos focar apenas na célula T citotóxica e sua interação com o MHC-I.

4. Rota de apresentação de antígeno

Toda infecção viral e algumas bacterianas ocorrem intracelularmente. Porém, os anticorpos não conseguem acessar o interior das células para eliminá-las. Para detectar e eliminar estas células infecciosas, faz-se necessário o sistema adaptativo citotóxico, no qual as células apresentam fragmentos de proteínas (epítomos) derivados dos patógenos, e o linfócito T citotóxico reconhecerá este epítomo então como não próprio e desencadeará uma reação que resultará na morte dessa célula infectada, impedindo assim a proliferação dos agentes invasores. Para tudo isto acontecer, faz-se necessário alguns passos desde a infecção até a sinalização de apoptose.

Proteínas citoplasmáticas, que consistem de proteínas próprias e não próprias (em caso de infecção), são direcionadas para um processo de ubiquitinação. As ubiquitinas são pequenas proteínas que têm como uma das suas diversas funções a de marcar proteínas para a sua degradação proteolítica. Para que isto ocorra, as ubiquitinas são adicionadas às proteínas com o auxílio de 3 enzimas: E1 (Enzima ativadora da ubiquitina), E2 (Enzima transportadora da ubiquitina) e E3 (Ubiquitina ligase) (Glickman & Ciechanover, 2002; Mukhopadhyay & Riezman, 2007).

Após a ubiquitinação, as proteínas são direcionadas para um complexo multiprotéico chamado proteossomo, onde serão clivadas em pequenos peptídeos. Nos humanos, esse complexo é composto por uma subunidade central denominada 20S e duas subunidades localizadas nas extremidades, denominadas de 19S. As subunidades 19S são responsáveis pelo reconhecimento das proteínas ubiquitinadas, em um processo dependente de ATP. Após a entrada na 19S, as proteínas sofrem uma modificação na sua estrutura secundária provocando a perda da sua forma nativa. Após estas modificações, as proteínas migram para a região 20S, onde se encontram os sítios catalíticos responsáveis pela clivagem das proteínas. As subunidades catalíticas $\beta 1$, $\beta 2$ e $\beta 5$, que estão presentes dentro da 20S, são as responsáveis pela quebra da ligação peptídica. Em um contexto de infecção viral, o proteossomo sofre modificações nas subunidades $\beta 1$, $\beta 2$ e $\beta 5$, modificando-as para $\beta 1i$, $\beta 2i$, e $\beta 5i$ ocorrendo assim uma alteração na especificidade do substrato. Com esta modificação, o proteossomo constitutivo passa a se chamar imunoproteossomo e este cliva as proteínas em regiões com a extremidade C-terminal mais hidrofóbica, gerando peptídeos ligantes que possuem uma maior afinidade com o MHC-I (Nassif et al., 2014; Murata et al., 2007). Além da modificação que resulta no imunoproteossomo, podem ocorrer três outras modificações nos proteossomos resultando no chamado proteossomo intermediários e também no timoproteossomo (sendo este presente apenas no timo). Ambos possuem modificações nas mesmas subunidades $\beta 1$, $\beta 2$ e $\beta 5$, porém de uma forma diferente ocorrido nos imunoproteossomos, ocorrendo assim uma diferença no padrão de clivagem dos peptídeos (Vigneron et al., 2017).

A clivagem proteica ocorrida no imunoproteossomo tem como resultado a formação de peptídeos com tamanho entre 3 a 30 aminoácidos. Porém, estes peptídeos podem sofrer rearranjos, onde determinados blocos de aminoácidos podem ser ligados com blocos que

estão distribuídos mais anteriormente ou posteriormente na sequência linear da proteína. Estes peptídeos, chamados de *spliced peptides* são produtos de uma transpeptidação, o que torna o conjunto formado ainda maior de peptídeos do que acreditava-se ao levar apenas a sequência linear em consideração. Este evento, ao primeiro momento, pode ser visto como uma forma de escape, pois gerando peptídeos que não possuem a sequência exata presente na proteína, o sistema imune poderia possuir uma dificuldade em gerar uma resposta mais específica a esta proteína. Porém, há evidências que 25% dos peptídeos apresentados são *spliced peptides* e que este fenômeno ocorre para uma produção de peptídeos com maior estabilidade quando apresentados na fenda do MHC-I (Vigneron et al., 2017).

Após a saída do proteossomo, alguns destes peptídeos ligam-se a transportadores associados à apresentação de antígenos (TAP), presentes na membrana do retículo endoplasmático. Esse transportador seleciona e transporta peptídeos para que sejam selecionados para apresentação no contexto de MHC-I. A TAP é uma proteína heterodimerica formada por duas subunidades, a TAP1 e a TAP2 (Antonίου et al., 2003). A região deste transportador que fica exposta ao meio citosólico possui uma abertura por onde o peptídeo, de tamanho entre 7 a 16 aminoácidos, pode se ligar e assim ser translocado para dentro do retículo. Sendo assim, a TAP seleciona epítomos que têm uma maior probabilidade de se ligarem e ficarem estáveis dentro da fenda do MHC-I (Lankat-Buttgereit & Tampe, 2002).

Dentro do retículo endoplasmático, o epítomo é preparado para ser integrado à fenda do MHC. Para isto acontecer, endoenzimas chamadas ERAP2, ERAAP e ERAP1 são responsáveis por clivar os epítomos maiores que 10 aminoácidos de forma que se estabilizem melhor na fenda do MHC. Encontram-se também proteínas como a tapasina, a calreticulina e a ERp57, que pertencem ao complexo de carregamento de peptídeo (PLC, do inglês *peptide loading complex*), sendo estas responsáveis por estabilizar o MHC e manter a fenda de ligação em uma conformação que favoreça o encaixe do epítomo (Blum et al., 2013). Após o epítomo estar devidamente alocado na fenda do MHC, o pMHC é migrado pelo complexo de golgi para a superfície celular, permitindo sua interação com as células T CD8⁺. A Figura 1 apresenta de uma forma simplificada toda a rota de apresentação de antígenos.

5. O MHC-I e o TCR: Suas estruturas e funções

O MHC é uma molécula encontrada em todos os vertebrados mandibulados, e é codificada no braço curto do cromossomo 6 em humanos (o MHC é denominado HLA em humanos) (Kulski et al., 2002). É ele dividido em MHC de classe I, II e III, os quais estão envolvidos em várias etapas da resposta imunológica, sendo o tipo I e II os mais relevantes para este processo. O MHC-I é uma proteína heterodímera, formada por duas cadeias polipeptídicas: uma cadeia alfa, definida como cadeia pesada e uma cadeia beta 2-microglobulina, sendo esta responsável pela estabilização do complexo. A cadeia alfa é subdividida nos domínios alfa 1, alfa 2 e alfa 3, onde alfa 1 e alfa 2 formam uma espécie de fenda, local onde são ancorados os peptídeos e onde o TCR interage para fazer o reconhecimento. Na região de alfa 3 acontece, além da ligação não covalente com a beta 2-microglobulina, o acoplamento da molécula acessória da célula T citotóxica, o CD8. Este acoplamento tem como função a manutenção da posição correta do MHC durante a interação entre TCR:pMHC (Antoniou et al., 2003).

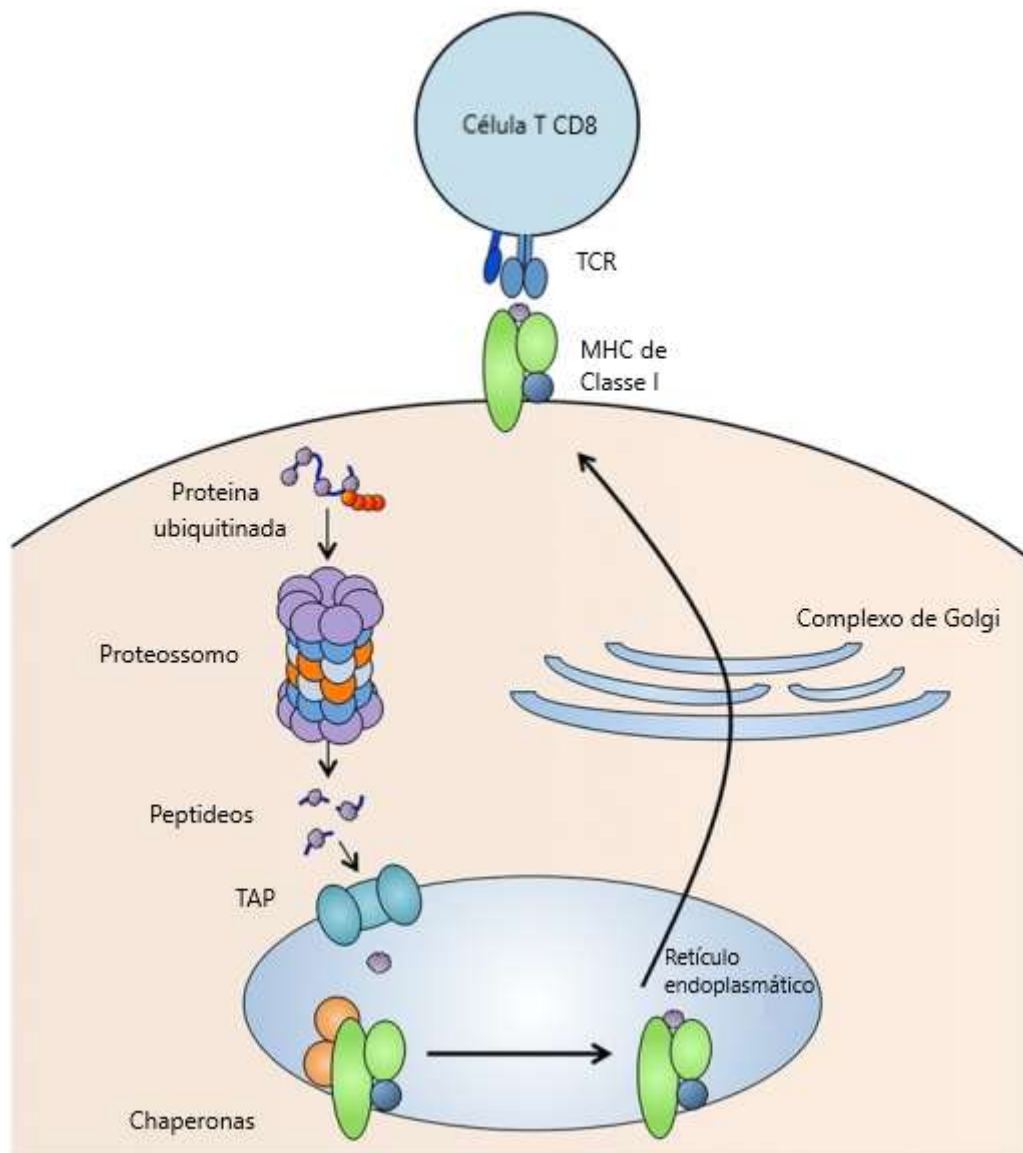


Figura 1: Ilustração da rota de apresentação de antígenos desde a ubiquitinação da proteína até o transporte para a membrana e a sua apresentação para o linfócito T CD8. Figura adaptada de Weinberg *et al.*, 2015

O MHC-I possui, nas regiões gênicas de alfa 1 e alfa 2, um alto grau de polimorfismo, permitindo assim que uma ampla variedade de peptídeos possa ancorar na sua fenda. Cada humano pode apresentar até 6 alelos de MHC-I diferentes, que são herdados sob a forma de haplótipos (Vandiedonck & Knight, 2009). Foram identificados cerca de 13.324 alelos que codificam o MHC de tipo I e 4.857 alelos que codificam o MHC de tipo II, totalizando 18.181 alelos diferentes envolvidos na geração de diversidade desta molécula, amplificando massivamente os peptídeos que podem ser apresentados (<http://hla.alleles.org/nomenclature/stats.html>). As regiões mais importantes para a

estabilidade do peptídeo na fenda são chamadas de *pockets* (bolsos), situadas nas regiões onde o segundo e o nono aminoácido se encaixam na fenda, os quais ficam mais enterrados que os outros aminoácidos, tendo esta região preferência para aminoácidos hidrofóbicos (Sidney et al., 2008).

O TCR é um heterodímero da superfamília das imunoglobulinas formado por uma cadeia α e uma cadeia β , ou por uma cadeia γ e uma cadeia δ . Estas cadeias são compostas por dois domínios extracelulares, variáveis e constantes (Attaf et al., 2015). A região constante fica mais próxima da membrana da célula T e a região variável contém três sítios hipervariáveis, denominados de CDR (CDR1, CDR2 e CDR3). Os CDR1 e CDR2 são responsáveis, geralmente, pelo reconhecimento de resíduos do MHC-I. A região CDR3, representada em vermelho e em verde nas regiões da alça na Figura 2, é a responsável pelo reconhecimento do epítipo. Para realizar esta função, a alça CDR3 possui uma maior variabilidade em relação a CDR1 e CDR2. A molécula CD8, também presente na superfície da célula T, auxilia na estabilização da interação TCR:pMHC, ligando-se à alfa 3 do MHC. Um mesmo TCR pode interagir com até um milhão de complexos pMHC distintos (Li et al., 2013). Se a célula T reconhecer um pMHC epítipo como não próprio, esta irá desenvolver uma resposta citotóxica contra a célula que está realizando a apresentação, desencadeando toda uma cascata de sinalização, e levando a célula infectada à sua apoptose (Milstein et al., 2011).

6. A reatividade cruzada e sua utilização na prospecção de novas terapias

O TCR é capaz de reconhecer uma gama enorme de peptídeos com diferentes intensidades. Ao ocorrer um reconhecimento degenerado, tem-se o que chamamos de reatividade cruzada, sendo este fenômeno definido como o processo onde um determinado TCR pode reconhecer e gerar resposta contra mais de um epítipo distinto (oriundo de patógenos diferentes, por exemplo) (Regner, 2001). Desta maneira, uma infecção por um patógeno seria capaz de gerar linfócitos T de memória capazes de produzir uma resposta contra um patógeno não relacionado posteriormente.

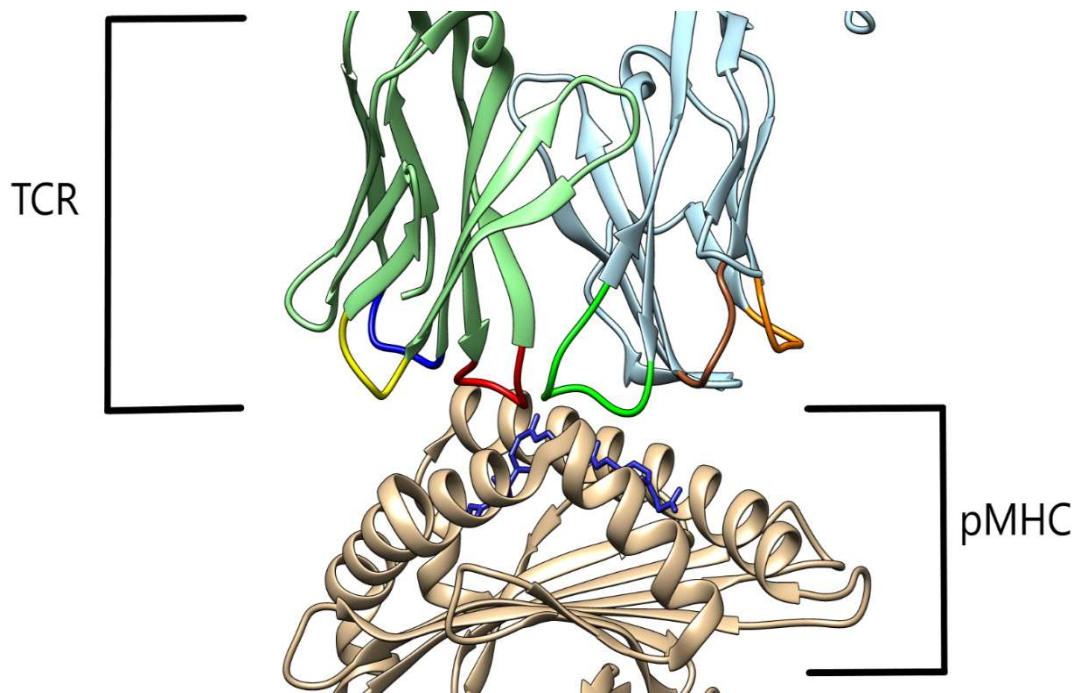


Figura 2: Ilustração mostrando como funciona o reconhecimento da célula T, através do seu TCR, de um epítipo apresentado pelo MHC de classe I. O TCR pintado da cor verde é equivalente ao alfa e o azul claro é o beta. Na região mais abaixo ocorrem as alças que são responsáveis pela interação com o pMHC. Na cadeia alfa do TCR, a alça em amarelo é a região CDR2, a alça em azul é a CDR1 e em vermelho é a CDR3. Na cadeia beta, em verde temos o CDR3, em marrom temos a CDR1 e em laranja temos a CDR2. Na parte do pMHC, em azul temos o epítipo e em bege temos as cadeias alfa 1 e alfa 2, formando a fenda.

Em um primeiro momento, quando pensamos em reatividade cruzada tem-se a ideia que esta reação ocorre devido ao alto grau de similaridade das sequências lineares do epítipo. Devido a isto, os primeiros trabalhos que estudaram o fenômeno da reatividade cruzada eram focados primariamente na busca por similaridades entre as sequências primárias de aminoácidos dos epítipos que a estimulavam. Entretanto, sabe-se que epítipos que compartilham menos de 50% de identidade entre seus resíduos também são capazes de gerar uma resposta cruzada (Fytily et al., 2008). Esse fato salienta a importância da avaliação de outras características além da sequência linear do epítipo, que contribuam para o desencadeamento da resposta imune, como por exemplo, as suas propriedades bioquímicas e estruturais.

No âmbito do desenvolvimento vacinal, é de grande interesse produzir uma vacina que confira proteção contra mais de um tipo de doença, ou pelo menos, vacinas que protejam

contra um maior número de linhagens ou genótipos virais em patógenos altamente variáveis como o Vírus da Hepatite C (HCV) ou o Vírus da Imunodeficiência Humana (HIV). Para isso, é necessário encontrar os padrões de similaridade que definem a ocorrência da reatividade cruzada. É importante salientar que nem sempre esse fenômeno de reatividade cruzada é benéfico. Podemos citar, por exemplo, doenças autoimunes como a esclerose múltipla, que são induzidas por mimetismo molecular, onde um antígeno exógeno faz com que células T gerem respostas cruzadas contra antígenos próprios (Tejada-Simon et al, 2003).

Em câncer a reatividade cruzada é uma das maiores preocupações dos pesquisadores. Ao utilizar um linfócito T para gerar resposta a um determinado conjunto de epítopos derivados de proteínas de células tumorais, tem-se o risco de que esses linfócitos acabem gerando também, por via de reatividade cruzada, uma resposta contra proteínas de células saudáveis, devido a características compartilhadas por peptídeos derivados destas duas células (Antunes et al., 2017; Linette et al., 2013). Em geral, uma das grandes diferenças entre células tumorais e saudáveis é o nível de expressão diferencial de diversas proteínas e a presença de proteínas de estágio embrionário e neoantígenos nas células tumorais. (Hanahan & Weinberg, 2011).

7. Bioinformática e Imunologia: O casamento perfeito

Graças aos avanços nas áreas da tecnologia, de processamento, e no avanço dos ambientes de simulação, as abordagens computacionais surgem como ferramentas poderosas em estudos biológicos. Assim, a imunoinformática (área da bioinformática que compreende preditores e abordagens *in silico* relacionados à imunologia) torna-se uma opção cada vez mais confiável e viável financeiramente em estudos envolvendo novas abordagens para o combate e tratamento de doenças infecciosas e tumores (Tomar & De, 2010).

Os bancos estruturais de proteínas, por exemplo, disponibilizam cristais de pMHCs (código 1HLA no Protein Data Bank por ex.) e TCRs (2V2W) (Berman et al., 2000) que podem ser utilizados para analisar os elementos que interagem na formação de um complexo pMHC, e que resíduos deste complexo são importantes na interação com o receptor de célula T. Além disso, o custo de uma cristalização experimental e sua taxa de sucesso é uma grande barreira para a geração de cristais no número de combinações de epítopos e alelos de MHC desejável

para uma compreensão mais completa da estimulação da imunogenicidade. Então, alternativamente, temos abordagens computacionais estruturais para suplantarem esse tipo de limitação técnica, como o ancoramento molecular. Esta ferramenta é utilizada na exploração de sítios e modos de ligação entre ligantes e seus receptores. Uma das ferramentas do nosso laboratório, chamada DockTope, realiza a ancoragem do epítipo, na sua estruturação alelo específica, no MHC de classe I (Rigo et al., 2015). Foi uma ferramenta desenvolvida por nós e que é de livre acesso.

Após o processo de modelagem dos complexos, pode-se utilizar uma abordagem complementar que forneça informações adicionais sobre as estruturas geradas. Uma delas é o cálculo de potencial eletrostático. Com ele, podemos inferir a carga residual em cada uma das diferentes regiões da superfície da molécula. Essa análise é de fundamental importância em diversos estudos, como nos desenhos de ligantes (fármacos), por exemplo (Kitchen, Decornez, Furr, & Bajorath, 2004). No caso das estruturas dos complexos pMHCs, esse é um dos parâmetros utilizados pela nossa ferramenta para encontrar padrões de similaridade, onde pMHCs que possuem um campo eletrostático mais similar são agrupados em ramos de um dendograma. Podemos inferir, com estes dados, uma possível reatividade cruzada entre os epítopos mais próximos, pois é a complementariedade de carga entre o TCR:pMHC que rege a resposta imunológica e a reatividade cruzada, sendo então estes valores inferidos pelo campo eletrostático do pMHC (Antunes et al., 2010; Mendes et al., 2015). Essa ferramenta, chamada MatchTope, foi o grande projeto deste doutorado.

Tem-se a disponibilidade de poucos softwares para a predição da reatividade cruzada. Além disso, a grande maioria destes softwares utilizam a sequência linear do epítipo de interesse, um epítipo derivado de uma proteína tumoral por exemplo, como entrada para tentar encontrar epítopos próprios que possam ser similares a ponto de desencadear uma reatividade cruzada (Moise et al., 2015; Z. H. Zhang et al., 2007). O maior problema desse tipo de abordagem é que, como já dissemos anteriormente, não se pode inferir reatividade cruzada apenas utilizando a sequência linear. Utilizando dados estruturais e usando valores como a topologia e a distribuição de cargas, que são os pontos-chaves do reconhecimento do TCR, tem-se uma confiabilidade muito maior nas predições. Com a nossa técnica, fomos capazes de prever reatividade cruzada de epítopos que possuíam similaridades menores que 50% (S. Zhang et al., 2015). Podendo a nossa técnica ser utilizada para uma gama de doenças,

iremos fazer uma rápida introdução a potenciais alvos que podem ser utilizadas como entrada para o MatchTope.

8. Dengue

O vírus da dengue é um arbovírus pertencente à família Flaviviridae. É formado por uma única fita simples de RNA, a qual codifica três proteínas estruturais e sete não estruturais (Rodenhuis-Zybert, Wilschut, & Smit, 2010). Possui cinco subtipos D1V, D2V, D3V e D4V e D5V que são diferenciados devido a sua antigenicidade (Normile, 2013). A infecção por qualquer um dos subtipos confere proteção permanente para o mesmo, porém a reinfeção por outro subtipo pode desencadear a dengue hemorrágica. Devido a este fenômeno, o maior problema que deve ser evitado na criação da vacina da dengue é a ocorrência da reatividade cruzada entre seus subtipos, sendo desejável buscar alvos que confirmam proteção para cada subtipo e que não tenham similaridades que possam desencadear resposta entre eles (Duan et al., 2012; Ranjit & Kissoon, 2011).

9. Hepatite C

O vírus da hepatite C é relativamente pequeno, envelopado, possuindo uma fita simples de RNA (Op De Beeck & Dubuisson, 2003). Pertencente à família Flaviviridae, possui seis diferentes genótipos e, pela sua alta taxa de mutação, estes genótipos podem apresentar ainda subdivisões (Simmonds et al., 1993). Os genótipos diferem entre 30 a 35% do genoma entre si, dificultando assim uma busca por uma vacina que abranja todos os genótipos (Ohno et al., 1997).

10. Zika

O vírus da Zika pertence à família Flaviviridae, é envelopado e apresenta uma fita simples de RNA, a qual codifica sete proteínas não estruturais e três proteínas estruturais (Malone et al., 2016). Possui ele duas principais linhagens: A linhagem Africana e a Asiática, tendo sido a segunda introduzida no Brasil, causando o surto de infecção que está ocorrendo desde 2015 (Sikka et al., 2016). Até o momento, não há vacina disponível para este vírus.

11. Câncer

O câncer é definido como um grupo de doenças envolvendo crescimento anormal das células, que podem ter potencial para invadir novas regiões do corpo além do seu sítio original. Existem mais de 100 tipos de tumores, tendo uma classificação de acordo com o seu local de origem e, se este tumor está se espalhando por outras regiões, é denominado de maligno, enquanto se apenas está se mantendo em sua região de origem, é classificado de benigno (Hanahan & Weinberg, 2011).

Há diversos tipos de tratamento para o câncer. Entre os mais comuns podemos citar a radioterapia, a quimioterapia e a cirurgia. Porém, tem-se utilizado a cada dia novas terapias, como a imunoterapia, onde se utilizam células do sistema imune para o tratamento do câncer (Alsaab et al., 2017; Ledford, 2017; Palumbo et al., 2013; Smyth et al., 2002). Essa abordagem explora o fato de células tumorais possuírem, em sua superfície, moléculas que podem ser reconhecidas pelo sistema imune, sendo chamadas de antígenos associados a tumores (*tumour-associated antigens* ou TAA). Essas moléculas são em geral proteínas, lipídios, carboidratos ou outros tipos de macromoléculas.

Uma das abordagens da imunoterapia baseia-se na utilização de linfócitos T citotóxicos para reconhecer células cancerígenas. Há várias aplicações do tratamento utilizando as células T CD8⁺ (Antunes et al., 2017; Ledford, 2017; Linette et al., 2013). Porém, o tratamento pode ocasionar efeitos colaterais, pois ao utilizar um linfócito T para gerar resposta a um determinado conjunto de epítopos derivados de proteínas de células tumorais, tem-se o risco de esses linfócitos acabarem gerando também, por via de reatividade cruzada, uma resposta contra proteínas de células saudáveis. Podemos citar, como exemplo, a reatividade

cruzada que ocasionou óbitos de pacientes ao serem submetidos a uma imunoterapia utilizando linfócitos T citotóxicos. Essas células foram selecionadas pois geravam resposta a um determinado epítipo de uma proteína de uma célula tumoral (proteína MAGE-A3). Porém, em alta concentração, esses linfócitos geraram uma resposta contra um epítipo de uma proteína própria do músculo esquelético (proteína titina) (Raman et al., 2016) (Morgan et al., 2013) . A sequência linear desses dois epítipos apresenta baixa similaridade (55%) (Linette et al., 2013), e provavelmente nenhum software disponível no mercado conseguiria prever essa reatividade cruzada.

Dada a quantidade de patógenos, o número de possíveis alvos, e a gama de possíveis tipos de abordagem para o tratamento que podem ser realizados, faz-se extremamente necessária uma ferramenta capaz de auxiliar o pesquisador na prospecção de novos alvos vacinais ou abordagens diferenciadas no âmbito da apresentação via MHC-I e resposta citotóxica.

Objetivos

Conhecendo a rota de apresentação de antígenos combinada com o reconhecimento do pMHC pelo TCR do linfócito T CD8+ e sabendo a sua importância na resposta imune contra patógenos intracelulares e células tumorais, este trabalho norteou-se nos seguintes objetivos:

12. Objetivo Geral

Desenvolver uma ferramenta capaz de prever similaridades entre diversos alvos prospectados e painéis de alvos imunogênicos no contexto do MHC-I, inferindo assim novos alvos que possam ser usados em abordagens vacinais e processos imunoterapêuticos.

13. Objetivos Específicos

- I. Desenvolver um *workflow* envolvendo a modelagem de alvos no contexto do MHC-I, utilizando o cálculo do potencial eletrostático e a análise das similaridades, contextualizando assim agrupamentos formados pelos pMHCs e indicando os possíveis alvos imunogênicos. As atividades programadas foram:
 - Modelagem dos alvos;
 - Edição do arquivo PDB e modificação em sua posição tridimensional;
 - Cálculo do campo eletrostático do pMHC;
 - Utilização como entrada para o agrupamento hierárquico os dados de carga da região da fenda; e
 - Apresentação do agrupamento através de um dendrograma e um *heatmap* como resultado da similaridade entre os alvos.

- II. Criar novas abordagens, utilizando o MatchTope, para o desenvolvimento racional de novas terapias e vacinas, através do seguinte:

- Busca por alvos que possuem similaridade entre vários genótipos e até espécies diferentes, para a detecção de reatividades cruzadas que auxiliem na pesquisa de novas vacinas;
- Busca por alvos da mesma espécie viral que não possuem similaridades, para o desenvolvimento de vacinas que não desencadeariam reatividade cruzada;
- Estudo de proteínas tumorais e próprias de células saudáveis, para prever uma possível reatividade cruzada não desejada entre as mesmas; e
- Utilização de dados de microarranjos de pacientes com tumor, para uma possível inferência sobre quais epítomos estariam sendo reconhecidos pelos linfócitos T CD8⁺ dos mesmos.

Capítulo II

Improved structural method for T-cell cross-reactivity prediction

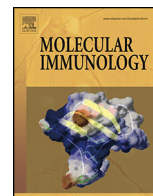
(Artigo “in extenso” publicado na revista “*Molecular Immunology*”)



Contents lists available at ScienceDirect

Molecular Immunology

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



Improved structural method for T-cell cross-reactivity prediction

Marcus F.A. Mendes^{a,b,1}, Dinler A. Antunes^{a,b,1}, Maurício M. Rigo^{a,b},
Marialva Sinigaglia^{a,b}, Gustavo F. Vieira^{a,b,*}

^a NBLI – Núcleo de Bioinformática do Laboratório de Imunogenética, Departamento de Genética, Universidade Federal do Rio Grande do Sul, Av. Bento Gonçalves 9500, Building 43323, room 225, Brazil

^b Programa de Pós-Graduação em Genética e Biologia Molecular (PPGBM), Universidade Federal do Rio Grande do Sul (UFRGS), Rio Grande do Sul, Porto Alegre, Brazil

ARTICLE INFO

Article history:

Received 9 April 2015

Received in revised form 3 June 2015

Accepted 16 June 2015

Available online xxx

Keywords:

Cross-reactivity

pMHC-I

HCA

ASA

Pvclust

Vaccine development

ABSTRACT

Cytotoxic T-lymphocytes (CTLs) are the key players of adaptive cellular immunity, being able to identify and eliminate infected cells through the interaction with peptide-loaded major histocompatibility complexes class I (pMHC-I). Despite the high specificity of this interaction, a given lymphocyte is actually able to recognize more than just one pMHC-I complex, a phenomenon referred as cross-reactivity. In the present work we describe the use of pMHC-I structural features as input for multivariate statistical methods, to perform standardized structure-based predictions of cross-reactivity among viral epitopes. Our improved approach was able to successfully identify cross-reactive targets among 28 naturally occurring hepatitis C virus (HCV) variants and among eight epitopes from the four dengue virus serotypes. In both cases, our results were supported by multiscale bootstrap resampling and by data from previously published *in vitro* experiments. The combined use of data from charges and accessible surface area (ASA) of selected residues over the pMHC-I surface provided a powerful way of assessing the structural features involved in triggering cross-reactive responses. Moreover, the use of an R package (pvclust) for assessing the uncertainty in the hierarchical cluster analysis provided a statistical support for the interpretation of results. Taken together, these methods can be applied to vaccine design, both for the selection of candidates capable of inducing immunity against different targets, or to identify epitopes that could trigger undesired immunological responses.

© 2015 Elsevier Ltd. All rights reserved.

1. Introduction

Cellular immunity is one of the two main branches of the adaptive immunologic response, focused on specific functions of the cytotoxic T-lymphocytes (CTLs). Although both cellular and humoral immunity are desired for an ideal and longstanding immunization, CTL response plays a central role in regard to antiviral immunity (Brehm et al., 2004). After infecting a host cell, the virus

will use the host molecular machinery to replicate its genome and produce new virions. In addition to all the mechanisms that allow virus escape from circulating neutralizing antibodies, during its intracellular replication cycle the virus is virtually hidden from the action of humoral immunity. However, some viral proteins will unavoidably be marked to enter the endogenous antigen presentation pathway. Through this route, virus-derived peptides will be presented at the cell-surface in the context of major histocompatibility complex (MHC) class I molecules, forming stable peptide:MHC-I (pMHC-I) complexes. Each CTL produced by the host has one specific T-cell receptor (TCR), which is able to recognize pMHC-I complexes presenting nonself peptides. Therefore, through the interaction between pMHC-I complexes and TCRs, CTLs are able to identify and eliminate infected cells.

The TCR/pMHC-I interaction is highly specific, which allows the development of memory T-cells that will be once again triggered in future challenges with the same target. However, a given lymphocyte is able to recognize more than just one pMHC-I complex. This capacity of a CTL to recognize non-related peptides derived from the same virus, or even peptides from heterologous

Abbreviations: CTLs, cytotoxic T-lymphocytes; MHC, major histocompatibility complex; pMHC-I, peptide: major histocompatibility complex class I; TCR, T-cell receptor; D1–EM–D2, docking 1–energy minimization–docking 2; HCV, hepatitis C virus; ASA, accessible surface area.

* Corresponding author at: Programa de Pós-Graduação em Genética e Biologia Molecular (PPGBM), Universidade Federal do Rio Grande do Sul (UFRGS), Rio Grande do Sul, Porto Alegre, Brazil. Tel.: +55 51 33089938.

E-mail addresses: marcus.famendes@gmail.com (M.F.A. Mendes), dinler@gmail.com (D.A. Antunes), mauriciomr985@gmail.com (M.M. Rigo), msinigaglia@gmail.com (M. Sinigaglia), gusforavanti@yahoo.com.br (G.F. Vieira).

¹ These authors contributed equally to this work.

<http://dx.doi.org/10.1016/j.molimm.2015.06.017>

0161-5890/© 2015 Elsevier Ltd. All rights reserved.

viruses, was defined as cross-reactivity (Vieira and Chies, 2005). As expected, cross-reactivity has direct implications over vaccine development, autoimmunity and heterologous immunity, a process by which the immunization with one pathogen confers protection against another (Cornberg et al., 2010; Selin et al., 1994; Welsh and Fujinami, 2007; Welsh and Selin, 2002). Understanding of the molecular features driving these cross-reactivities became a major goal for several immunologists, but the system's complexity has delayed progress in the field. Wedemeyer et al. (2001) have proposed that cross-recognition of two heterologous epitopes could be triggered by the high amino acid sequence similarity between them. Similarity in terms of biochemical properties was also proposed as being the key for cross-recognition (Vieira and Chies, 2005), and was even applied with some success to predict cross-reactivity (Frankild et al., 2008; Moise et al., 2013). However, structural studies have shown that even epitopes with low sequence and biochemical similarity might present quite identical pMHC-I surfaces (Antunes et al., 2011; Sandalova et al., 2005), indicating that this structural similarity should account for the cross-stimulation of a given T-cell population.

Structural analysis of pMHC-I complexes can provide a level of information much closer to that presented *in vivo* for the interaction with the TCR. On the other hand, structural approaches are frequently limited by the number of pMHC-I structures already produced by experimental methods, such as X-ray crystallography and NMR (nuclear magnetic resonance). Our group has used structural bioinformatics tools to build *in silico* models of pMHC-I complexes that were not yet determined by experimental methods. This approach, referred as *D1-EM-D2 (docking 1-energy minimization-docking 2)*, was previously validated through the successful reproduction of several crystal structures (Antunes et al., 2010; Sinigaglia et al., 2013) and has been used to provide novel complexes for the CrossTope Data Bank for cross-reactivity assessment (Sinigaglia et al., 2013). Our group has also combined this approach with the use of multivariate statistical methods to make structural-based cross-reactivity predictions (Antunes et al., 2011). In a previous study, we used images of the electrostatic potential distribution over the pMHC-I surface to predict the cross-reactivity pattern among 28 naturally occurring hepatitis C virus (HCV) variants, in the context of HLA-A*02:01 (Antunes et al., 2011). Hierarchical clustering of proteins based on electrostatic potential over the entire surface has been previously used to protein functional assignment and protein classification, as performed by the webPIPSA server (Richter et al., 2008). This approach, however, is not suitable for cross-reactive prediction since most of the pMHC surface will not be contacted by the TCR and only few residues from the TCR-interacting face will play a key role in triggering a T cell response. The innovative image-based clustering of pMHC-I complexes here described has been shown to be a fast and efficient way to predict cross-reactivity using structural information, being able to identify cross-reactive targets even between epitopes which shared no amino acids in sequence (Zhang et al., 2015).

In a previous study, one region over the pMHC-I surface was defined, based on the observation of the main spots of variation among the 28 complexes analyzed. Based on the extracted information from the pMHC-I structures, we were able to predict the same clusters of cross-reactivity observed *in vitro* (Antunes et al., 2011). Despite the success of this approach, the same parameters could not be applied to other subsets, since different regions of the pMHC-I surface might have diverse influence over the TCR recognition. In this context, we presented here an improved and standardized structural-based method for T-cell cross-reactivity prediction of HLA-A*02:01-restricted epitopes. In the present work, we aimed to provide a generic set of "gates" that could be applied to any subset of epitopes restricted to HLA-A*02:01. These

gates were defined considering the key TCR interactions regions, which could be involved in cross-reactive responses.

Another improvement we implemented in this work was the inclusion of topography prediction. There are experimental evidences suggesting that charge similarity is more important than subtle topographic differences between the cross-reactive complexes (Jorgensen et al., 1992; Kessels et al., 2004). However, pMHC-I complexes are 3D structures and, hence, topography variation certainly has some influence over the TCR recognition. The accessible surface area (ASA) of a residue can provide a quantitative measure of how exposed or buried its side chain is, which will have impact over the pMHC-I topography. ASA values of the epitope residues, for instance, were previously related to immunogenicity (Meijers et al., 2005) and were also able to identify non-cross-reactive complexes (Antunes et al., 2010).

The predictive capacity of our method was enhanced by the inclusion of these new features such as mapping interaction zones in TCR/pMHC complexes that are responsible for cytotoxic response, topography prediction, and a bootstrap-based statistical method to validate the hierarchical clusters. Our results with the analysis of hepatitis C virus and dengue virus epitopes support its use as an important guidance tool for rational vaccine development.

2. Results and discussion

2.1. Identification of conserved contacts among TCR-HLA-A*02:01 crystal structures

The human HLA-A*02:01 is largely studied for being the most frequent MHC-I allele in human populations (<http://www.allelefrequencies.net/>) (Fernandez-Vina et al., 1992). For this reason, the protein encoded by this specific allele (called allotype) also presents the larger number of crystal structures available at the Protein Data Bank (PDB). Aiming to identify the residues involved in the recognition of this allotype by different TCRs, we performed an extensive search for all available crystal structures of TCR/HLA-A*02:01 complexes. This search returned 31 complexes (Table A.1), presenting 16 different TCRs and 20 different epitopes. Despite this variability, five epitope positions (p4–p8 – gates 1–3) and four MHC-I residues were consistently indicated as involved with TCR interactions, being present in more than 85% of these complexes. The P4–P6 positions of the epitope had already been observed as being directly involved in the stimulation of immunogenicity (Calis et al., 2012, 2013; Frankild et al., 2008; Hoof et al., 2010; Rudolph et al., 2006; Wucherpfennig et al., 2009). Several residues over the pMHC-I surface might participate in the interaction with the TCR, influencing the specific level of T-cell stimulation that will be triggered by each pMHC-I. However, we here postulate that changes in these nine conserved contacts might have greater impact over the T-cell recognition, therefore influencing cross-reactivity.

Supplementary material related to this article can be found, in the online version, at <http://dx.doi.org/10.1016/j.molimm.2015.06.017>

2.2. Inclusion of ASA values

We decided to include ASA values together with electrostatic potential information to improve our prediction method. It is important to note that the epitope amino acids composition will affect not only the charges and the ASA values of the epitope itself, but also of surrounding MHC-I residues. For that reason, in addition to the ASA values for the nine epitope residues, we also included ASA values from 28 frequently TCR-interacting MHC-I residues in our approach (Fig. 1B).

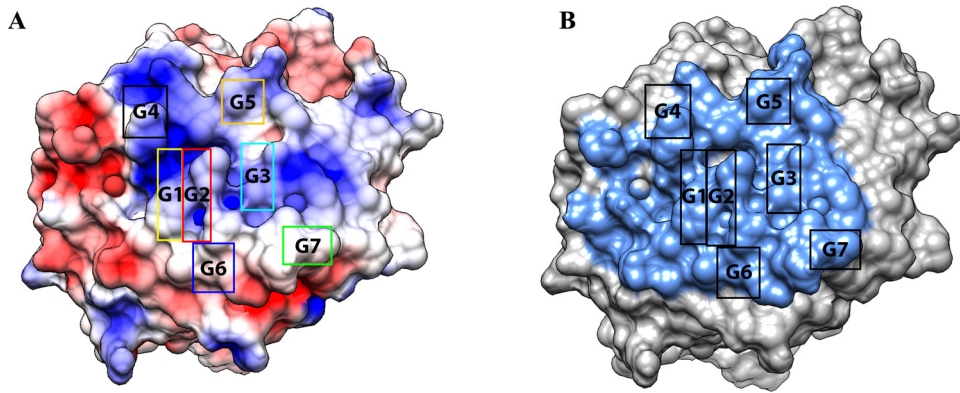


Fig. 1. Seven gates defined to obtain color histograms and selected residues for ASA assessment. Top view of a pMHC-I complex presenting a dengue-derived epitope in the cleft of HLA-A*02:01, obtained with the UCSF Chimera package (Trott et al., 2010). In (A), electrostatic potential over the surface was computed with the Delphi program and represented as red (negative charges) and blue (positive charges) spots, with a range from -3 to $+3$ kT. The seven gates (G1–G7) relate to conserved contacts with different TCRs, as observed in the crystal structures available, and were selected for the RGB analysis with ImageJ. In (B), the complex surface is depicted in grey while the surface of all residues selected for ASA assessment are indicated in blue. Black rectangles indicate the seven gates (from G1–G7) used in the RGB analysis. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Comparing results with and without ASA, we observed a better definition in the clusters, making the results more consistent with *in vitro* data. To exemplify this improvement, G6.26 (not including ASA values) appears in other branch, outside of the cross reactive cluster, being now included in the correct cross reactive cluster. For a full comparison, an image of clusterization analysis without ASA values can be viewed in Fig. A.1.

Supplementary material related to this article can be found, in the online version, at <http://dx.doi.org/10.1016/j.molimm.2015.06.017>

2.3. Method validation with a previously studied subset

Twenty-eight variants, covering all six HCV genotypes, were tested *in vitro* against the same T-cell population, which was

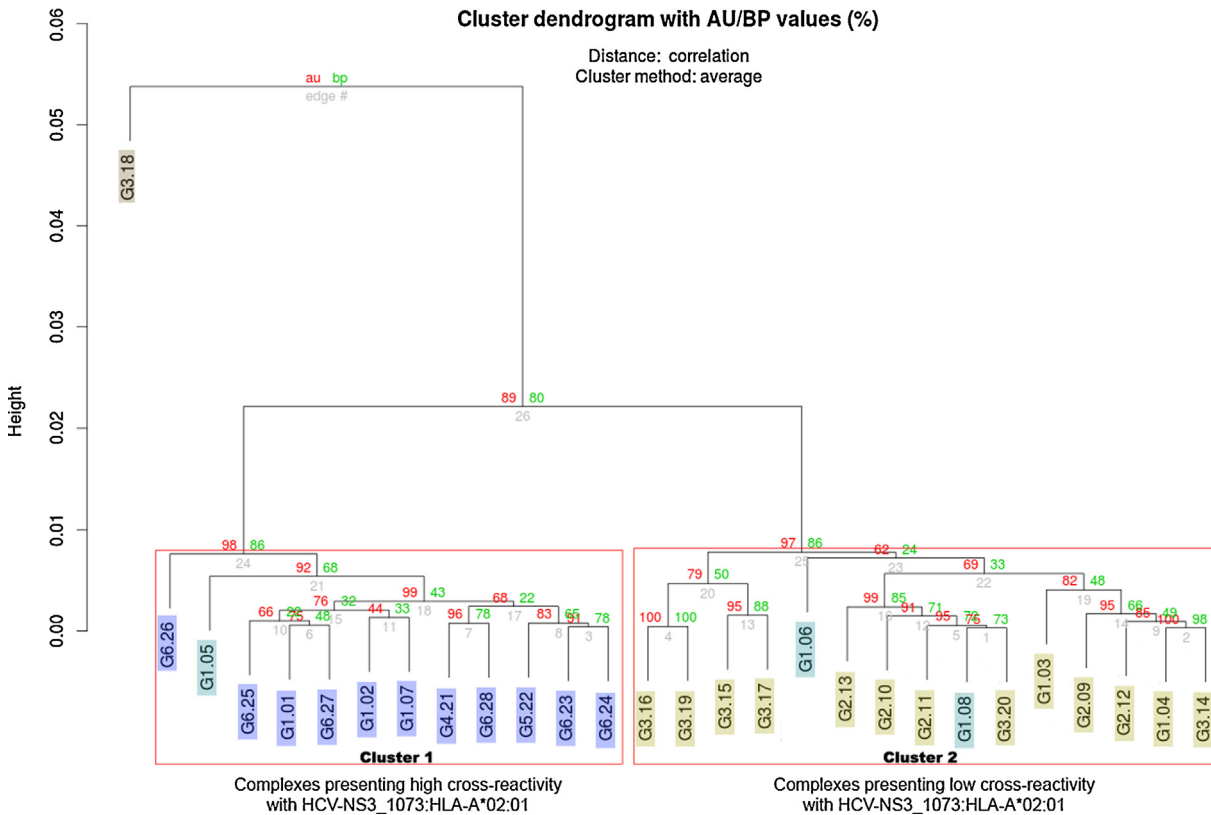


Fig. 2. HCA of 28 HCV natural occurring variants. Dendrogram representing the hierarchical cluster analysis (HCA) of 28 pMHC-I complexes loaded with HCV-derived epitopes covering all six HCV genotypes (from G1–G6). The input data was accessible surface area values and color histograms (RGB) for each pMHC-I, which provided information on topography and charges distribution over the surface. Red boxes indicate the main clusters identified ($\alpha = 0.95$). High (Cluster 1) and low (Cluster 2) G1.01 cross-reactive complexes fell in independent main clusters. The only complex that presented no response *in vitro*, G3-18, fell alone in an independent branch. The strong (dark blue), intermediate (light blue), low (yellow) and without (brown) cross-reactive targets in respect to G1.01 are represented inside individual boxes. Information on the specific response presented by each complex in cross-reactivity tests (*in vitro*) is provided in Additional file 3. G, HCV genotype; AU, approximately unbiased; BP, bootstrap probability. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

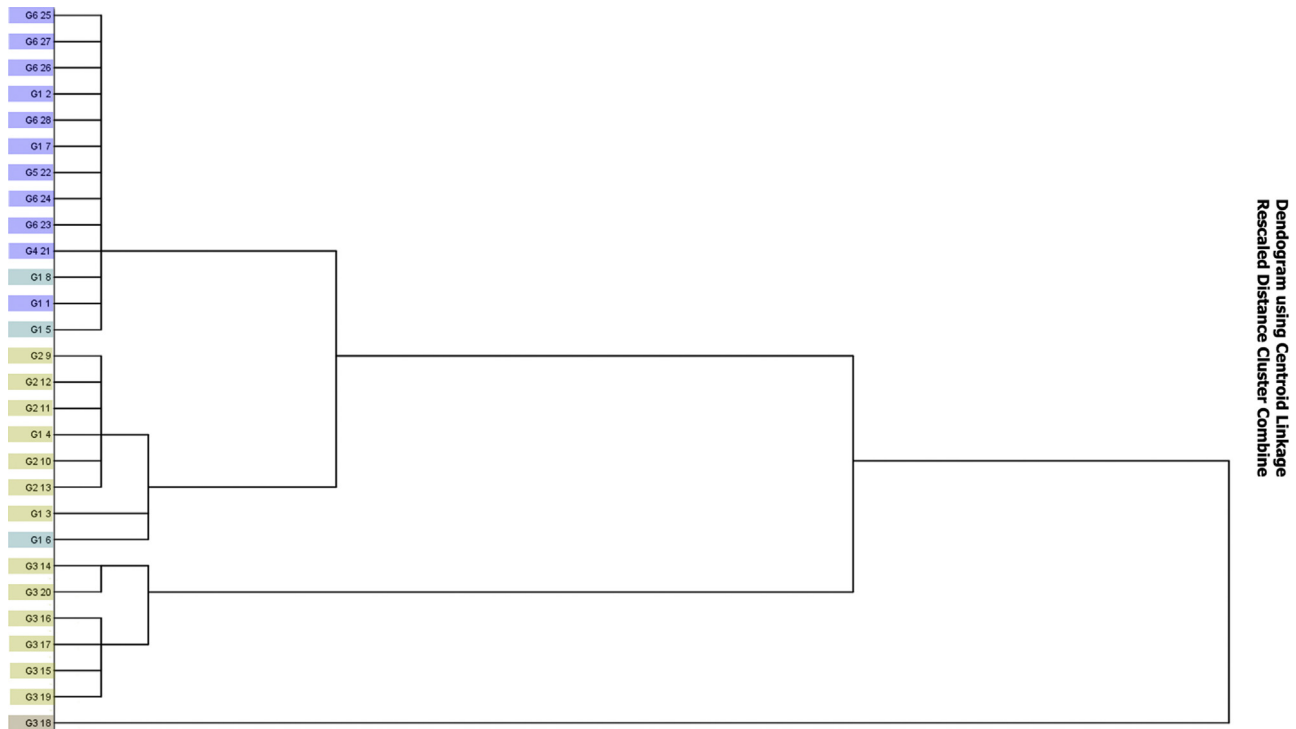


Fig. 3. HCA of 28 HCV naturally occurring variants from previous article. A modified figure from our previous article (Antunes et al., 2011), representing a hierarchical cluster analysis (HCA) of 28 pMHC-I complexes loaded with HCV-derived epitopes covering all six HCV genotypes (from G1–G6). The input data was extracted from a single spot in the surface, and provided information on charges distribution using color histograms (RGB) values. The dark blue boxes indicate the G1-01 cross-reactive complexes, light blue boxes depict the intermediate targets, yellow boxes indicate targets with low cross-reactives and brown boxes indicate the target with no cross-reactives. The dendrogram was generated by the SPSS software, using hierarchical clustering, with centroid method and squared euclised. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

obtained from an individual vaccinated with the wild-type epitope HCV-NS3₁₀₇₃ (CINGVCWTV) (Fyttili et al., 2008). The level of IFN-gamma production stimulated against a highly cross-reactive variant from genotype 1 (G1-01: CVNGVCWTV) was defined as a reference of high response, which was used to classify the other variants into high, intermediate or low cross-reactive complexes (Table A.2).

Supplementary material related to this article can be found, in the online version, at <http://dx.doi.org/10.1016/j.molimm.2015.06.017>

A HCA based in our improved approach was able to divide the complexes into two main clusters (Cluster 1 and Cluster 2) and one out-group represented by G3-18 (Fig. 2). A threshold was defined with the *pvrect* function to highlight these groups ($\alpha=0.95$), which are corroborated by AU *p*-values with low standard errors (Fig. 2). The variant G3-18 (from genotype 3) fell in a completely independent branch. This result is in agreement with our previous analysis and with the experimental data, since G3-18 was the only among the 28 complexes that presented no detectable response *in vitro* (Fyttili et al., 2008). All the high G1-01 (HCV-NS3₁₀₇₃) cross-reactive complexes fell in Cluster 1 (AU = 98). Of note, in the *in vitro* assay, the complexes with the higher IFN- γ levels within the cross-reactive complexes were G1-02, G1-07, G5-22, G6-25 and G6-27 (Antunes et al., 2010, 2011; Fyttili et al., 2008; Sinigaglia et al., 2013). With the exception of G5-22, all other complexes fell in the same sub-cluster of the reference variant G1-01 (AU = 76). It is important to note that this level of information was not contemplated by our previous work (Fig. 3). The high responder variant G6-26 and the intermediate responder G1-05 fell in separate branches, but still within the main cluster of the cross-reactive complexes (AU = 98). It is also important to mention that our previous analysis of these complexes presented the intermediate responder G1-05 as the closest related complex to the reference complex G1-01

(Antunes et al., 2011). We explained this unexpected result by suggesting that despite the surface charges distribution other issues might account for the lower response presented by G1-05. Our improved approach was able to identify neglected structural differences between G1-01 and G1-05, and correctly placed G1-05 outside the sub-clusters of high responders.

All low cross-reactive complexes fell in Cluster 2 (AU = 97). The low responders from genotype 1, G1-03 and G1-04, fell correctly into this main low responders cluster, as well as the intermediate responders G1-06 and G1-08. The complex G1-06 was also placed within the low responders in the original analysis (Fig. 3). Of note, a trend to the separation of the variants according to their genotypes is also observed, since we have a sub-cluster only with G3 complexes (AU = 79) and a sub-cluster with the majority of G2 complexes (AU = 99). Our HCA results also provide other suggestions, such as that G1-08 is more closely related to G2-11 and G3-20 (AU = 95) than to G1-06. However, to these new cross reactive suggested targets, there is no experimental background in Fyttili's paper to support this level of speculation (Fyttili et al., 2008). Note that the *in vitro* assay with these 28 HCV variants was performed to verify the cross-reactivity against the wild-type HCV-NS3₁₀₇₃. Cross-reactivity also depends on the T-cell population involved, so to evaluate the cross-reactivity against G1-08, an assay with a G1-08-specific T-cell population would be needed.

2.4. Cross-reactivity prediction among dengue virus serotypes

Dengue virus (DV) represents a major challenge for vaccine development (Halstead, 2013). Despite effective immunization against one serotype is easy to achieve, and protective T-cell response is observed, challenge of an immunized individual with an heterologous serotype often leads to severe symptoms, such as dengue hemorrhagic fever and dengue shock syndrome (DHF/DSS).

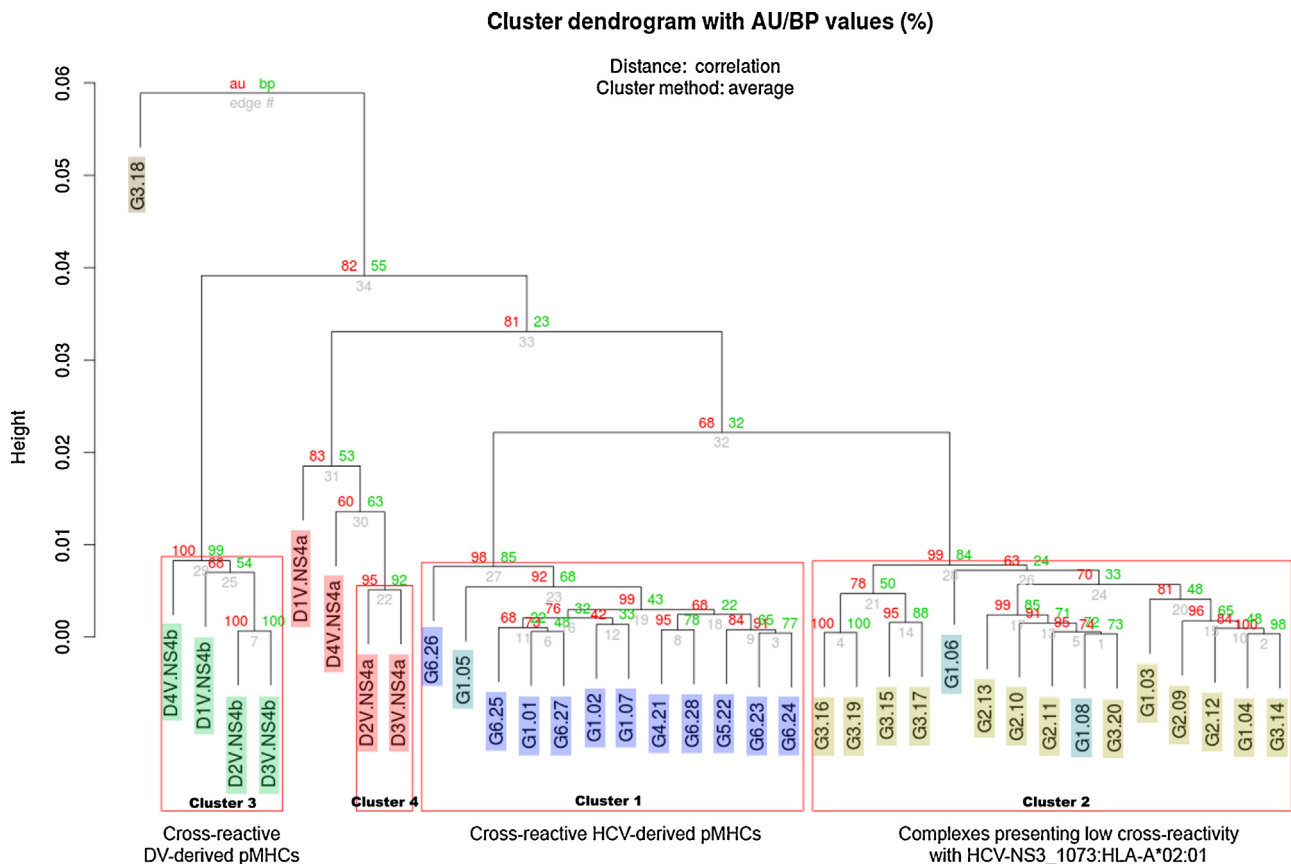


Fig. 4. Structure-based hierarchical clustering of pMHC-I complexes. Dendrogram of 36 pMHC-I complexes representing the hierarchical cluster analysis performed with the Pvcust R package. The input data was accessible surface area values and color histograms (RGB) for each pMHC-I, which provided information on topography and charges distribution over the surface. Red boxes indicate the main clusters identified ($\alpha=0.95$). Cross-reactive and non-cross-reactive complexes of both subsets (HCV and DV) fell in independent clusters. AU, approximately unbiased; BP, bootstrap probability. Dark blue box indicates the G1.01 cross-reactive complexes, light blue box indicates the intermediate targets, yellow box indicates targets with low cross-reactives and brown box indicate the target with no cross-reactives. Green box indicates NS4b targets and red box indicates NS4a targets. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

In this context, cross-reactive T-cells are believed to mediate the immunopathogenesis of DHF/DSS during secondary heterologous challenge (Duan et al., 2012). Therefore, the identification of non-cross-reactive immunogenic targets, specific for each DV serotype, is one way to develop a combined tetravalent vaccine. In a recent publication, Duan et al. (2012) identified HLA-A*02:01-restricted peptides from the four DV serotypes, and examined their immunogenicity and cross-reactivity. From their data, we extracted the epitope sequence of two groups of targets, one being identified as (i) cross-reactive variants, and the other as (ii) non-cross-reactive variants (Fig. A.2).

Supplementary material related to this article can be found, in the online version, at <http://dx.doi.org/10.1016/j.molimm.2015.06.017>

We performed new predictions with the combined data from both subsets (HCV and DV), totaling 36 pMHC-I complexes. The HCV and DV variants fell in independent main clusters, HCV maintaining the same complexes in Clusters 1 and 2 and defining two more groups (Clusters 3 and 4). The same threshold ($\alpha=0.95$) was able to identify cross-reactive and non-cross-reactive complexes within these groups (Fig. 4). All four NS4b variants fell in the same cluster (Cluster 3) ($AU=100$). This was expected, since cross-reactive *in vitro* response was indeed observed for these four variants. The same level of clustering was not observed for the NS4a variants ($AU=83$), a group that did not present cross-reactivity in the study of Duan et al. (2012).

The variants D1V-NS4a₁₄₀ and D4V-NS4a₁₄₀ fell in independent branches, while the other two (D2V-NS4a₁₄₀ and D3V-NS4a₁₄₀) fell in the same cluster ($AU=95$). Our HCA, therefore, indicates a possible cross-reactivity between D2V-NS4a₁₄₀ and D3V-NS4a₁₄₀, which could be understood as a false positive result. However, it is important to highlight that cross-reactivity is also dependent on the specific T-cell population involved, and normally produces responses with lower intensity when compared to the challenge with the cognate peptide. Of note, the D2V-NS4a₁₄₀ presented really low levels of response even upon challenge with the cognate epitope (Fig. A.1) (Duan et al., 2012). Despite of a possible structural similarity (Fig. 5), a cross-reactive response would be possibly undetectable with this T cell population. However, our approach relies exclusively on structural features of the pMHC-I surface, such as charges distribution and ASA values, and therefore is capable of identifying the closer related complexes. Also, other features in antigen processing might prevent the T cell stimulation process.

Finally, the combined HCA (HCV and DV) was able to reproduce the same results observed in the independent HCV analysis. This combined approach corroborates the consistency of our method, even with a greater number of complexes, suggesting its possible use in a larger scale as a virtual screening method. In this sense, we also explored an alternative way to present our HCA results. Instead of a dendrogram, this data can be used as input for relational networks, which can provide more intuitive information about the

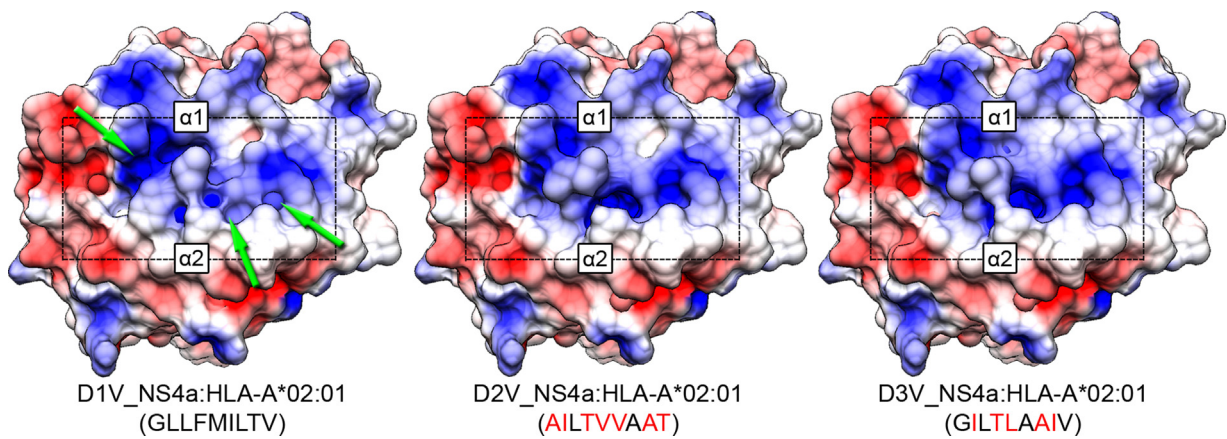


Fig. 5. Topography and electrostatic potential comparison among pMHCs presenting dengue-derived epitopes. “TCR-interacting surfaces” of three pMHC-I complexes presenting epitopes derived from three different Dengue Virus serotypes are depicted. Regions with positive (blue) and negative (red) charges are represented with a scale from -3 to $+3$ kT. Sequences of presented peptides are depicted below each complex, with mutations in relation to “D1V” indicated in red. Alpha-1 and Alpha-2 MHC domains are also shown. TCR-interacting surfaces of complexes “D2V” and “D3V” share greater similarity in terms of electrostatic potential, while “D1V” presents some differences in three positively charged spots (green arrows). Images were obtained with the UCSF Chimera package (Trott et al., 2010). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

cross-reactive-networks studied (Fig. 6), indicating, however, the same relationships.

2.5. Applicability to vaccine development

Several immunogenic targets were identified and successful immunization can be achieved, but HCV diversity remains a major challenge. The identification of targets capable of triggering cross-genotype responses could drive the efforts to develop a new generation of vaccines, improving vaccination coverage.

On the other hand, cross-reactivity is an issue to be avoided in a DV vaccine development, since it is involved in the immunopathogenesis of DHF/DSS. Once again, our improved structurally based prediction could be applied as a virtual screening method to identify undesirable cross-reactive responses that are unknown, and must be tested before the use of predicted targets in an anti-DV vaccine.

Traditional methods of vaccine development provided some successful results, but have been unable to overcome some of the major challenges for global health, such as the control of HIV and HCV. In that context, a new generation of rationalized vaccines is

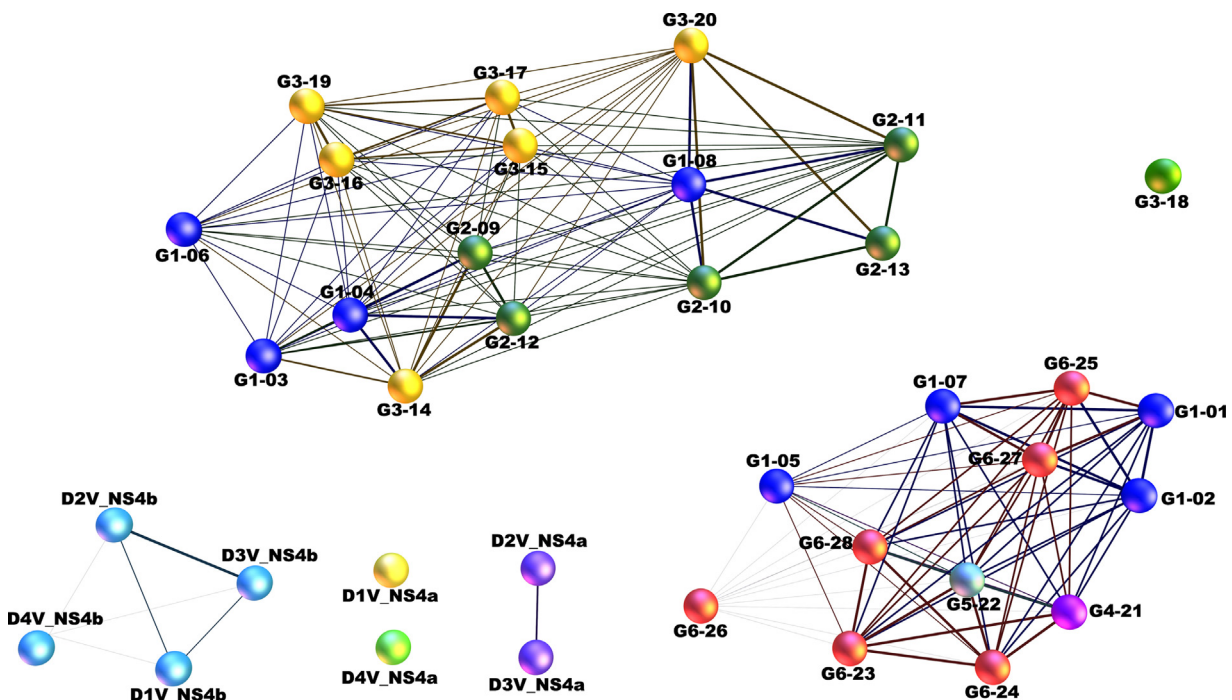


Fig. 6. Relational network of 36 pMHC-I complexes. Relational network generated with the Gephi program, based on the dendrogram of 36 pMHC-I complexes (Fig. 4). Each sphere represents a given pMHC-I and different colors indicate different HCV genotypes or DV serotypes. For instance, red spheres indicate pMHC-I complexes loaded with HCV genotype six epitopes. Lines (edges) indicate cross-reactivity between the connected complexes (nodes), complexes without connections are considered non-cross-reactive. The strength of each line indicates the similarity between the connected complexes, being a structure-based indicative of the strength of the cross-reactivity between them. The distribution of the clusters is merely representative, and distance between nodes in the picture has no meaning. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

starting to be planned, and bioinformatics tools are playing a major role in this process (Donati and Rappuoli, 2013; Dormitzer et al., 2012). Combined *in silico* approaches can save time and money, identifying the candidates more likely to stimulate the desired immune response, which can then be tested with *in vitro* and *in vivo* experiments to confirm its safety and efficacy for the use in a new vaccine.

3. Conclusions

The CD8+ T-cell cross-reactivity is a complex phenomenon triggered by the structural similarity between two different pMHC-I complexes that are recognized by the same TCR. Despite the enormous variability of TCRs and epitopes involved in these interactions, there are few conserved contacts that are shared by all TCR/pMHC-I crystal structures available, providing a map of the most important regions over the pMHC-I surface. Moreover, cross-reactivity between two pMHC-I complexes can be predicted based on the electrostatic potential over these selected regions. Although there are many studies about possible characteristics that trigger cross-reactivity, our method applying electrostatic potential (Antunes et al., 2011) and topology data to predict cross reactivity is a new one in this field. Our innovative approach showed that use of ASA values can improve this prediction, adding valuable information on the topography of these complexes. Finally, the use of an R package to assess the uncertainty of the hierarchical clustering provided a statistical validation of the results. Our method can be applied in rational vaccines construction, allowing to predict the impact of heterologous immunity and anticipate individual response to vaccination (Włodarczyk et al., 2009; Zhang et al., 2015). It can also be used to predict unexpected off-target toxicity in T cell based immunotherapies for cancer, field in which cross-reactivity has become a major concern (Linette, 2013; Stone et al., 2015).

The presented results demonstrate that our technique is on the right track. The next steps to consolidate this approach will come with the increase on analyzed cross-reactive networks, through the recovery and inclusion of *in vivo* experimental data available in scientific literature. This increase in the number of networks will strengthen the specificity of the approach, decreasing the number of false positive results. Alternatively, we aim to implement a strategy using neural network or Support Vector Machine algorithms to infer immunogenicity in pMHC complexes considering their charge distribution and topographic patterns. These different tools will become available in our immunoinformatics platform Crosstope – Structural Data Bank for Cross-Reactivity Assessment (<http://www.crosstope.com.br>).

4. Materials and methods

4.1. Identification of conserved contacts between TCRs and pMHCs

An extensive search for all available crystal structures of TCR/pMHC-I complexes restricted to HLA-A*02:01 with 9 residues epitopes was performed in the Protein Data Bank and IMGT/3D structure-DB (Kaas et al., 2004). Curated and calculated contacts between TCR and pMHC, for each complex, were obtained from IEDB-3D (Ponomarenko et al., 2011). An arbitrary cut-off of 85% and 60% was used to select TCR-interacting residues of the pMHC to retrieve electrostatic potential and ASA (Fig. 1) values, respectively. Information on included complexes is provided in Table A.1. Considering the nine key positions identified in crystal structures, we defined a group of seven regions over the pMHC-I surface (Fig. 1A). These regions, or “gates”, were defined considering the

specific contribution of each one of these residues to the pMHC-I surface. Three regions were defined covering the epitope surface. The contribution of epitope positions p4 and p5 were collected by two independent gates (G1 and G2). In the case of positions p6, p7 and p8, only one gate was defined, centered over p7 (G3). This was decided because p7 is much more exposed to the contact with the TCR, while p6 and p8 have a lower contribution to the pMHC-I surface. Other four gates were defined over selected MHC-I residues (G4, G5, G6 and G7). These seven key regions are in agreement with previously described “TCR footprints” for this allotype (Gras et al., 2009, 2012; Rudolph et al., 2006) and, therefore, will be probably involved in cross-reactive responses.

4.2. Construction of pMHC-I complexes

All our structural analysis were performed with pMHC-I complexes obtained through the previously described *D1-EM-D2* approach (Antunes et al., 2010). Briefly, only the FASTA sequence of the epitopes was recovered from the reference studies (Duan et al., 2012; Fyttili et al., 2008) and used as input to produce 3D structures of these epitopes, with PyMOL scripts. A “donor” structure of an empty HLA-A*02:01 was obtained by removing the epitope from a reference PDB structure (Protein Data Bank code 2V2W). The new pMHC-I structure, harboring the epitope of interest in the context of HLA-A*02:01, was then obtained by a combined sequence of molecular docking and energy minimization steps. These steps were performed with AutodockVina (Trott et al., 2010) and GROMACS 4.5.1 (Pronk et al., 2013), respectively. The accuracy and reliability of this *D1-EM-D2* approach was tested in previous studies (Antunes et al., 2010; Sinigaglia et al., 2013).

4.3. Electrostatic potential and ASA calculations over the pMHC-I complexes

Electrostatic potential for each pMHC-I structure was calculated with Delphi (Li et al., 2012), with custom parameters (e.g.: $indi = 1.0$, $exdi = 80.0$, $prbrad = 1.4$, $salt = 0.2$). Accessible surface area (ASA) from each pMHC-I complex was calculated with NACCESS V2.1.1 (<http://www.bioinf.manchester.ac.uk/naccess/>), which in a simplified explanation calculates the atomic accessible surface by rolling a probe of specific size around a van der Waals surface, of the selected residues. In this work, we used a probe size 1.40 Å.

4.4. Image acquisition and data extraction

Images of the electrostatic potential distribution over the “TCR-interacting surface” of each pMHC-I were obtained with the UCSF Chimera package from the Resource for Biocomputing, Visualization, and Informatics of the University of California, San Francisco (Pettersen et al., 2004). The “Electrostatic surfacing coloring” option of Chimera was used to import and visualize the electrostatic potential calculated with Delphi, using a range from -3 to $+3$ kT. Selected regions over these images were defined, and color histograms (RGB) of these areas were obtained with ImageJ 1.43u software (National Institute of Health, USA, <http://rsb.info.nih.gov/ij/>). In total, 42 values were obtained from the seven histograms of each image, such as color mean and standard deviation for each RGB component. Figures included in the article were edited with Adobe Photoshop CS2 v.9.0. program (Adobe, San Jose, CA).

4.5. Clustering analysis

As previously described, our prediction method was based on the use of pMHC-I structural features as input for multivariate statistical methods (Antunes et al., 2011). Originally, only information on electrostatic potential was used to define the clusters of putative

cross-reactive complexes. Now, we combined additional information on ASA values and improved our approach with the use of an R package (*pvcust*) to assess the uncertainty of the hierarchical cluster analysis (HCA) (Suzuki and Shimodaira, 2006). This package provides both bootstrap probability (BP) and approximately unbiased (AU) *p*-values, which are computed by multiscale bootstrap resampling, and has been shown to be less biased than other methods in typical cases of phylogenetic tree selection (Shimodaira, 2002). The “average” linkage method was used with “correlation” distance, and the number of bootstrap replications was set to 10,000. Results were plotted as dendrograms with bootstrap probabilities (BP) and approximately unbiased (AU) *p*-values. Main clusters were identified with *pvrct* ($\alpha=0.95$) and standard errors for AU *p*-values were obtained with *seplot*. Relational networks were plotted with the open-source platform Gephi (<https://gephi.org>).

This improvement adds a statistical validation to the dendrogram, enriching the discussion of the results, and avoiding unsubstantiated conclusions.

Acknowledgements

We thank Jader Peres da Silva, Artur Krumberg Schüller and Marina Roberta Scheid for collaboration in some steps of this work. This work was supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

References

- Antunes, D.A., Rigo, M.M., Silva, J.P., Cibulski, S.P., Sinigaglia, M., Chies, J.A.B., Vieira, G.F., 2011. Structural in silico analysis of cross-genotype-reactivity among naturally occurring HCV NS3-1073-variants in the context of HLA-A*02:01 allele. *Mol. Immunol.* 48, 1461–1467.
- Antunes, D.A., Vieira, G.F., Rigo, M.M., Cibulski, S.P., Sinigaglia, M., Chies, J.A.B., 2010. Structural allele-specific patterns adopted by epitopes in the MHC-I cleft and reconstruction of MHC:peptide complexes to cross-reactivity assessment. *PLoS ONE* 5, e10353.
- Brehm, M.A., Selin, L.K., Welsh, R.M., 2004. CD8 T cell responses to viral infections in sequence. *Cell. Microbiol.* 6, 411–421.
- Calis, J.J.A., Boer, R.J., Kesmir, C., 2012. Degenerate T-cell recognition of peptides on MHC molecules creates large holes in the T-cell repertoire. *PLoS Comput. Biol.* 8, e1002412.
- Calis, J.J.A., Maybeno, M., Greenbaum, J.A., Weiskopf, D., De Silva, A.D., 2013. Properties of MHC class I presented peptides that enhance immunogenicity. *PLoS Comput. Biol.* 9 (10), e1003266.
- Cornberg, M., Clute, S.C., Watkin, L.B., Saccoccio, F.M., Kim, S.-k., Naumov, Y.N., Brehm, M.A., Aslan, N., Welsh, R.M., Selin, L.K., 2010. CD8 T cell cross-reactivity networks mediate heterologous immunity in human EBV and murine vaccinia virus infections. *J. Immunol.* 184, 2825–2838.
- Donati, C., Rappuoli, R., 2013. Reverse vaccinology in the 21st century: improvements over the original design. *Ann. N. Y. Acad. Sci.* 1285, 115–132.
- Dormitzer, P.R., Grandi, G., Rappuoli, R., 2012. Structural vaccinology starts to deliver. *Nat. Rev. Microbiol.* 10, 807–813.
- Duan, Z.L., Li, Q., Wang, Z.B., Xia, K.D., Guo, J.L., Liu, W.Q., Wen, J.S., 2012. HLA-A*0201-restricted CD8+ T-cell epitopes identified in dengue viruses. *Virology* 439, 259.
- Fernandez-Vina, M.A., Falco, M., Sun, Y., Stastny, P., 1992. DNA typing for HLA class I alleles: I. Subsets of HLA-A2 and of -A28. *Hum. Immunol.* 33, 163–173.
- Frankild, S., de Boer, R.J., Lund, O., Nielsen, M., Kesmir, C., 2008. Amino acid similarity accounts for T cell cross-reactivity and for “holes” in the T cell repertoire. *PLoS ONE* 3, e1831.
- Fytilli, P., Dalekos, G.N., Schlaphoff, V., Suneetha, P.V., Sarrazin, C., Zauner, W., Zachou, K., Berg, T., Manns, M.P., Klade, C.S., Cornberg, M., Wedemeyer, H., 2008. Cross-genotype-reactivity of the immunodominant HCV CD8 T-cell epitope NS3-1073. *Vaccine* 26, 3818–3826.
- Gras, S., Burrows, S.R., Turner, S.J., Sewell, A.K., McCluskey, J., Rossjohn, J., 2012. A structural voyage toward an understanding of the MHC-I-restricted immune response: lessons learned and much to be learned. *Immunol. Rev.* 250, 61–81.
- Gras, S., Saulquin, X., Reiser, J.-B., Debeaupuis, E., Echasserieau, K., Kissenpennig, A., Legoux, F., Chouquet, A., Le Gorrec, M., Machillot, P., Neveu, B., Thielens, N., Malissen, B., Bonneville, M., Housset, D., Gorrec, M.L., Alerts, E., 2009. Structural bases for the affinity-driven selection of a public TCR against a dominant human cytomegalovirus epitope. *J. Immunol.* 183, 430–437.
- Halstead, S.B., 2013. Identifying protective dengue vaccines: guide to mastering an empirical process. *Vaccine* 31, 4501–4507.
- Hoof, I., Perez, C.L., Buggert, M., Gustafsson, R.K.L., Nielsen, M., 2010. Interdisciplinary analysis of HIV-specific CD8+ T cell responses against variant epitopes reveals restricted TCR promiscuity. *J. Immunol.* 184, 5383–5391.
- Jorgensen, J.L., Esser, U., Fazekas de St Groth, B., Reay, P.A., Davis, M.M., 1992. Mapping T-cell receptor-peptide contacts by variant peptide immunization of single-chain transgenics. *Nature* 355, 224–230.
- Kaas, Q., Ruiz, M., Lefranc, M.-P., 2004. IMGT/3D structure-DB and IMGT/structural query, a database and a tool for immunoglobulin, T cell receptor and MHC structural data. *Nucleic Acids Res.* 32, D208–D210.
- Kessels, H.W.H.G., de Visser, K.E., Tirion, F.H., Coccoris, M., Kruisbeek, A.M., Schumacher, T.N.M., 2004. The impact of self-tolerance on the polyclonal CD8+ T cell repertoire. *J. Immunol.* 172, 2324–2331.
- Li, L., Li, C., Sarkar, S., Zhang, J., Witham, S., Zhang, Z., Wang, L., Smith, N., Petukh, M., Alexov, E., 2012. DelPhi: a comprehensive suite for DelPhi software and associated resources. *BMC Biophys.* 5, 9.
- Linette, Gerald P., et al., 2013. Cardiovascular toxicity and titin cross-reactivity of affinity-enhanced T cells in myeloma and melanoma. *Blood* 122 (6), 863–871.
- Meijers, R., Lai, C.-C.C., Yang, Y., Liu, J.-H.H., Zhong, W., Wang, J.-H.H., Reinherz, E.L., 2005. Crystal structures of murine MHC class I H-2 D(b) and K(b) molecules in complex with CTL epitopes from influenza A virus: implications for TCR repertoire selection and immunodominance. *J. Mol. Biol.* 345, 1099–1110.
- Moise, L., Gutierrez, A.H., Bailey-Kellogg, C., Terry, F., Leng, Q., Abdel Hady, K.M., Verberkmoes, N.C., Sztejn, M.B., Losikoff, P.T., Martin, W.D., Rothman, A.L., De Groot, A.S., 2013. The two-faced T cell epitope: examining the host–microbe interface with JanusMatrix. *Hum. Vaccin. Immunother.* 9, 1577–1586.
- Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., Ferrin, T.E., 2004. UCSF chimera – a visualization system for exploratory research and analysis. *J. Comput. Chem.* 25, 1605–1612.
- Ponomarenko, J., Papangelopoulos, N., Zajonc, D.M., Peters, B., Sette, A., Bourne, P.E., 2011. IEDB-3D: structural data within the immune epitope database. *Nucleic Acids Res.* 39, D1164–D1170.
- Pronk, S., Pall, S., Schulz, R., Larsson, P., Bjelkmar, P., Apostolov, R., Shirts, M.R., Smith, J.C., Kasson, P.M., van der Spoel, D., Hess, B., Lindahl, E., 2013. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* 29, 845–854.
- Richter, S., Wenzel, A., Stein, M., Gabdoulline, R.R., Wade, R.C., 2008. webPIPSA: a web server for the comparison of protein interaction properties. *Nucleic Acids Res.* 36 (Suppl. 2), W276–W280.
- Rudolph, M.G., Stanfield, R.L., Wilson, I.A., 2006. How TCRs bind MHCs, peptides, and coreceptors. *Annu. Rev. Immunol.* 24, 419–466.
- Sandalova, T., Michaelsson, J., Harris, R.A., Odeberg, J., Schneider, G., Karre, K., Achour, A., Michaëlsson, J., Kärre, K., 2005. A structural basis for CD8+ T cell-dependent recognition of non-homologous peptide ligands: implications for molecular mimicry in autoreactivity. *J. Biol. Chem.* 280, 27069–27075.
- Selin, L.K., Nahill, S.R., Welsh, R.M., 1994. Cross-reactivities in memory cytotoxic T lymphocyte recognition of heterologous viruses. *J. Exp. Med.* 179, 1933–1943.
- Shimodaira, H., 2002. An approximately unbiased test of phylogenetic tree selection. *Syst. Biol.* 51, 492–508.
- Sinigaglia, M., Antunes, D.A., Rigo, M.M., Chies, J.A., Vieira, G.F., 2013. CrossTope: a curate repository of 3D structures of immunogenic peptide: MHC complexes. *Database (Oxford)*, bat002.
- Stone, J.D., Harris, D.T., Kranz, D.M., 2015. TCR affinity for pMHC formed by tumor antigens that are self-proteins: impact on efficacy and toxicity. *Curr. Opin. Immunol.* 33, 16–22.
- Suzuki, R., Shimodaira, H., 2006. *Pvcust*: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* 22, 1540–1542.
- Trott, O., Olson, A.J., News, S., 2010. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* 31, 455–461.
- Vieira, G.F., Chies, J.A.B., 2005. Immunodominant viral peptides as determinants of cross-reactivity in the immune system – can we develop wide spectrum viral vaccines? *Med. Hypotheses* 65, 873–879.
- Wedemeyer, H., Mizukoshi, E., Davis, A.R., Binnink, J.R., Rehmann, B., 2001. Cross-reactivity between hepatitis C virus and Influenza A virus determinant-specific cytotoxic T cells. *J. Virol.* 75, 11392–11400.
- Welsh, R.M., Fujinami, R.S., 2007. Pathogenic epitopes, heterologous immunity and vaccine design. *Nat. Rev. Microbiol.* 5, 555–563.
- Welsh, R.M., Selin, L.K., 2002. No one is naive: the significance of heterologous T-cell immunity. *Nat. Rev. Immunol.* 2, 417–426.
- Włodarczyk, M.F., Kraft, A., Chen, H., Selin, L.K., 2009. Protection or immunopathology upon heterologous virus infection: a decision of memory cells. *J. Immunol.* 182, 43.15 (meeting abstract supplement).
- Wucherpfennig, K.W., Call, M.J., Deng, L., Mariuzza, R., 2009. Structural alterations in peptide–MHC recognition by self-reactive T cell receptors. *Curr. Opin. Immunol.* 21, 590–595.
- Zhang, S., Bakshi, R., Suneetha, P., Fytilli, P., Antunes, D., Vieira, G., Jacobs, R., Klade, C., Manns, M., Kraft, A., Wedemeyer, H., Schlaphoff, V., Cornberg, M., 2015. Frequency, privacy and cross-reactivity of pre-existing HCV-specific CD8+ T-cells in HCV seronegative individuals: implication for vaccine responses. *J. Virol.* (in press).

Capítulo III

MatchTope: A tool to predict peptides complexed in Major Histocompatibility Complex I cross reactivity

(Artigo “in extenso” a ser submetido para a revista “*Scientific Reports*”)

MatchTope: A tool to predict peptides complexed in Major Histocompatibility Complex I cross reactivity

Marcus Fabiano de Almeida Mendes¹, Marcelo de Souza Bragatte¹, Martiela Vaz de Freitas¹, Ina Pöhner², Stefan Richter², Rebecca C. Wade^{2,3}, Francisco Mauro Salzano¹ & Gustavo Fioravanti Vieira¹

¹Bioinformatic Nucleus, Immunogenetics Laboratory, Genetics Department, Biosciences Institute, Federal University of Rio Grande do Sul, Caixa Postal 15053, 91501-970 Porto Alegre, RS, Brazil. ²Molecular and Cellular Modeling Group, Heidelberg Institute for Theoretical Studies (HITS), Heidelberg, Baden-Württemberg, Germany. ³Center for Molecular Biology (ZMBH), DKFZ-ZMBH Alliance and Interdisciplinary Center for Scientific Computing (IWR), Heidelberg University, Heidelberg, Germany. Correspondence and requests for materials should be addressed to G.F.V. (gusfioravanti@yahoo.com.br) or M.F.A.M. (marcus.famendes@gmail.com).

The development of new vaccines and treatments for cancer that target the immune system are a very active but challenging areas of research. There are several *in silico* tools for predicting immunogenicity based on analysis of the sequences of peptides that bind to the Major Histocompatibility Complex (MHC). Few of these bioinformatics tools make use of the three-dimensional structure of the peptides. Here, we have describe a new tool, MatchTope, for predicting the cross-reactivity of peptides by computing and analyzing the electrostatic potentials of MHC-peptide complexes. We validated MatchTope against previously published data and showed that it can be used to predict cross-reactivity between targets derived from several viral subtypes or even a possible cross response between tumor-derived peptides and self-peptides.

The immune system protects an organism against a wide range of exogenous pathogens, like viruses, bacteria and fungi, as well as endogenous pathological entities, like tumor cells¹. However, to avoid autoimmune diseases, the immune system should not generate a response to the organism's own healthy cells². Various types of cells, receptors, chemokines and interleukins are involved in the immune response and the complex interactions between these components drive the human immune system³.

One of the main ways by which the human immune system recognizes a pathogen and creates a response to it is by loading the Major Histocompatibility Complex (MHC) with a peptide (pMHC)⁴. The presented epitope can be derived from a self-protein, a protein from a pathogen or a tumor cell protein⁵. There are two main MHC types: MHC class I (MHC-I) and MHC class II (MHC-II). In humans, the MHC loci are called 'Human leukocyte antigens' (HLA).

Lymphocyte T cells are responsible for the interaction with pMHCs. There are two types of T lymphocytes, classified as CD8⁺ or cytotoxic and CD4⁺ or helper T cells. The cytotoxic CD8⁺ lymphocytes that interact with MHC-I are the focus of the current work. The presented epitope can be recognized as self or non-self. If the epitope is recognized as non-self, a signaling cascade will be triggered, which ultimately leads to the apoptosis of the infected or tumor cell⁶. This recognition is not strictly specific: the T-cell receptor (TCR) can not only recognize an exact match of the epitope but also similar ones. This latter event is called cross-reactivity⁷. Obviously, expanding the broadness of the recognition allows reduction of the number of TCRs. However, an epitope derived from a virus protein can mimic a self-epitope and thus trigger an autoimmune disease^{2,8}. Furthermore, the high similarity between proteins of normal cells and tumor cell proteins makes it difficult for lymphocyte CD8⁺ cells to properly respond to tumor cells^{1,9,10}.

Cross-reactivity becomes particularly important in the development of a new vaccine when it is crucial to check whether the vaccine will be effective against all subtypes of the pathogen. Likewise, when developing a new immunotherapeutic approach, it is necessary to ensure that the target will not trigger cross-reactivity with a self-protein. Unfortunately, it is impossible to test all possible complexes between epitopes and MHC-I in experiments, and therefore *in silico* analyses are helpful. Some cross-reactivity predictors are available, that use linear sequences as input, and were mostly designed for the prediction of allergic processes¹¹⁻¹³. However, it is already known that some epitopes show cross-response despite sharing

fewer than 50% of the amino acid residues in their linear sequence, which implies substantial difficulties for such predictors to correctly predict cross-reactivity^{14,15}. For this reason, we developed a new cross-reactivity prediction tool, MatchTope, which uses protein structural information to predict similarities between pMHC-I targets, therefore facilitating the development of new vaccines and immunotherapies. Using several available datasets, we verified that MatchTope achieves excellent agreement with experimental results, proving that this tool can greatly facilitate vaccine development for diverse diseases and cancer immunotherapeutic treatments.

Results

Opening the MatchTope black box. The MatchTope tool is based on the calculation of the molecular electrostatic potentials (MEP) of MHC class I loaded with different peptides, and then clustering the different peptide-MHC class I complexes based on the similarity of their MEPs. The concept of using the MEP as a measure of pMHC class I similarity was described in our previous articles^{16,17}.

Fig. 1 displays the steps involved in our analysis. We used PIPSA (Protein Interaction Property Similarity Analysis), a software that is an established tool for analyzing protein electrostatic interaction similarities (<http://pipsa.h-its.org/pipsa/>)¹⁸. We added some modifications to the standalone version to account for the cylindrical shape of the pMHC binding cleft. To cluster the targets by electrostatic similarity, MatchTope uses the R clustering package¹⁹.

Prior to the analysis, the user should provide a set of pMHC class I pdb files (a minimum of two files is required). Since only a few crystallographic complexes exist to date, the input pdb file will often be derived by modeling. The pdb file should include several columns holding the 3D coordinates of each protein atom and some additional information, such as occupancy, temperature factor, element name, charge, radius or other properties, depending on the source. Since some columns of non-standard pdb files for modeled complexes were found to cause problems during the PIPSA run, these are deleted in a pre-processing step using a bash script.

The next step involves a shift of the 3D coordinates of all provided complexes superimpose them. This process is important to ensure the comparison of the same

electrostatic regions in different pMHCs. To achieve this, we use a Python script to call the PyMOL 'Fitting' function²⁰. This function superimposes the pdb input with a predefined model pdb structure.

After the fitting process, MatchTope starts to calculate the electrostatic similarity of the complexes by using the PIPSA standalone version. PIPSA first calculates the MEPs using the University of Houston Brownian Dynamics (UHBD) program²¹. There are several available programs for calculating the MEP of a protein, but we have chosen UHBD due to its convenient short runtime compared to the APBS program²². PIPSA creates a 'skin' around each pMHC and then the MEPs of the pMHC complexes are compared. Besides calculating overall electrostatic similarities for the full proteins in the complete skins, the algorithm also allows for calculating similarities in a focused region. For this study, a cylinder in the cleft of the pMHC was considered and only regions of the protein skins residing within this cylinder were used for computing similarity indices, as shown in Fig. 2. Using this focused region, we can reduce the noise and thereby avoid erroneous clustering of the results.

The final part of the analysis, the clustering process, uses the similarity indices calculated during the PIPSA run as input. To group electrostatically similar pMHCs together, MatchTope uses an R package called 'cluster', which performs a hierarchical clustering of the input data and a package called 'gplots' to create a heat map with a color gradient to visualize the level of similarity between different targets. The cluster package requires some user-defined arguments. We used maximum distance as the metric and complete-linkage clustering as the linkage criterion. This choice of arguments yielded the best results for our data.

MatchTope validation. To validate MatchTope, we used two data sets from previously published articles^{23,24}. A list with all epitopes is shown in Table S1, and data on the superposition of input and model pdbs are shown in Table S2. The low RMSDs (average 0.019 Angström) indicate that all structures were well superimposed. The first data set is from a Hepatitis C target cross-reactivity study²⁴. In this study, 28 epitopes presented to HLA-A*02-01 were tested against a wild type epitope from the virus. We modeled these targets using the DockTope tool²⁵. The clustering of the structures on the basis of MEP resulted in two large groups and one outlier (Fig. 3). One cluster was previously demonstrated to show cross-reactivity in *in vitro* analyses, proving that our tool grouped the targets correctly. The second

group does not have *in vitro* data in support of the predicted cross-reactivity, but manual analysis of the complexes reveals striking similarities between them. This indicates that members of the group would probably trigger cross-reactivity if tested against a specific TCR. The outlier is a pMHC complex that is very distinct in comparison with the other 27 pMHCs, and thus, this result fits well with our expectations.

The second data set we have chosen was from a study on Dengue virus²³. Eight pMHCs with peptides derived from two different proteins, NS4a and NS4b, of the four dengue virus serotypes, are considered. *In vitro* data showed that the epitopes generated from NS4b had cross-reactivity, while epitopes from NS4a did not. Using DockTope to model the complexes and MatchTope to compute their MEP similarity resulted in two distinct clusters, exactly corresponding to NS4a- and NS4b (Fig. 3). The NS4b group shows very small electrostatic distances, indicating that the group members were highly similar. The NS4a group, on the other hand, appears dispersed, showing considerable intragroup distances, indicating low similarity. That explains why this group will probably not trigger a cross response. Again, we found that the results of the *in silico* analysis generated by our tool matched the *in vitro* data.

In addition, we also used a set of distinct complexes studied *in vitro* and deposited in the CrossTope data bank (<http://crosstope.com/>)²⁶, for which cross-reactivity has not yet been demonstrated experimentally. Images of the MEPs on to the molecular surfaces of the pMHCs are available, in Cross Tope, and it is possible to detect similar patterns and manually group pMHCs into various clusters using these similarities. The automated MatchTope analysis again led to the expected result with clustering of pMHCs with a similar electrostatic charge distribution in the same group (data not shown).

We tested around 10,000 different combinations of settings for PIPSA as well as statistical parameterization. For PIPSA, we varied the probe size, the skin thickness and the radius of the cylindrical shape by which the focused region is defined and, for the statistical analysis, we tested a number of clustering options. In this tool, we finally implemented the parameters and settings that yielded the best results.

MatchTope availability. Upon publication of this article, a standalone version of MatchTope will be made available for download free of charge via the CrossTope interface at <http://www.crosstope.com/Home/Tools>. In addition, users can send epitope sequences or

pdb structures of pMHCs to MatchTopetool@gmail.com; we will then perform the analysis described herein and send the results back. In the near future, we will in addition release a MatchTope web server version, where pdb files or complexes modelled using DockTope can be uploaded and then directly subjected to MatchTope analysis. The results will then be displayed on the web page and be available for download.

Discussion

We here describe a fully automated tool for comparing and clustering pMHCs by MEP similarity for cross-reactivity prediction. Using previously published data sets as input data, we were able to correctly group the targets showing cross-reactivity. MatchTope allows the user to analyze up to 100 pMHC structures at once, calculates the MEPs, and groups similar complexes together. The resulting distances in electrostatic potential space enable the user to draw conclusions about whether cross-reactivity is likely to occur for the analyzed complexes or not.

MatchTope makes use of various bash scripts, R scripts and a customized version of PIPSA. In addition, as an external tool, PyMOL is required. We recommend using the DockTope tool for modeling targets²⁵. With this tool, the user can model peptides complexed with HLA-A*02:01, HLA-B*27:05, H2-Db or H2-Kb, but any pMHC of class I allele can be used as input for MatchTope. MEPs are always calculated with the same settings, even if the pMHC allele differs between different complexes. The pdb file with the selected focused region used to determine the electrostatic distances is defined by placing a cylinder in the pMHC cleft, and can be exported to a separate pdb file.

In a previous article¹⁷, we discussed how one of the TCR variable domains CDR3, recognizes if peptides originate from a self-protein or a non-self-protein. In our new tool MatchTope, we solely make use of these regions to calculate the MEP similarities. We did not consider the rest of the complex because this would only increase the noise in the analysis. We plan to include topographic features combined with the MEP data in a future MatchTope implementation, to improve the robustness of the analysis.

The final step in our determinations involves clustering of the structures by electrostatic potential similarity as determined by PIPSA using the R cluster package. This package allows for the choice of a variety of metrics and linkage criteria, but the best results were achieved using maximum distance as the metric and complete-linkage clustering as the linkage

criterion. We confirmed the suitability of our parameter choice by clustering published data^{24,27} into groups consistent with the results of *in vitro* analysis.

In our data bank, CrossTope (<http://www.crosstope.com/>)²⁶, hundreds of immunogenic pMHC models are available, for which a pdb file can be downloaded and images of the MEPs can be viewed. It had previously been observed that these immunogenic pMHCs show common patterns of electrostatic charge distributions when manually comparing the pictures. With MatchTope, however, a comparison on a much larger scale becomes feasible. MatchTope was able to point to us to similarities between immunogenic targets which were not previously observed, and thus may be helpful in the field of reverse vaccine development.

The field of cancer immunology is rapidly developing and immunotherapeutic approaches are becoming more and more common, showing promising results. One methodology makes use of TCR modifications to enhance affinity against tumor-specific peptides^{1,9}. However, one major risk of using the TCR modification approach is the cross-reactivity with a normal cell presenting self-peptides. A well-known case²⁸ is the cross-reactivity between the melanoma-associated antigen MAGE-A3 and a titin-derived antigen expressed by healthy cardiac cells, which led to the death of two patients. These two peptides have a low sequence similarity, sharing just 5 amino acid residues out of 9, but X-ray crystallography showed structural similarity between them. MatchTope is able to demonstrate their similarity without making use of crystallography, proving that is a powerful tool to predict an undesirable cross-reactivity.

This result, together with our validation, demonstrates that similarity between pMHCs can be predicted from the MEP of the cleft region of the structure only. Since electrostatic similarity can trigger cross-reactivity events, our tool can be used as a cross-reactivity predictor. MatchTope overcomes issues inherent to predictors using just linear sequences as input. Even with low sequence similarity, *e.g.* less than 50% shared amino acid residues, our tool can properly cluster the targets and seems less prone to yield erroneous classifications.

Working with this tool as a cross-reactivity predictor may open new prospects in the field of vaccine prediction. For example, in a TCR modification-based vaccine development approach, combining MatchTope with our previously published modeling tool DockTope can substantially enhance the process of finding new safe vaccine candidates in a shorter time

span. It is also possible to predict cross-reactivity between a tumor epitope and a self-epitope, which can be used in the context of research on new immunotherapies.

Methods

MatchTope Automation. MatchTope is a software built to seek similarities in the MEPs of pMHC molecules and to group similar MEP patterns by hierarchical clustering. In order to do this, we developed a workflow involving 3 bash scripts, a Python script, the PyMOL program²⁰, a modified PIPSA standalone version¹⁸ and 2 R packages, to perform the following steps: (i) edit pdb files to remove unnecessary columns; (ii) superimpose all the pdb files; (iii) use these pdb files as input for PIPSA to calculate the MEPs and corresponding similarity indices; (iv) use the PIPSA results as input for the R package to perform hierarchical clustering. The MatchTope tool was tested on Linux Ubuntu 14.04 and Ubuntu 16.04 systems. The average run time of MatchTope for an input of 30 pdb files is 4 minutes.

Pre-processing the input pdb files. Since the pMHC pdb files are often the output of modeling software, they typically have some columns with unnecessary information, which can cause problems for the PIPSA software. To avoid any issues, a bash script removes these columns using shell instructions. After this process, the pdb files retain nine columns, namely the ATOM or HETATOM identifier for proteins and other groups, respectively, atom number, type of atom, the corresponding amino acid residue, chain information, amino acid residue number and the Cartesian x-, y- and zcoordinates.

Fitting. To avoid the problem of comparing different regions of different pMHCs due to a nonuniform orientation in the 3D space, we implemented a fitting routine in a Python script making use of the PyMOL software²⁰. We employ a pdb model that gives the reference position and all input structures are fitted to the orientation and positions of this model. The script repeats this superposition process twice to ensure a good result.

PIPSA calculation. PIPSA first computes similarity indices for the electrostatic potential analytically from the pdb files, making use of monopole and dipole terms. Hydrogen atoms are added using WHATIF (<https://swift.cmbi.umcn.nl/servers/html/index.html>) as necessary²⁹. Next, the input for the UHBD calculation is generated and the electrostatic potential grids computed with UHBD. The PIPSA program then computes the Hodgkin similarity index for all pairs of electrostatic potential grids^{18,30}. This is done on the molecular

skin and within a 20 Angström radius cylindrical region, defined in the pMHC cleft using the 3D coordinates of this cleft. Due to the fitting step, the program can use the same 3D coordinates for all pMHCs. The 'skin' represents the remaining layer, after excluding any region inside the solvent-accessible surface area defined with a certain probe radius, and has a defined thickness. Everything outside this region is also excluded. Corresponding points on the potential grids within the skins of the two proteins to be compared are used for computing similarity indices. Potential values lying outside of this skin or outside of the cylinder created in the region of interest of the pMHCs will not be used. The thickness of the skin and the probe radius are adjusted to 15 and 0.5 Angström, respectively, for best results. Images are generated with UCSF Chimera 1.12.

Hierarchical clustering and heat map. After the PIPSA calculation is finished, the program uses the resulting similarities of the MEPs as input to the R package. Using the 'cluster' package, R creates a hierarchical clustering of the results, grouping most similar pMHCs in the same cluster. The package uses the maximum distance as a metric and complete-linkage clustering as the linkage criterion. Together with the clustering represented in a tree-based format, a heat map is displayed, calculated using the 'gplots' package and using a color gradient to visualize how similar pMHCs are.

Validation methodology. To validate our tool, we used two distinct data sets. All targets were nonamers and modeled using the DockTope software²⁵. A list of all epitope sequences used in our validation step is shown in Table S1. We modeled all epitopes in HLA*A-0201 complex options, using the standard settings. We used the given interferon-gamma results from the published data sets to determine cross-reactivity and confirm the validity of our *in silico* analysis^{23,24}. The interferon-gamma information is available on the respective articles.

References

1. Zamora, A. E., Crawford, J. C. & Thomas, P. G. Hitting the target: How T cells detect and eliminate tumors. *J Immunol* **200**, 392-399 (2018).
2. O'Byrne, K. J. & Dalglish, A.G. Chronic immune activation and inflammation as the cause of malignancy. *Br J Cancer* **85**, 473-483 (2001).
3. Uematsu, S. & Akira, S. Toll-like receptors and type I interferons. *J Biol Chem* **282**, 15319-15323 (2007).
4. Purcell, A. W., Croft, N. P. & Tschärke, D. C. Immunology by numbers: Quantitation of antigen presentation completes the quantitative milieu of systems immunology! *Curr Opin Immunol* **40**, 88-95 (2016).
5. Sei, J. J. *et al.* Peptide-MHC-I from endogenous antigen outnumber those from exogenous antigen, irrespective of APC phenotype or activation. *PLoS Pathog* **11**, e1004941 (2015).
6. Attaf, M. *et al.* The T cell antigen receptor: The Swiss army knife of the immune system. *Clin Exp Immunol* **181**, 1-18 (2015).
7. Regner, M. Cross-reactivity in T-cell antigen recognition. *Immunol Cell Biol* **79**, 91-100 (2001).
8. Schwimbeck, P. L. *et al.* Molecular mimicry and myasthenia gravis. An autoantigenic site of the acetylcholine receptor alpha-subunit that has biologic activity and reacts immunochemically with herpes simplex virus. *J Clin Invest* **84**, 1174-1180 (1989).
9. Antunes, D. A. *et al.* Interpreting T-cell cross-reactivity through structure: Implications for TCR-based cancer immunotherapy. *Front Immunol* **8**, 1210 (2017).
10. Pardoll, D. M. The blockade of immune checkpoints in cancer immunotherapy. *Nat Rev Cancer* **12**, 252-264 (2012).
11. Moise, L. *et al.* iVAX: An integrated toolkit for the selection and optimization of antigens and the design of epitope-driven vaccines. *Hum Vaccin Immunother* **11**, 2312-2321 (2015).
12. Negi, S. S. & Braun, W. Cross-React: A new structural bioinformatics method for predicting allergen cross-reactivity. *Bioinformatics* **33**, 1014-1020 (2017).
13. Zhang, Z. H. *et al.* AllerTool: A web server for predicting allergenicity and allergic cross-reactivity in proteins. *Bioinformatics* **23**, 504-506 (2007).

14. Cornberg, M. *et al.* CD8 T cell cross-reactivity networks mediate heterologous immunity in human EBV and murine vaccinia virus infections. *J Immunol* **184**, 2825-2838 (2010).
15. Cornberg, M. *et al.* Narrowed TCR repertoire and viral escape as a consequence of heterologous immunity. *J Clin Invest* **116**, 1443-1456 (2006).
16. Antunes, D. A. *et al.* Structural allele-specific patterns adopted by epitopes in the MHC-I cleft and reconstruction of MHC:peptide complexes to cross-reactivity assessment. *PLoS One* **5**, e10353 (2010).
17. Mendes, M. F. *et al.* Improved structural method for T-cell cross-reactivity prediction. *Mol Immunol* **67**, 303-310 (2015).
18. Wade, R. C., Gabdouliline, R. R. & De Rienzo, F. Protein interaction property similarity analysis. *Int J Quantum Chem* **83**, 122-127 (2001).
19. Ihaka, R. & Gentleman, R. R: A language for data analysis and graphics. *J Computat Graph Stat* **5**, 299 (1996).
20. Schrodinger, L. L. C. The PyMOL Molecular Graphics System, Version 1.8. (2015).
21. Madura, J. D. *et al.* Electrostatics and diffusion of molecules in solution: Simulations with the University of Houston Brownian Dynamics program. *Comp Physics Communic* **91**, 57-95 (1995).
22. Baker, N. A. *et al.* Electrostatics of nanosystems: Application to microtubules and the ribosome. *Proc Natl Acad Sci U S A* **98**, 10037-10041 (2001).
23. Duan, Z. L. *et al.* HLA-A*0201-restricted CD8+ T-cell epitopes identified in dengue viruses. *Virology* **9**, 259 (2012).
24. Fyttili, P. *et al.* Cross-genotype-reactivity of the immunodominant HCV CD8 T-cell epitope NS3-1073. *Vaccin* **26**, 3818-3826 (2008).
25. Rigo, M. M. *et al.* DockTope: A web-based tool for automated pMHC-I modelling. *Sci Rep* **5**, 18413 (2015).
26. Sinigaglia, M. *et al.* CrossTope: A curate repository of 3D structures of immunogenic peptide: MHC complexes. Database (Oxford) **2013**, bat002 (2013).
27. Clute, S. C. *et al.* Cross-reactive influenza virus-specific CD8+ T cells contribute to lymphoproliferation in Epstein-Barr virus-associated infectious mononucleosis. *J Clin Invest* **115**, 3602-3612 (2005).

28. Raman, M. C. *et al.* Direct molecular mimicry enables off-target cardiovascular toxicity by an enhanced affinity TCR designed for cancer immunotherapy. *Sci Rep* **6**, 18851 (2016).
29. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res* **28**, 235-242 (2000).
30. Good, A. C. The calculation of molecular similarity: Alternative formulas, data manipulation and graphical display. *J Mol Graph* **10**, 144-151 (1992).
31. Pettersen, E. F. *et al.* UCSF Chimera – a visualization system for exploratory research and analysis. *J Comput Chem* **25**, 1605-1612 (2004).

Acknowledgements

This work was supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). We also thank Aruã Ramos Metello de Assis, Mauricio Meneghatti Rigo, and Dinler Amaral Antunes for their involvement in this project.

Author Contributions

M.F.A.M., M.S.B., M.V.F. and G.F.V. conceived the study. M.F.A.M. wrote the paper. M.F.A.M., R.C.W. and S.R. developed the MatchTope scripts. M.F.A.M., M.V.F. and M.S.B. conducted experiments. M.F.A.M., M.V.F., I.P., G.F.V. and F.M.S. analysed and interpreted the data. I.P., G.F.V., R.C.W. and F.M.S. were responsible for general revision of the manuscript.

Additional Information

Supplementary information accompanies this paper.

Competing Interests: The authors declare no competing interests.

Figure Captions

Fig. 1. The flowchart of the MatchTope tool showing the analysis process from the first step of inputting the pdb files to the final step of generating the results. Each step is described in greater detail in the Methods section.

Fig. 2. A sample image of one pMHC showing the exact region and size of the cylindrical region that is used by PIPSA to compare the MEPs of various pMHCs. The pMHC is shown in cartoon representation with the alpha chain in light blue color, the beta chain in dark blue color, the epitope in orange with amino acid side chains in stick representation and the cylindrical region used for calculation shown by a gray semi-transparent surface. The pMHC was modeled with the DockTope tool using a dengue epitope as input.

Fig. 3. Final result of the MatchTope analysis: The hierarchical clustering is represented in a tree-based format together with a heat map representation of the resulting similarities. To test the tool, we used previously published data. Epitopes that have the name highlighted with a green box are derived from Hepatitis C virus (HCV) and show cross-reactivity. The yellow boxes indicate HCV-derived epitopes that lack cross-reactivity against the epitopes in green boxes. The red epitope is an outlier. Epitopes highlighted with brown and blue boxes are derived from proteins of the four serotypes of the Dengue virus; the ones shown in brown have no reactivity between them and the blue ones do show cross-reactivity. These boxes were manually added to facilitate interpretation of the results and are not implemented in the program. The color coding of the heat map, as indicated by the color key in the upper left corner, ranges from red color to indicate highest similarity to purple, indicating lowest similarity in the current data set.

Table S1. Table with the list of MHC peptide linear sequences and their respective names.

Table S2. RMSD values (in Angström) of pMHCs used for MatchTope validation compared to the pdb model after fitting. The low RMSD values shown here (below 0.1 Angström) prove that all pMHCs are in the same position. The last row presents the mean overall results.

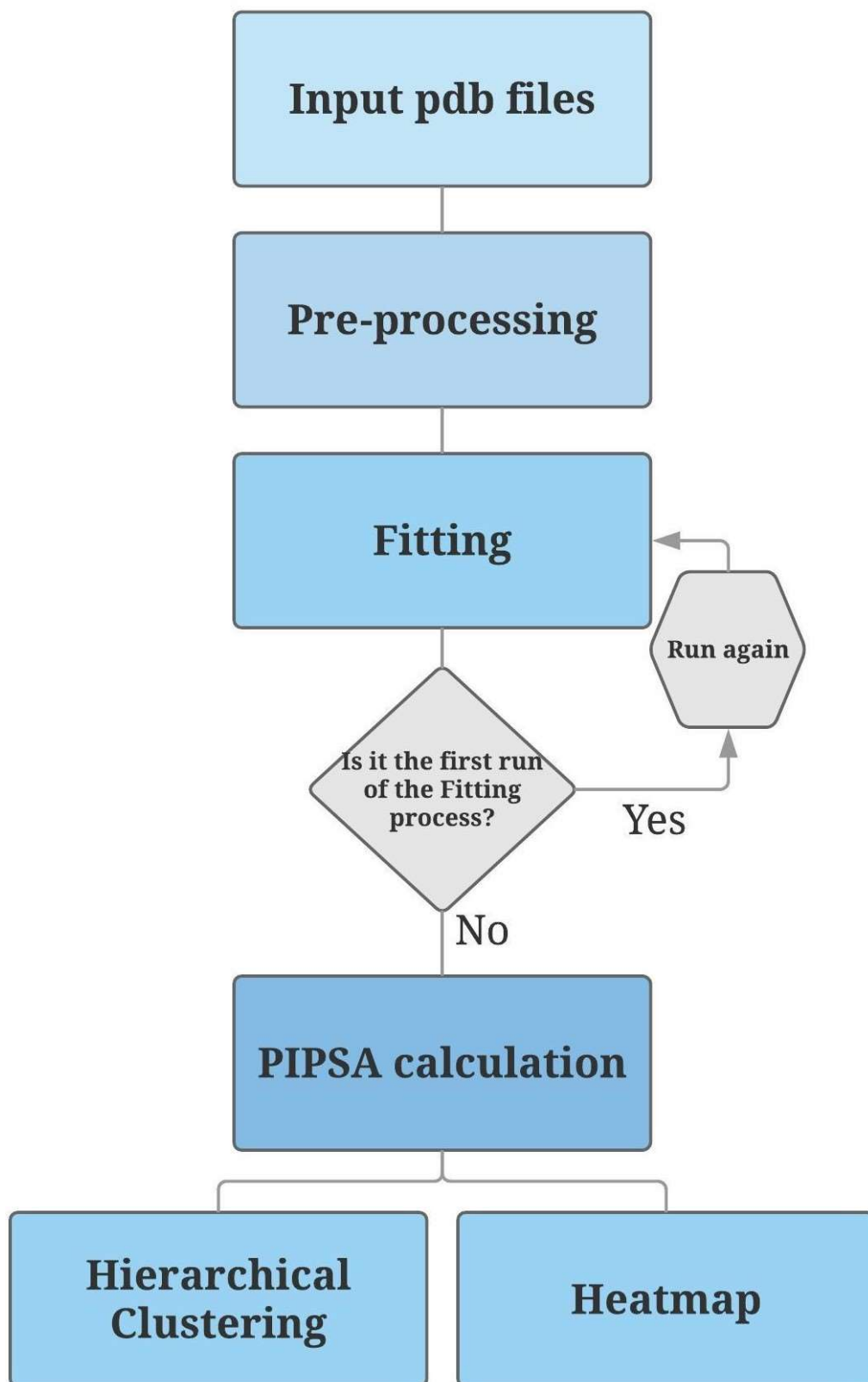


Figure 1

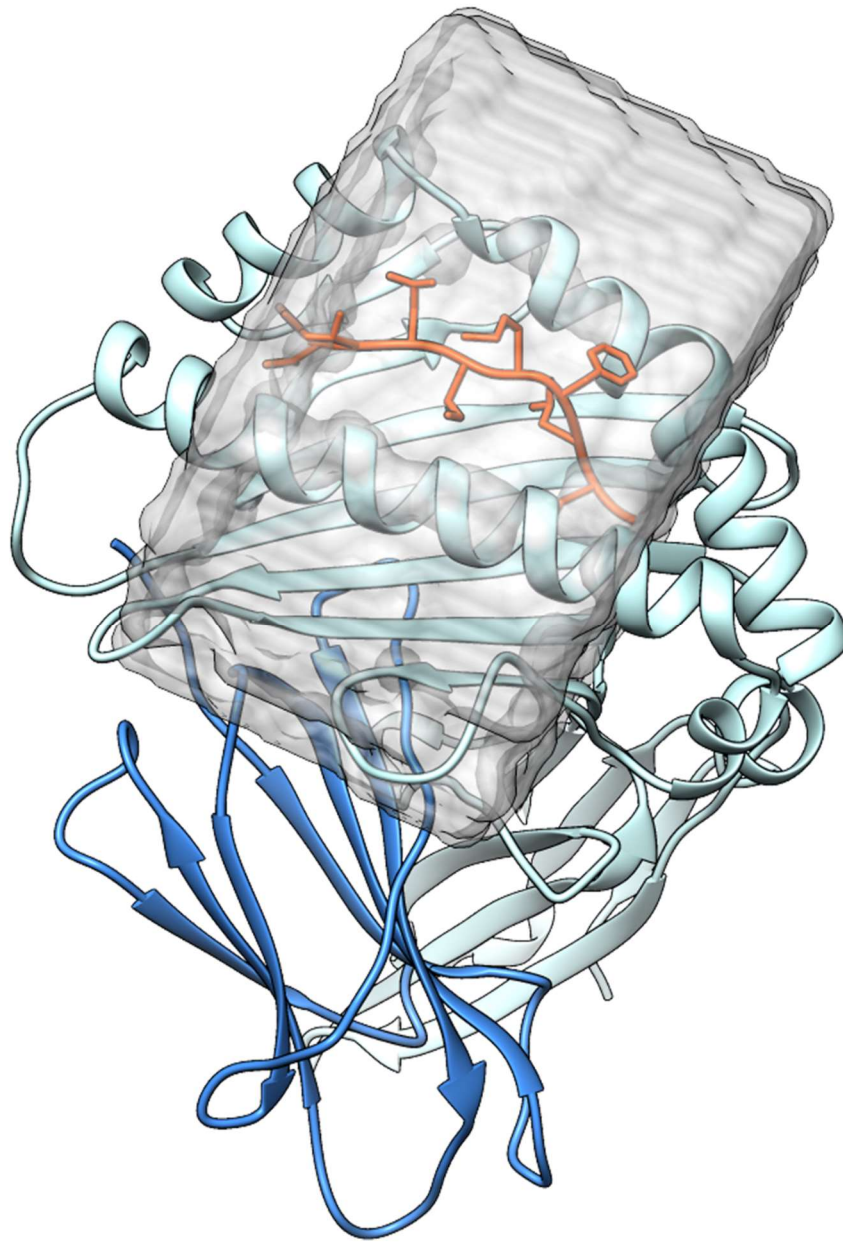


Figure 2

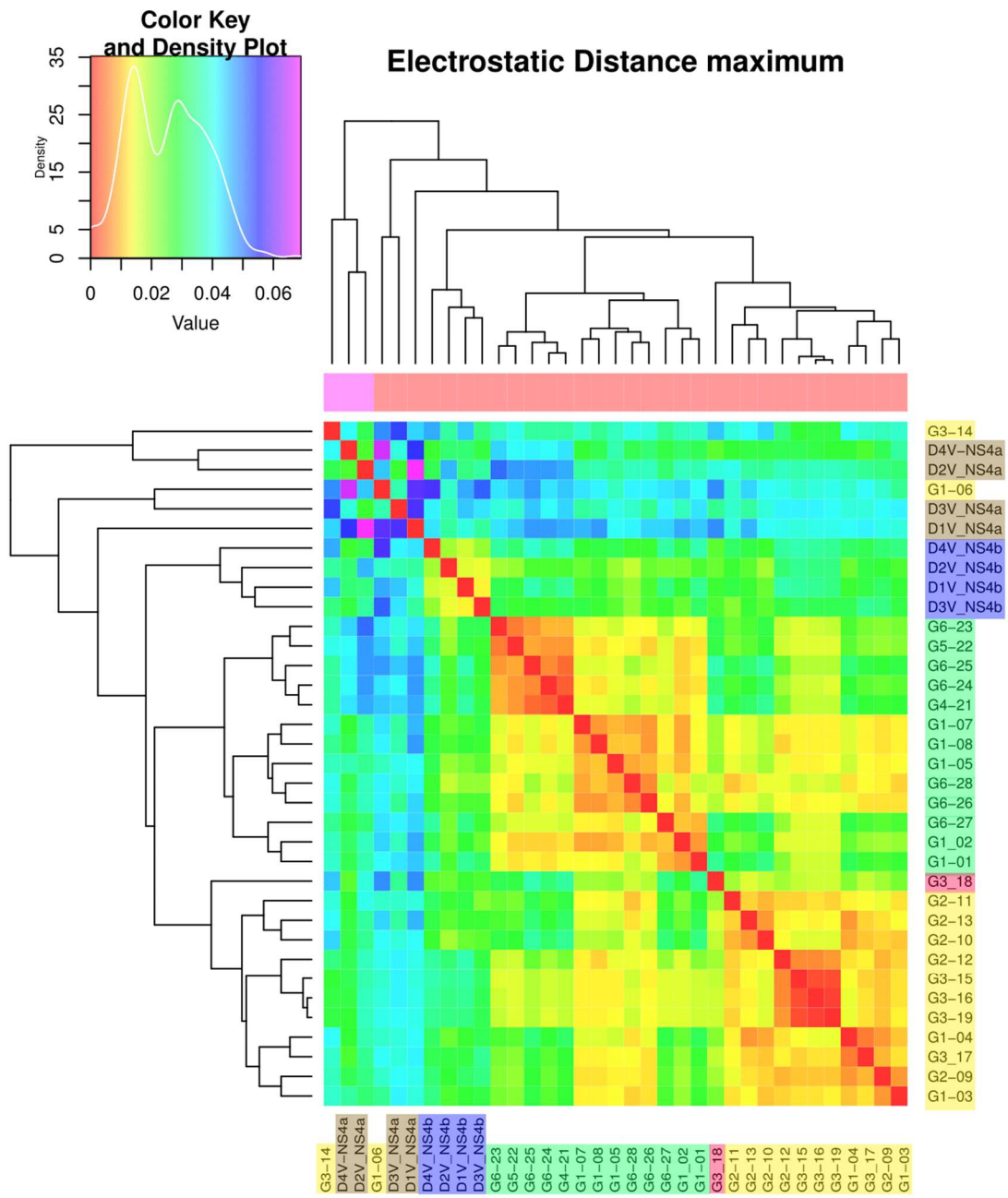


Figure 3

Hepatitis C epitopes			Dengue epitopes		
Genotype 1	G1_01	CVNGVCWTV	Serotype 1	D1V_NS4a	GLLFMILTV
	G1_02	CTNGVCWTV		D1V_NS4b	LMRTTWAL
	G1_03	CVSGACWTV	Serotype 2	D2V_NS4a	AILTVVAAT
	G1_04	CISGVCWTV		D2V_NS4b	LMMRTTWAL
	G1_05	CINGACWTV	Serotype 3	D3V_NS4a	GILTLAAIV
	G1_06	CVNGACMTV		D3V_NS4b	LLMRTSWAL
	G1_07	CINGVCWSV	Serotype 4	D4V_NS4a	TILTIIGLI
	G1_08	CINGVCWSI		D4V_NS4b	LMRTTWAF
Genotype 2	G2_09	CISGVLWTV			
	G2_10	TISGVLWTV			
	G2_11	SISGVLWTV			
	G2_12	SIAGVLWTV			
	G2_13	TISGILWTV			
Genotype 3	G3_14	TVGGVTWTV			
	G3_15	SVGGVMWTV			
	G3_16	TIGGVMWTV			
	G3_17	AIGGVMWTV			
	G3_18	TVGDVMWTV			
	G3_19	TVGGVMWTV			
	G3_20	TVGGVIWTV			
Genotype 4	G4_21	AVNGVMWTV			
Genotype 5	G5_22	CINGVMWTL			
Genotype 6	G6_23	SINGVMWTV			
	G6_24	AINGVMWTV			
	G6_25	TVNGVMWTV			
	G6_26	AVNGVLWTV			
	G6_27	TINGVLWTV			
	G6_28	TVNGVLWTV			

Table S1. Table with the list of linear peptide sequences and their respective names.

Epitope	RMSD
D1V_NS4a	0.016
D1V_NS4b	0.023
D2V_NS4a	0.018
D2V_NS4b	0.016
D3V_NS4a	0.014
D3V_NS4b	0.021
D4V_NS4a	0.010
D4V_NS4b	0.017
G1_01	0.016
G1_02	0.022
G1_03	0.018
G1_04	0.029
G1_05	0.040
G1_06	0.029
G1_07	0.018
G1_08	0.009
G2_09	0.020
G2_10	0.025
G2_11	0.021

Epitope	RMSD
G2_12	0.020
G2_13	0.015
G3_14	0.017
G3_15	0.021
G3_16	0.010
G3_17	0.021
G3_18	0.015
G3_19	0.026
G3_20	0.020
G4_21	0.017
G5_22	0.013
G6_23	0.027
G6_24	0.017
G6_25	0.026
G6_26	0.016
G6_27	0.017
G6_28	0.012
Mean	0.019

Table S2. RMSD values (in Angström) of pMHCs used for MatchTope validation compared to the pdb model after fitting. The low RMSD values shown here (below 0.1 Angström) prove that all pMHCs are in the same position. The last row presents the mean overall results.

Capítulo IV

Discussão Geral e Referências Complementares

Discussão Geral

O objetivo deste trabalho foi a criação de uma ferramenta de cálculo de similaridade utilizando dados numéricos de campo eletrostático como entrada, para assim realizar uma predição de possíveis alvos que possam desencadear o fenômeno de reatividade cruzada ou que apresentem similaridade a alvos imunogênicos e possam ser utilizados em abordagens vacinais e imunoterapêuticas.

No âmbito do desenvolvimento de vacinas com o auxílio da imunoinformática, podemos voltar aos primórdios do final da década de 90, onde começaram a surgir, graças ao *boom* de sequenciamentos e disponibilização de dados biológicos que houve naquele momento, ferramentas voltadas para a imunologia (Lefranc, 2014). Surge então o termo vacinologia reversa, onde começou-se a usar a sequência do proteoma dos patógenos na busca de alvos para o desenvolvimento de novas vacinas (Rappuoli et al., 2016). Com o passar do tempo houve o avanço da tecnologia de sequenciamento e do poder computacional dos processadores, permitindo assim as análises utilizando *machine learning*, e originando ferramentas como NetMHC (Lundegaard et al., 2008), SYFPEITHI (Rammensee et al., 1999), o banco de dados do PDB (Berman et al., 2000) e o IEDB (Vita et al., 2015). Com isso, muito do desenvolvimento de novas vacinas passam agora primeiramente por análises *in silico* para apenas depois de um resultado positivo nesta etapa progredir-se para análises *in vitro/in vivo*.

Na imunoinformática, tem-se poucas ferramentas que conseguem prever se linfócitos públicos, isto é, presentes em grande parte da população, vão gerar resposta ou não se confrontados com determinados pMHCs. Ferramentas de predição de imunogenicidade baseada em dados estruturais são escassas (Moise et al., 2015; Z. H. Zhang et al., 2007), sendo que apenas uma utiliza a estrutura tridimensional da proteína como entrada, e mesmo assim é um preditor de reatividade cruzada de alvos imunogênicos apenas no contexto de anticorpos contra alergênicos (Z. H. Zhang et al., 2007). Neste sentido, uma ferramenta de predição de reatividade cruzada, utilizando dados estruturais como entrada e no contexto da resposta citotóxica é algo inédito e que pode ser muito bem utilizado em vários estudos, possuindo assim um grande nicho a ser explorado.

A disponibilidade de cristais de pMHC-I é pequena frente ao grande número de complexos possíveis, muito devido ao elevado custo, tempo longo e necessidade de mão de

obra especializada para realizar a cristalografia. Devido a isso surgiram algumas ferramentas que podem ser utilizadas em sua modelagem (Khan & Ranganathan, 2010; Todman et al., 2008). Dentre estas destaca-se o DockTope (Rigo et al., 2015), que é uma ferramenta que modela peptídeos apresentados no contexto do MHC-I. O DockTope modela quatro alelos de MHC-I, sendo 2 de humanos (HLA*A-02:01 e HLA*B-27:05) e dois de murinos (H2-Kb e H2-Db).

Graças ao DockTope possuímos então a capacidade de obtermos vários pMHCs modelados em um curto espaço de tempo. Algo que demoraria em torno de 6 meses para ser produzido utilizando a cristalografia por Raio X, pode ser obtido em torno de 3 horas utilizando a nossa ferramenta. Esta técnica foi validada reproduzindo cristais com um RMSD médio de 1,9 Å. Com isto torna-se possível a análise de grande quantidade de alvos modelados, abrindo assim uma gama de possibilidade em estudos que analisam a base estrutural da imunogenicidade, com potencial uso do MatchTope. Para maiores informações sobre o DockTope leia-se o Anexo 1 no final desta Tese.

Com uma grande quantidade de dados em mãos e um *know-how* em estudos utilizando pMHC-I, nosso grupo começou a estudar possíveis aplicações para estas estruturas modeladas. No início utilizou-se arquivos em formato de imagens, com o equivalente a uma foto do campo eletrostático do pMHC, como forma de calcular similaridade, sendo que utilizamos regiões que eram normalmente discordantes. Eram consideradas regiões onde diversas estruturas possuísem um maior grau de diferença entre si. Com isso foram escolhidos em torno de 7 zonas diferentes, e foram extraídos a média e o desvio padrão dos valores de cores das imagens. Estes dados serviam como entrada para a análise estatística, havendo a geração de um dendrograma em que se indicava a similaridade entre os alvos (Antunes et al., 2010).

Além desta técnica, também foi criado um banco de dados chamado CrossTope (Sinigaglia et al., 2013), para hospedar pMHCs modelados por nosso grupo, sendo o critério de seleção dos epítomos a sua imunogenicidade, avaliada por artigos recuperados no banco de dados do IEDB (Vita et al., 2015). Estão disponíveis estruturas modeladas em quatro alelos diferentes, os mesmos alelos utilizados pelo DockTope. Este banco de dados é aberto a qualquer pessoa, de forma gratuita, e continuamente vem sendo atualizado com a inclusão de novas estruturas.

O Capítulo II desta tese mostra um avanço na nossa metodologia de análise de similaridade, sendo esta versão denominada de *legacy*. A ideia de utilizar a imagem do campo eletrostático e extrair os valores de cores (caracterizados pela sigla RGB, do inglês *Red Green and Blue*) continua válida, porém a metodologia sofreu várias alterações. Em um primeiro momento, adquirimos um total de 37 cristais de complexos TCR:pMHC, sendo os MHC expressos no contexto de A*02:01, para realizarmos um levantamento sobre as regiões em que as alças do TCR contatam o pMHC. A partir destas informações, foram selecionadas sete regiões, sendo três destas na porção onde se localiza o epítopo (aminoácido 3, 4 e um quadrante maior abrangendo os aminoácidos 5, 6 e 7) e quatro quadrantes nas alças da fenda, sendo dois na alfa 1 e o restante na alfa 2. Com isto, tem-se a certeza que são regiões de grande importância para o reconhecimento do linfócito T CD8⁺, aumentando assim a confiabilidade das análises.

Outra modificação desta técnica, em comparação com a técnica previamente utilizada, foi a forma de obtenção dos campos eletrostáticos. Na técnica atual utilizou-se um software separadamente, o Delphi (L. Li et al., 2012). Esta ferramenta implementa a equação de Poisson-Boltzmann para o cálculo dos valores eletrostáticos do pMHC. Ao usar este software separadamente foi possível customizar determinadas opções, obtendo assim um arquivo com uma melhor resolução e produzindo um ganho de confiabilidade na análise.

Para ser gerada a imagem do pMHC com o campo eletrostático, foi utilizado o software UCSF Chimera (Pettersen et al., 2004). Esta ferramenta tem como principal função a visualização de proteínas tridimensionais, como o arquivo PDB por exemplo, e também possui a função de mostrar a superfície da molécula, podendo adicionar o campo eletrostático previamente calculado pelo Delphi. Com isto, gera-se a imagem que é utilizada para a extração dos valores de RGB.

A próxima etapa do processo é a obtenção dos dados de cores, o que é realizado com o *software* ImageJ (Scheider et al., 2012). Através do programa foram selecionadas as 7 regiões indicadas acima, sendo extraídos a média e o desvio padrão dos valores de RGB. Os valores são adicionados a uma tabela juntamente com aqueles da área de acesso ao solvente calculado pelo programa Naccess, sendo esta uma forma indireta de calcular a topologia.

Após a obtenção dos valores de carga e topologia, utilizou-se um pacote do instrumento R para pvclust (Suzuki & Shimodaira, 2006), que adiciona uma análise mais

robusta ao utilizar o bootstrap para validar os agrupamentos mostrados no dendrograma. Com isto, temos um valor de probabilidade agregado aos conjuntos formados.

Para validar nossas análises utilizamos um conjunto de dados já publicados. Em um dos artigos foram selecionados 28 alvos derivados de seis genótipos de hepatite C, sendo que 27 deles são derivados de cinco genótipos e o restante derivado da linhagem selvagem (normal) do vírus (Fytilli et al., 2008). Através então da informação da quantidade de interferon gama liberado pelos linfócitos T CD8⁺, selecionados por gerar resposta contra o epítipo selvagem; se eles interagem com um outro epítipo, pode-se afirmar que estes alvos possuem reatividade cruzada.

Em um outro artigo, foram selecionados oito peptídeos, derivados dos subtipos 1 ao 4 de dengue (DENV-1, DENV-2, DENV-3, DENV-4), sendo dois epítipos de cada subtipo (Duan et al., 2012). Foram eles testados utilizando o mesmo parâmetro de quantificação de interferon gama dos linfócitos T citotóxicos, tendo sido encontrados quatro epítipos que possuem reatividade cruzada e quatro que não a possuem; estes alvos foram utilizados como entrada para a análise de similaridade.

Somando os 2 artigos temos 36 alvos, os quais foram analisados pela nossa técnica para aferir se esta funciona corretamente. No Capítulo II desta Tese encontram-se os dendrogramas que indicam eventos de reatividade cruzada, demonstrando que a nossa metodologia é válida para inferirmos similaridade entre os alvos e para buscarmos alvos imunogênicos através de análises de similaridade (Mendes et al., 2015).

Com os resultados positivos e o artigo sobre a metodologia publicado, o próximo passo era transformar o *legacy* em uma ferramenta de predição de similaridade. Esta ferramenta em princípio estaria disponível online e utilizaria a mesma metodologia empregada no Capítulo II, porém de uma forma automatizada. Os passos efetuados para realizar esta ideia serão apresentados abaixo.

A primeira tentativa foi utilizar uma forma modificada do arquivo gerado pelo Delphi. Uma das opções de saída deste programa é um arquivo em formato de texto plano, contendo valores que seriam equivalentes ao campo eletrostático. Este arquivo possui em torno de 100 mil linhas, em formato de matriz. Utilizamos uma função de distância de matriz para calcular a similaridade, usando assim este valor como entrada para um dendrograma. Infelizmente, apesar de várias tentativas, os resultados preliminares não foram satisfatórios. Uma das

possibilidades que levantamos é que as matrizes de campo eletrostático estivessem com valores de regiões distintas entre os complexos, isto é, uma matriz de uma região em um pMHC não necessariamente seria a mesma região em um outro pMHC, fazendo com que o cálculo de distância entre as matrizes não funcionasse corretamente, pois estavam sendo comparadas regiões diferentes. Para averiguarmos se as matrizes seriam das mesmas regiões entre os pMHCs, tentou-se converter essas matrizes em um mapa de *voxels*, isto é, pontos na tela (*pixels*) com volume, onde cada matriz de carga resultava em pontos de diferentes cores. Porém, o processo para realizarmos esse mapeamento tinha um custo computacional tão elevado que tornou inviável.

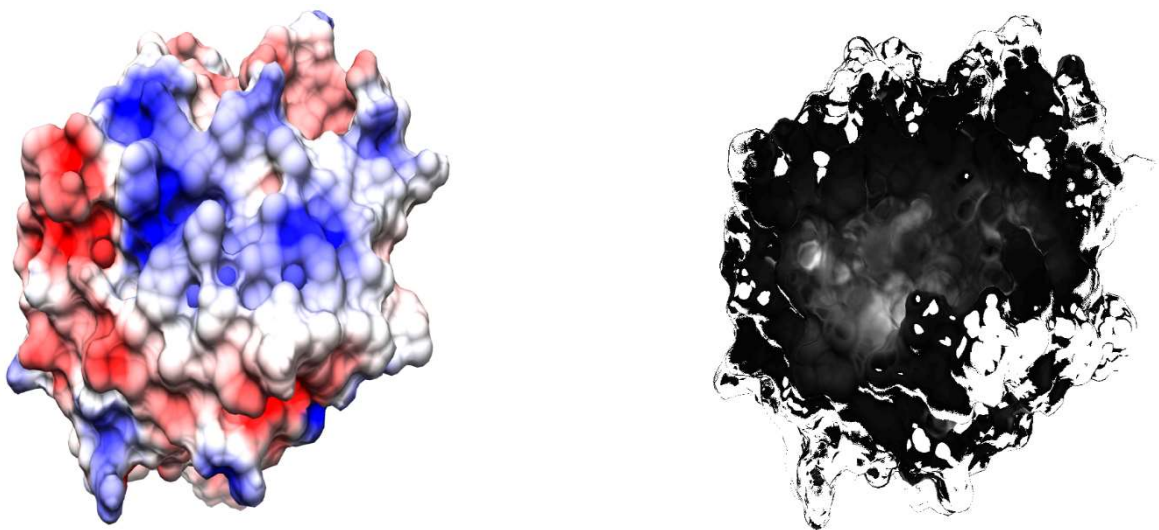


Figura 3: À esquerda temos uma imagem de um pMHC com o seu campo eletrostático calculado na superfície. À direita encontra-se o resultado da sobreposição entre 94 figuras de pMHCs distintas. A parte central, mais acinzentada, equivale às regiões que mais diferem entre os diversos complexos, sendo estes pixels utilizados para a obtenção dos valores de RGB.

Com todos estes contratempos, voltamos então com o plano de trabalhar com as imagens. A princípio, teríamos um script do ImageJ para a obtenção automática dos valores de RGB. No entanto, este script nunca chegou a ser implementado, devido a uma série de fatores que fogem ao escopo desta discussão. Partimos então para uma nova abordagem, desta vez utilizando um algoritmo que sobrepõe todas as imagens do campo eletrostático dos pMHCs e calcula os pixels de maior divergência entre estes, usando os valores RGB destes pixels como entrada para o agrupamento hierárquico. Um exemplo de uma saída deste programa pode ser vista na Figura 3.

Apesar de este algoritmo estar funcionando corretamente, ocorreu um outro problema que inviabilizou a automatização do processo de criação das imagens, impossibilitando assim a geração da ferramenta. Ao criar um script para realizar as etapas de transporte do arquivo PDB, cálculo da superfície, transporte do campo eletrostático e salvamento da imagem, houve um erro no programa UCSF Chimera. Em alguns complexos gerou-se uma deformação na imagem em que alguns pMHCs possuíam um buraco na estrutura, sem nenhuma superfície sendo mostrada. Infelizmente, este é um erro no algoritmo do UCSF Chimera que seus criadores não conseguiram corrigir. Contato com os responsáveis pelo programa, que propuseram algumas soluções, foram inúteis e com isso toda a metodologia por trás da ferramenta teria que ser repensada. Embora estagnado neste ponto o método continuava sendo válido, apesar de não estar automatizado, sendo atualmente utilizado em alguns estudos conduzidos pelo nosso grupo ou por colaboradores.

O Capítulo III apresenta a ferramenta chamada MatchTope. Esta ferramenta foi desenvolvida em colaboração com o laboratório MCM, localizado no *Heidelberg Institute of Theoretical Science*. O MatchTope utiliza como base uma versão modificada do software PIPSA (*Protein Interaction Property Similarity Analysis*) (Wade, Gabdoulline, & De Rienzo, 2001). Esta ferramenta trabalha de uma forma similar à metodologia abordada pelo nosso grupo, utilizando também o cálculo do campo eletrostático, sendo este realizado pelo software UHBD (Madura et al., 1995) para obter a análise de similaridade entre as proteínas. O PIPSA é utilizado geralmente para estimar propriedades de ligações entre proteínas e averiguar parâmetros cinéticos de diversas enzimas. No entanto, em um primeiro momento o PIPSA não funcionou como planejávamos. Como o pMHC possui regiões de maior importância para o reconhecimento do linfócito T CD8⁺, sendo que o campo eletrostático de regiões além da fenda não tem uma relevância fundamental na sinapse imunológica, o PIPSA não conseguiu prever de uma forma precisa as similaridades entre os alvos utilizados no teste, pois utilizou a molécula inteira como entrada para a análise introduzindo assim um alto grau de ruído, pois apenas a região da fenda é significativa. Mesmo editando o PDB do pMHC para apenas a região da fenda ou selecionando apenas os carbonos alfa dos epítopos para utilização como entrada, as análises continuavam a apresentar resultados errôneos.

Devido a estes problemas desistimos de utilizar o PIPSA por um tempo. Porém, com a possibilidade de realizar o Doutorado sanduiche, pude ir pessoalmente ao laboratório de

Heidelberg e com isto tentar aprimorar a ferramenta para funcionamento com a nossa metodologia. Desenvolvemos então uma versão modificada do PIPSA, onde implementou-se um algoritmo que desenha um cilindro na posição da fenda e apenas aquela região é utilizada para o cálculo de similaridade. Com isto conseguimos resultados concordantes com os obtidos anteriormente, podendo assim utilizar este método como base para a nossa ferramenta de similaridade entre pMHCs.

Apesar de ter sido implementado o cilindro para a obtenção dos campos eletrostáticos na fenda, foram realizados diversos testes até chegarmos à utilização deste parâmetro. Foram várias tentativas implementando uma área cônica e uma área esférica dentro da fenda, formatos que existiam previamente, mas que não estavam disponíveis na versão online para a obtenção dos valores de carga. Além disso, foram várias tentativas utilizando diversos parâmetros como variação no raio da esfera, ângulo do cone, raio do cilindro, além de vários tamanhos da sonda que o programa utiliza para mapear o complexo. Diferentes tamanhos de *skin* também foram avaliados pelo *software* para definir a altura da região de interação eletrostática. No Capítulo III há uma figura mostrando a região cilíndrica que utilizamos e uma melhor explicação sobre o funcionamento da ferramenta. A Figura 4 mostra o formato cônico desenhado pelo PIPSA para a obtenção dos valores de carga, sendo este formato diversamente testado sobre vários parâmetros diferentes.

Para a análise estatística foram experimentados vários padrões disponíveis no pacote *cluster*, pertencentes ao R (Ihaka & Gentleman, 1996), para chegarmos a um que conseguisse realizar um agrupamento condizente com os dados de bancada. Existem 35 parametrizações distintas no pacote *cluster*, sendo 7 opções quanto à forma de cálculo da métrica de similaridade, e 5 opções quanto à função utilizada para calcular a ligação entre os diversos elementos. Os resultados variam muito dependendo dos parâmetros escolhidos, como pode ser comparado utilizando o resultado apresentado no Capítulo III com a Figura 5.

No total, se somarmos a quantidade de vezes que foram testados os mais diversos parâmetros, tanto no cálculo do potencial eletrostático quanto no agrupamento hierárquico, chegaremos a um número em torno de 10 mil resultados que foram analisados para calibrarmos a ferramenta em um nível aceitável de precisão. Para uma maior descrição do *workflow* do MatchTope, o Capítulo III apresenta maiores detalhes.

Com a ferramenta pronta abre-se um leque de possibilidades para a sua utilização. Desde a prospecção de novas vacinas até um banco de dados de ligandomas de tumores, a gama de opções é grande e a sua aplicação é simples, rápida e gera bons resultados, auxiliando assim o pesquisador em seu estudo. Segue abaixo uma pequena lista de potenciais funções para o MatchTope:

- **Busca por alvos similares entre diversos genótipos:** Em doenças como a Hepatite C, por exemplo, tem-se um leque de genótipos com muitas diferenças genômicas, o que dificulta a prospecção de uma vacina que abranja a todos, ou pelo menos o maior número de indivíduos daquela espécie. Ao utilizarmos o proteoma de várias amostras de diferentes genótipos é possível inferir, através de um preditor de processamento e apresentação de antígeno, diversos alvos potenciais que podem ser imunogênicos. Em seguida, a partir da modelagem destes alvos utiliza-se ferramentas como o DockTope e o resultado da modelagem como entrada para o MatchTope. Com isto temos como resultado diversos agrupamentos, onde é escolhido um que possua alvos de diversos genótipos distintos, os quais podem ser testados *in vitro* para buscar alvos com respostas de amplo espectro;
- **Busca por alvos dissimilares dentro de uma mesma espécie:** A dengue hemorrágica tem como uma das suas causas a reatividade cruzada. A prospecção de alvos vacinais para esta doença deve seguir a ideia de que, nestes casos, aquele alvo não seja semelhante o suficiente com algum outro dentro dos outros subtipos. Para isto utiliza-se o mesmo raciocínio do tópico acima, porém seleciona-se apenas alvos que não agruparam com nenhum outro, indicando assim uma baixa probabilidade de estimulação de reatividade cruzada de células T;
- **Prospecção de possíveis alvos em vírus recém descobertos:** Uma recente onda de infecção pelo até então não tão conhecido Zika vírus fez com que começasse uma procura rápida por possíveis alvos vacinais para esta doença. Uma abordagem que poderia ser realizada neste caso é a busca por espécies próximas que possuam dados descritos de alvos imunogênicos. Estes alvos então podem ser utilizados como referências imunogênicas para a procura de alvos similares dentro da doença a ser estudada. Por exemplo, ao procurar por alvos potenciais em Zika pode-se utilizar dados de proteínas similares com dados positivos de imunogenicidade na

febre amarela, pois a febre amarela é da mesma família do Zika Vírus. Através da utilização de uma região do genoma da febre amarela reconhecidamente imunogênico por estudos prévios, seriam realizados alinhamentos de sequência considerando-se organismos similares, como o genoma de outro flavivirus. Desta forma obtém-se proteínas alvo mais promissoras comparado a uma abordagem inicial tradicional, que teria que analisar o genoma total de um vírus sem dados prévios descritos de imunogenicidade. A partir deste ponto seguem-se as etapas de modelagem do alvo imunogênico previamente descrito e corroborado na literatura da febre amarela, mais os alvos selecionados após alinhamentos em regiões de interesse definidos como promissores no Zika vírus. Com estes complexos pMHCs disponíveis aplica-se a ferramenta MatchTope que dará como resultado de saída uma métrica de similaridade das proteínas comparadas. Para este exemplo considera-se a imunogenicidade do modelo de referência, permitindo prospectar o mesmo no alvo de estudo (Bragatte et al., 2018, comunicação pessoal);

- **Banco de dados de ligandomas de câncer:** Com o surgimento de novas e promissoras abordagens de imunoterapias é fundamental o desenvolvimento de uma ferramenta “in silico” para auxiliar na pesquisa por novos alvos. Como mencionado anteriormente nesta Tese, a reatividade cruzada de células T entre células tumorais e saudáveis é um assunto importante e há dificuldade em se prever sua ocorrência. Como o DockTope e o MatchTope são ferramentas com eficiência validada através de dados experimentais e que apresentam resultados relativamente rápidos, é possível a análise de dados em larga escala para este tipo de prospecção. Um dos nossos projetos futuros é a criação de um banco de dados de ligandomas de tumores, isto é, um banco de dados com diversos alvos derivados de proteínas tumorais e saudáveis que podem ser expressos e apresentados via MHC-I. O pesquisador terá acesso a informações de quais proteínas tumorais têm semelhança com proteínas de células saudáveis que indiquem uma probabilidade de ocorrência de uma reatividade cruzada. Para isto inicialmente serão selecionados um conjunto de dados de um tumor específico, e serão modelados pMHCs com peptídeos derivados de proteínas

destes tumores e de proteínas de células normais dos tecidos circundantes (serão modelados os peptídeos com altos escores preditos em diversas ferramentas da via de processamento de antígenos endógenos). As sequências das proteínas serão resgatadas através de bancos de dados, como por exemplo o The Human Protein Atlas e o UNIPROT (Uhlen et al., 2015; The UniProt, 2018);

- **Busca por epítomos imunogênicos utilizando dados de RNA-Seq:** Em uma parceria com um grupo de pesquisas do INCA e da Fiocruz-Rio, sob a coordenação do professor Martin Bonamino, desenvolvemos uma estratégia para encontrar que alvos dentro dos tumores estavam sendo reconhecidos e por qual população linfocítica. Para isto utilizamos as sequências de CDR3 prevalentes nos tumores e sequências de proteínas superexpressas no ambiente tumoral. Modelamos os pMHCs de peptídeos contidos nestas proteínas em um alelo de HLA frequente na população e que foi genotipado para os pacientes estudados (HLA-A*02:01). Paralelamente, procuramos cristais contendo o complexo ternário pMHC:TCR e que continham as sequências CDR3 similares às encontradas nos pacientes, supondo que sequências muito similares reconhecem pMHCs estruturalmente semelhantes. As propriedades físico-químicas das regiões de contato com o TCR dos cristais de pMHC foram calculadas e comparadas com as superfícies dos modelos contendo os peptídeos das proteínas superexpressas melhores ranqueados. Através desta técnica fomos capazes de identificar alvos correspondentes nos modelos e nos cristais, o que nos permite inferir o alvo no tumor e a respectiva sequência de CDR3 do linfócito que provavelmente o está reconhecendo. Uma explicação completa pode ser encontrada no artigo que está sendo submetido para a revista *Clinical Cancer Research*. A Figura 6 traz detalhes de como utilizar a ferramenta.

Os resultados das análises geradas pelo MatchTope com alvos derivados do dengue e da hepatite C confirmaram os testes *in vitro* previamente publicados. A diferença crucial em relação às nossas abordagens anteriores é que neste caso desenvolvemos um método de

análise que trabalha diretamente com os dados de potencial eletrostático. Isto nos dá ganho de tempo em processamento e análise e permite a automatização do processo, incluindo o uso concatenado com outras de nossas ferramentas.

Através destas perspectivas da utilização do MatchTope tem-se uma ferramenta pronta e disponível para uso aberto a quaisquer pesquisadores que queiram desenvolver trabalhos envolvendo a compreensão de eventos imunogênicos de estimulação de respostas citotóxicas relacionadas a alvos derivados de vírus, tumores e proteínas próprias (no caso de distúrbios autoimunes).

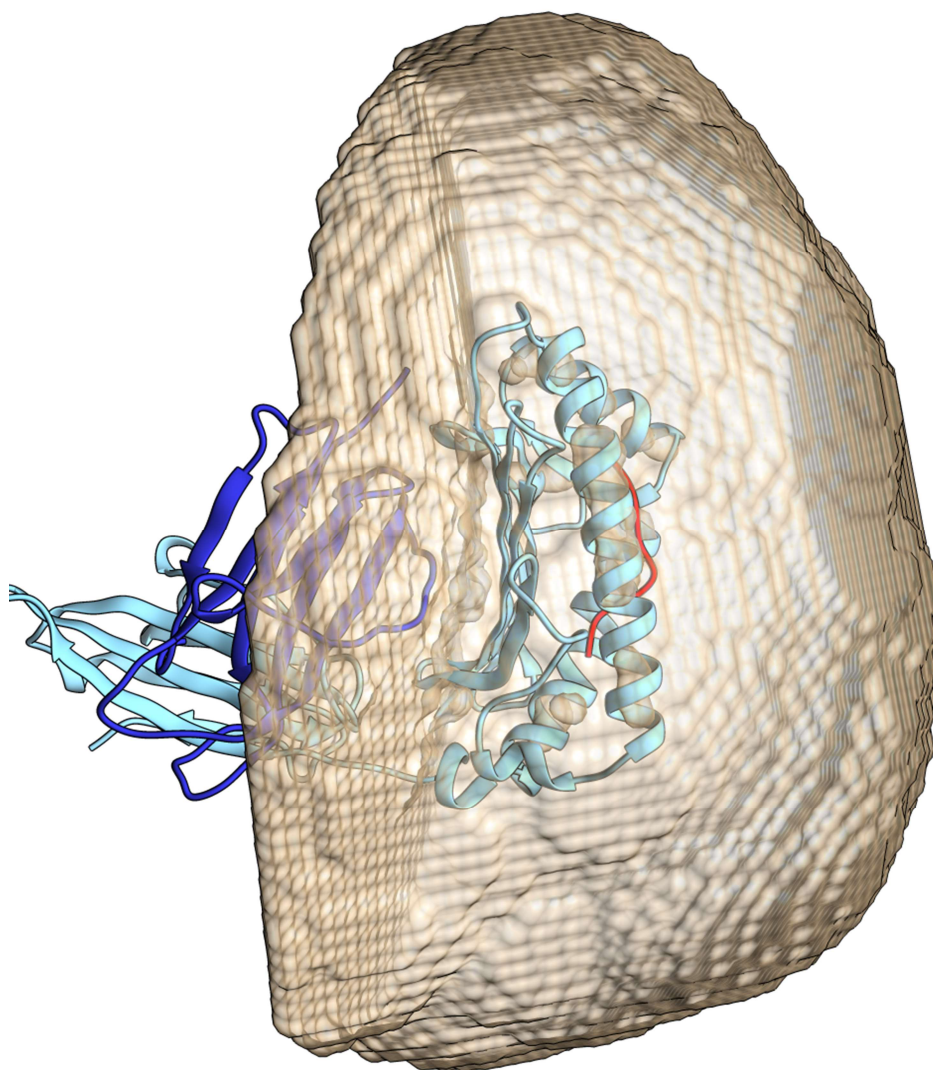


Figura 4: Representação gráfica de um pMHC (as cadeias alfa 1, alfa 2 e alfa 3 estão representadas em azul claro, a beta-2 microglobulina em azul escuro e o epítipo está indicado em vermelho) com a região onde é desenhado o cone (superfície semi transparente pintado em bege), sendo deste local extraídos os valores eletrostáticos, na cor cinza.

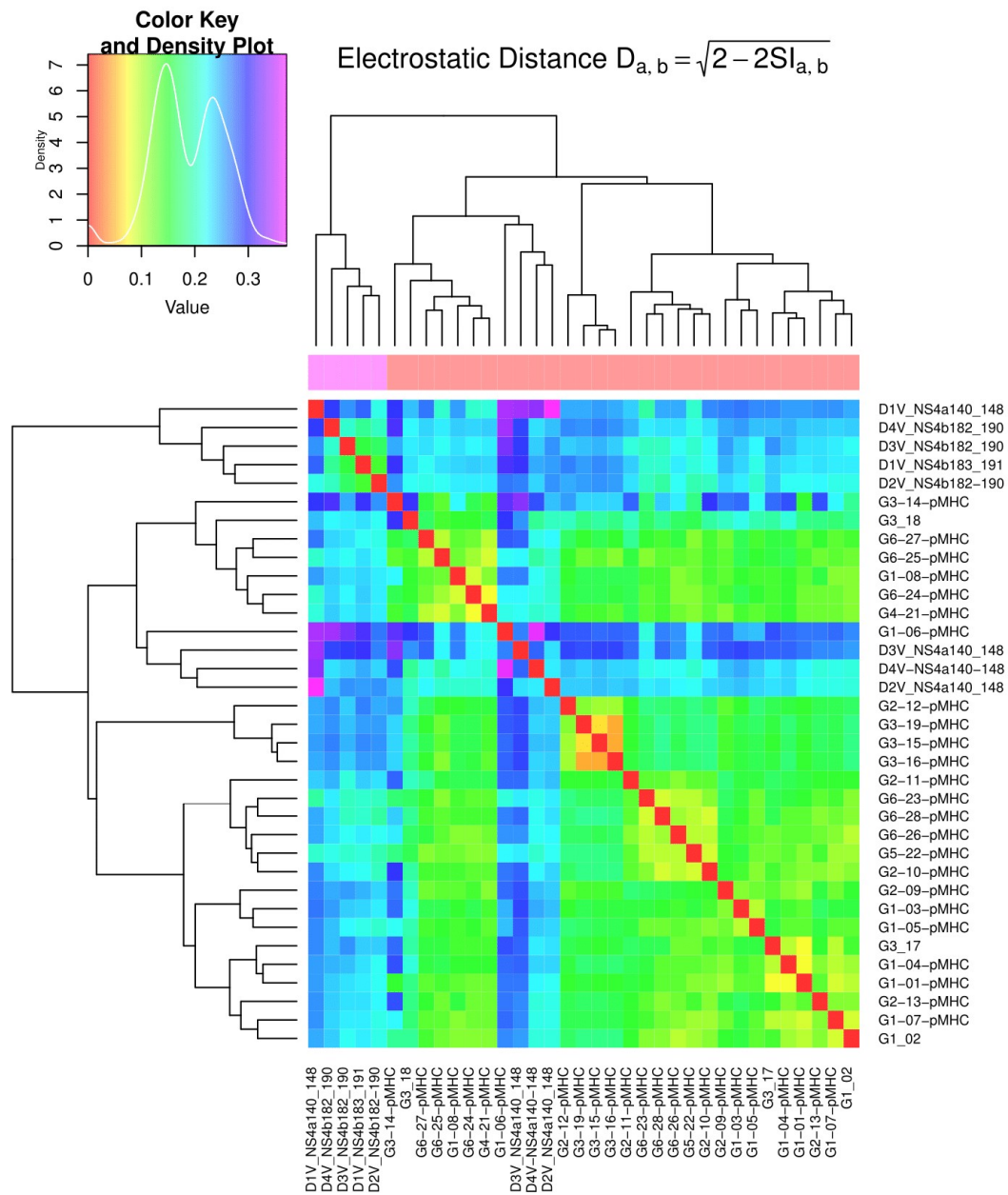


Figura 5: Dendrograma e mapa de calor, utilizando os mesmos parâmetros para o cálculo da região eletrostática utilizado no Capítulo III, porém empregando outros parâmetros para a operação estatística. Comparando-se com a Figura 3 do Capítulo III percebe-se uma grande diferença, mostrando a importância da utilização de um método estatístico adequado.

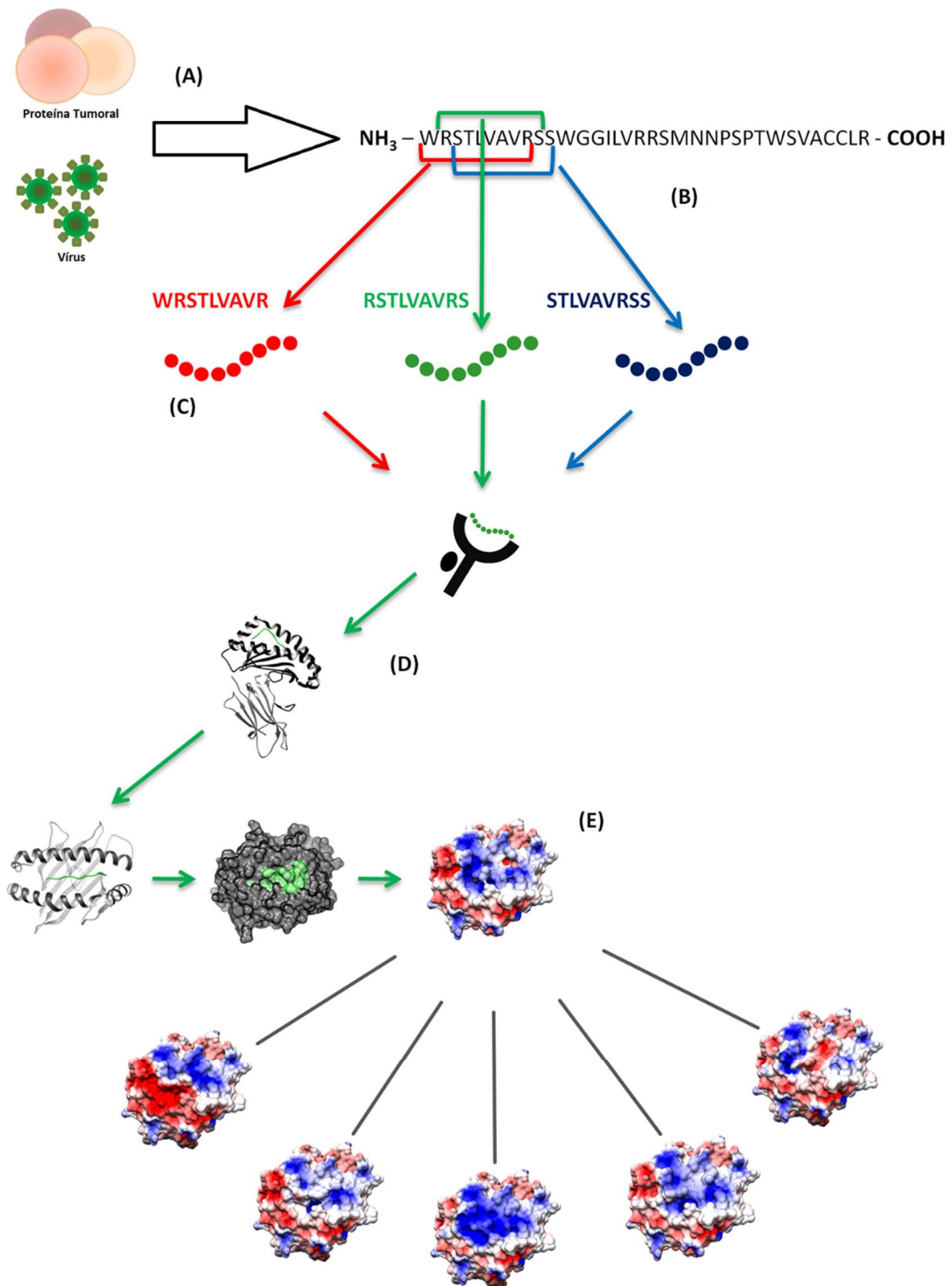


Figura 6: Uma esquematização da utilização do MatchTope em um estudo específico. A) Seleção da proteína de estudo, podendo ser ela viral, tumoral ou derivada de célula sadia. B) Obtenção do proteoma desta proteína. C) Utilizando o proteoma como entrada, emprega-se o uso de preditores de processamento e apresentação de antígeno, para identificar possíveis alvos que possam ser apresentados. D) Modelagem destes alvos utilizando alguma ferramenta específica, como por exemplo o DockTope, gerando assim a estrutura do pMHC. E) Utilizando os vários pMHCs gerados como entrada, calcula-se a similaridade destes usando o MatchTope, obtendo como resultado um dendrograma mostrando quais alvos possuem similaridade a ponto de desencadear uma reatividade cruzada. Imagem adaptada de Vieira et al. (submetido).

Referências complementares (Capítulos I e IV)

- Alder, M. N., Rogozin, I. B., Iyer, L. M., Glazko, G. V., Cooper, M. D., & Pancer, Z. (2005). Diversity and function of adaptive immune receptors in a jawless vertebrate. *Science*, *310*(5756), 1970-1973. doi: 10.1126/science.1119420.
- Alsaab, H. O., Sau, S., Alzhrani, R., Tatiparti, K., Bhise, K., Kashaw, S. K., & Iyer, A. K. (2017). PD-1 and PD-L1 checkpoint signaling inhibition for cancer immunotherapy: mechanism, combinations, and clinical outcome. *Front Pharmacol*, *8*, 561. doi: 10.3389/fphar.2017.00561.
- Antoniou, A. N., Powis, S. J., & Elliott, T. (2003). Assembly and export of MHC class I peptide ligands. *Curr Opin Immunol*, *15*(1), 75-81. doi: 10.1016/s0952-7915(02)00010-9
- Antunes, D. A., Rigo, M. M., Freitas, M. V., Mendes, M. F. A., Sinigaglia, M., Lizee, G., . . . Vieira, G. F. (2017). Interpreting T-Cell cross-reactivity through structure: implications for TCR-based cancer immunotherapy. *Front Immunol*, *8*, 1210. doi: 10.3389/fimmu.2017.01210.
- Antunes, D. A., Vieira, G. F., Rigo, M. M., Cibulski, S. P., Sinigaglia, M., & Chies, J. A. (2010). Structural allele-specific patterns adopted by epitopes in the MHC-I cleft and reconstruction of MHC:peptide complexes to cross-reactivity assessment. *PLoS One*, *5*(4), e10353. doi: 10.1371/journal.pone.0010353.
- Attaf, M., Legut, M., Cole, D. K., & Sewell, A. K. (2015). The T cell antigen receptor: the Swiss army knife of the immune system. *Clin Exp Immunol*, *181*(1), 1-18. doi: 10.1111/cei.12622.
- Ausubel, F. M. (2005). Are innate immune signaling pathways in plants and animals conserved? *Nat Immunol*, *6*(10), 973-979. doi: 10.1038/ni1253.
- Beck, G., & Habicht, G. S. (1996). Immunity and the invertebrates. *Scient Am*, *275*(5), 60-66. doi: 10.1038/scientificamerican1196-60.
- Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., . . . Bourne, P. E. (2000). The Protein Data Bank. *Nucl Ac Res*, *28*(1), 235-242.
- Blum, J. S., Wearsch, P. A., & Cresswell, P. (2013). Pathways of antigen processing. *Annu Rev Immunol*, *31*, 443-473. doi: 10.1146/annurev-immunol-032712-095910.
- Boehm, T., Hirano, M., Holland, S. J., Das, S., Schorpp, M., & Cooper, M. D. (2018). Evolution of alternative adaptive immune systems in vertebrates. *Annu Rev Immunol*, *36*, 19-42. doi: 10.1146/annurev-immunol-042617-053028.
- Boyton, R. J. (2002). Pulmonary defences to acute respiratory infection. *Brit Med Bull*, *61*(1), 1-12. doi: 10.1093/bmb/61.1.1.
- Duan, Z. L., Li, Q., Wang, Z. B., Xia, K. D., Guo, J. L., Liu, W. Q., & Wen, J. S. (2012). HLA-A*0201-restricted CD8+ T-cell epitopes identified in dengue viruses. *Virology*, *9*, 259. doi: 10.1186/1743-422X-9-259.
- Ferrero-Miliani, L., Nielsen, O. H., Andersen, P. S., & Girardin, S. E. (2007). Chronic inflammation: importance of NOD2 and NALP3 in interleukin-1beta generation. *Clin Exp Immunol*, *147*(2), 227-235. doi: 10.1111/j.1365-2249.2006.03261.x.
- Flajnik, M. F., & Kasahara, M. (2010). Origin and evolution of the adaptive immune system: genetic events and selective pressures. *Nat Rev Genet*, *11*(1), 47-59. doi: 10.1038/nrg2703.
- Fyttili, P., Dalekos, G. N., Schlaphoff, V., Suneetha, P. V., Sarrazin, C., Zauner, W., . . . Wedemeyer, H. (2008). Cross-genotype-reactivity of the immunodominant HCV CD8

- T-cell epitope NS3-1073. *Vaccine*, 26(31), 3818-3826. doi: 10.1016/j.vaccine.2008.05.045.
- Glickman, M. H., & Ciechanover, A. (2002). The ubiquitin-proteasome proteolytic pathway: destruction for the sake of construction. *Physiol Rev*, 82(2), 373-428. doi: 10.1152/physrev.00027.2001.
- Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell*, 144(5), 646-674. doi: 10.1016/j.cell.2011.02.013.
- Ihaka, R., & Gentleman, R. (1996). R: A language for data analysis and graphics. *J Comput Graph Stat*, 5(3), 299. doi: 10.2307/1390807.
- Khan, J. M., & Ranganathan, S. (2010). pDOCK: a new technique for rapid and accurate docking of peptide ligands to Major Histocompatibility Complexes. *Immunome Res*, 6 Suppl 1, S2. doi: 10.1186/1745-7580-6-S1-S2.
- Kitchen, D. B., Decornez, H., Furr, J. R., & Bajorath, J. (2004). Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat Rev Drug Discov*, 3(11), 935-949. doi: 10.1038/nrd1549.
- Kulski, J. K., Shiina, T., Anzai, T., Kohara, S., & Inoko, H. (2002). Comparative genomic analysis of the MHC: the evolution of class I duplication blocks, diversity and complexity from shark to man. *Immunol Rev*, 190(1), 95-122. doi: 10.1034/j.1600-065X.2002.19008.x.
- Lankat-Buttgereit, B., & Tampe, R. (2002). The transporter associated with antigen processing: function and implications in human diseases. *Physiol Rev*, 82(1), 187-204. doi: 10.1152/physrev.00025.2001.
- LeBien, T. W., & Tedder, T. F. (2008). B lymphocytes: how they develop and function. *Blood*, 112(5), 1570-1580. doi: 10.1182/blood-2008-02-078071.
- Ledford, H. (2017). Engineered cell therapy for cancer gets thumbs up from FDA advisers. *Nature*, 547(7663), 270. doi: 10.1038/nature.2017.22304.
- Lefranc, M. P. (2014). Immunoglobulin and T Cell receptor genes: IMGT((R)) and the birth and rise of immunoinformatics. *Front Immunol*, 5, 22. doi: 10.3389/fimmu.2014.00022.
- Li, L., Li, C., Sarkar, S., Zhang, J., Witham, S., Zhang, Z., . . . Alexov, E. (2012). DelPhi: a comprehensive suite for DelPhi software and associated resources. *BMC Biophys*, 5, 9. doi: 10.1186/2046-1682-5-9.
- Li, Y., Yin, Y., & Mariuzza, R. A. (2013). Structural and biophysical insights into the role of CD4 and CD8 in T cell activation. *Front Immunol*, 4, 206. doi: 10.3389/fimmu.2013.00206.
- Linette, G. P., Stadtmauer, E. A., Maus, M. V., Rapoport, A. P., Levine, B. L., Emery, L., . . . June, C. H. (2013). Cardiovascular toxicity and titin cross-reactivity of affinity-enhanced T cells in myeloma and melanoma. *Blood*, 122(6), 863-871. doi: 10.1182/blood-2013-03-490565.
- Lundegaard, C., Lamberth, K., Harndahl, M., Buus, S., Lund, O., & Nielsen, M. (2008). NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8-11. *Nucl Ac Res*, 36(Web Server issue), W509-512. doi: 10.1093/nar/gkn202.
- Madura, J. D., Briggs, J. M., Wade, R. C., Davis, M. E., Luty, B. A., Ilin, A., . . . McCammon, J. A. (1995). Electrostatics and diffusion of molecules in solution: simulations with the University of Houston Brownian Dynamics program. *Comp Phys Commun*, 91(1-3), 57-95. doi: 10.1016/0010-4655(95)00043-f.

- Malone, R. W., Homan, J., Callahan, M. V., Glasspool-Malone, J., Damodaran, L., Schneider Ade, B., . . . Zika Response Working, G. (2016). Zika virus: medical countermeasure development challenges. *PLoS Negl Trop Dis*, *10*(3), e0004530. doi: 10.1371/journal.pntd.0004530.
- Medzhitov, R. (2007). Recognition of microorganisms and activation of the immune response. *Nature*, *449*(7164), 819-826. doi: 10.1038/nature06246.
- Mendes, M. F., Antunes, D. A., Rigo, M. M., Sinigaglia, M., & Vieira, G. F. (2015). Improved structural method for T-cell cross-reactivity prediction. *Mol Immunol*, *67*(2 Pt B), 303-310. doi: 10.1016/j.molimm.2015.06.017.
- Mills, C. (2012). M1 and M2 macrophages: oracles of health and disease. *Crit Rev Immunol*, *32*(6), 463-488. doi: 10.1615/CritRevImmunol.v32.i6.10.
- Milstein, O., Hagin, D., Lask, A., Reich-Zeliger, S., Shezen, E., Ophir, E., . . . Reisner, Y. (2011). CTLs respond with activation and granule secretion when serving as targets for T-cell recognition. *Blood*, *117*(3), 1042-1052. doi: 10.1182/blood-2010-05-283770.
- Moise, L., Gutierrez, A., Kibria, F., Martin, R., Tassone, R., Liu, R., . . . De Groot, A. S. (2015). iVAX: an integrated toolkit for the selection and optimization of antigens and the design of epitope-driven vaccines. *Hum Vaccin Immunother*, *11*(9), 2312-2321. doi: 10.1080/21645515.2015.1061159.
- Morgan, R. A., Chinnasamy, N., Abate-Daga, D., Gros, A., Robbins, P. F., Zheng, Z., . . . Rosenberg, S. A. (2013). Cancer regression and neurological toxicity following anti-MAGE-A3 TCR gene therapy. *J Immunother*, *36*(2), 133-151. doi: 10.1097/CJI.0b013e3182829903.
- Mukhopadhyay, D., & Riezman, H. (2007). Proteasome-independent functions of ubiquitin in endocytosis and signaling. *Science*, *315*(5809), 201-205. doi: 10.1126/science.1127085.
- Murata, S., Sasaki, K., Kishimoto, T., Niwa, S., Hayashi, H., Takahama, Y., & Tanaka, K. (2007). Regulation of CD8+ T cell development by thymus-specific proteasomes. *Science*, *316*(5829), 1349-1353. doi: 10.1126/science.1141915.
- Nassif, N. D., Cambray, S. E., & Kraut, D. A. (2014). Slipping up: partial substrate degradation by ATP-dependent proteases. *IUBMB Life*, *66*(5), 309-317. doi: 10.1002/iub.1271.
- Nonaka, M. (2014). Evolution of the complement system. *Subcell Biochem*, *80*, 31-43. doi: 10.1007/978-94-017-8881-6_3.
- Normile, D. (2013). Tropical medicine. Surprising new dengue virus throws a spanner in disease control efforts. *Science*, *342*(6157), 415. doi: 10.1126/science.342.6157.415.
- Ohno, O., Mizokami, M., Wu, R. R., Saleh, M. G., Ohba, K., Orito, E., . . . Lau, J. Y. (1997). New hepatitis C virus (HCV) genotyping system that allows for identification of HCV genotypes 1a, 1b, 2a, 2b, 3a, 3b, 4, 5a, and 6a. *Journal of Clinical Microbiology*, *35*(1), 201-207.
- Palumbo, M. O., Kavan, P., Miller, W. H., Jr., Panasci, L., Assouline, S., Johnson, N., . . . Batist, G. (2013). Systemic cancer therapy: achievements and challenges that lie ahead. *Front Pharmacol*, *4*, 57. doi: 10.3389/fphar.2013.00057.
- Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., & Ferrin, T. E. (2004). UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem*, *25*(13), 1605-1612. doi: 10.1002/jcc.20084.
- Raman, M. C., Rizkallah, P. J., Simmons, R., Donnellan, Z., Dukes, J., Bossi, G., . . . Jakobsen, B. K. (2016). Direct molecular mimicry enables off-target cardiovascular toxicity by an


- enhanced affinity TCR designed for cancer immunotherapy. *Sci Rep*, 6, 18851. doi: 10.1038/srep18851.
- Rammensee, H.-G., Bachmann, J., Emmerich, N. P. N., Bachor, O. A., & Stevanović, S. (1999). SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics*, 50(3), 213-219. doi: 10.1007/s002510050595.
- Ranjit, S., & Kissoon, N. (2011). Dengue hemorrhagic fever and shock syndromes. *Pediatr Crit Care Med*, 12(1), 90-100. doi: 10.1097/PCC.0b013e3181e911a7.
- Rappuoli, R., Bottomley, M. J., D'Oro, U., Finco, O., & De Gregorio, E. (2016). Reverse vaccinology 2.0: Human immunology instructs vaccine antigen design. *J Exp Med*, 213(4), 469-481. doi: 10.1084/jem.20151960.
- Regner, M. (2001). Cross-reactivity in T-cell antigen recognition. *Immunol Cell Biol*, 79(2), 91-100. doi: 10.1046/j.1440-1711.2001.00994.x.
- Rigo, M. M., Antunes, D. A., Vaz de Freitas, M., Fabiano de Almeida Mendes, M., Meira, L., Sinigaglia, M., & Vieira, G. F. (2015). DockTope: a Web-based tool for automated pMHC-I modelling. *Sci Rep*, 5, 18413. doi: 10.1038/srep18413.
- Rodenhuis-Zybert, I. A., Wilschut, J., & Smit, J. M. (2010). Dengue virus life cycle: viral and host factors modulating infectivity. *Cell Mol Life Sci*, 67(16), 2773-2786. doi: 10.1007/s00018-010-0357-z.
- Schneider, C. A., Rasband, W. S., & Eliceiri, K. W. (2012). NIH Image to ImageJ: 25 years of Image Analysis. *Nature Meth*, 9(7), 671-675.
- Schwarz, B. A., & Bhandoola, A. (2006). Trafficking from the bone marrow to the thymus: a prerequisite for thymopoiesis. *Immunol Rev*, 209, 47-57. doi: 10.1111/j.0105-2896.2006.00350.x.
- Sidney, J., Peters, B., Frahm, N., Brander, C., & Sette, A. (2008). HLA class I supertypes: a revised and updated classification. *BMC Immunol*, 9, 1. doi: 10.1186/1471-2172-9-1.
- Sikka, V., Chattu, V. K., Popli, R. K., Galwankar, S. C., Kelkar, D., Sawicki, S. G., . . . Papadimos, T. J. (2016). The emergence of Zika virus as a global health security threat: a review and a consensus statement of the INDUSEM joint working group (JWG). *J Glob Infect Dis*, 8(1), 3-15. doi: 10.4103/0974-777X.176140.
- Simmonds, P., Holmes, E. C., Cha, T. A., Chan, S. W., McOmish, F., Irvine, B., . . . Urdea, M. S. (1993). Classification of hepatitis C virus into six major genotypes and a series of subtypes by phylogenetic analysis of the NS-5 region. *J Gen Virol*, 74 (Pt 11), 2391-2399. doi: 10.1099/0022-1317-74-11-2391.
- Sinigaglia, M., Antunes, D. A., Rigo, M. M., Chies, J. A., & Vieira, G. F. (2013). CrossTope: a curate repository of 3D structures of immunogenic peptide: MHC complexes. *Database (Oxford)*, 2013, bat002. doi: 10.1093/database/bat002.
- Smyth, M. J., Hayakawa, Y., Takeda, K., & Yagita, H. (2002). New aspects of natural-killer-cell surveillance and therapy of cancer. *Nat Rev Cancer*, 2(11), 850-861. doi: 10.1038/nrc928.
- Starr, T. K., Jameson, S. C., & Hogquist, K. A. (2003). Positive and negative selection of T cells. *Annu Rev Immunol*, 21, 139-176. doi: 10.1146/annurev.immunol.21.120601.141107.
- Stram, Y., & Kuzntzova, L. (2006). Inhibition of viruses by RNA interference. *Virus Genes*, 32(3), 299-306. doi: 10.1007/s11262-005-6914-0.
- Suzuki, R., & Shimodaira, H. (2006). Pvcust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, 22(12), 1540-1542. doi: 10.1093/bioinformatics/btl117.

- Tejada-Simon, M. V., Zang, Y. C., Hong, J., Rivera, V. M., & Zhang, J. Z. (2003). Cross-reactivity with myelin basic protein and human herpesvirus-6 in multiple sclerosis. *Ann Neurol*, *53*(2), 189-197. doi: 10.1002/ana.10425.
- Todman, S. J., Halling-Brown, M. D., Davies, M. N., Flower, D. R., Kayikci, M., & Moss, D. S. (2008). Toward the atomistic simulation of T cell epitopes automated construction of MHC: peptide structures for free energy calculations. *J Mol Graph Model*, *26*(6), 957-961. doi: 10.1016/j.jmglm.2007.07.005.
- Tomar, N., & De, R. K. (2010). Immunoinformatics: an integrated scenario. *Immunology*, *131*(2), 153-168. doi: 10.1111/j.1365-2567.2010.03330.x.
- Uhlen, M., Fagerberg, L., Hallstrom, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., . . . Ponten, F. (2015). Proteomics. Tissue-based map of the human proteome. *Science*, *347*(6220), 1260419. doi: 10.1126/science.1260419
- Vandiedonck, C., & Knight, J. C. (2009). The human Major Histocompatibility Complex as a paradigm in genomics research. *Brief Funct Genom Proteom*, *8*(5), 379-394. doi: 10.1093/bfpg/elp010.
- Vantourout, P., & Hayday, A. (2013). Six-of-the-best: unique contributions of gammadelta T cells to immunology. *Nat Rev Immunol*, *13*(2), 88-100. doi: 10.1038/nri3384.
- Vigneron, N., Ferrari, V., Stroobant, V., Abi Habib, J., & Van den Eynde, B. J. (2017). Peptide splicing by the proteasome. *J Biol Chem*, *292*(51), 21170-21179. doi: 10.1074/jbc.R117.807560.
- Vita, R., Overton, J. A., Greenbaum, J. A., Ponomarenko, J., Clark, J. D., Cantrell, J. R., . . . Peters, B. (2015). The immune epitope database (IEDB) 3.0. *Nucleic Acids Res*, *43*(Database issue), D405-412. doi: 10.1093/nar/gku938.
- Vivier, E., Raulet, D. H., Moretta, A., Caligiuri, M. A., Zitvogel, L., Lanier, L. L., . . . Ugolini, S. (2011). Innate or adaptive immunity? The example of natural killer cells. *Science*, *331*(6013), 44-49. doi: 10.1126/science.1198687.
- Wade, R. C., Gabdoulline, R. R., & De Rienzo, F. (2001). Protein interaction property similarity analysis. *Intern J Quant Chem*, *83*(3-4), 122-127. doi: 10.1002/qua.1204.
- Yuseff, M. I., Pierobon, P., Reversat, A., & Lennon-Dumenil, A. M. (2013). How B cells capture, process and present antigens: a crucial role for cell polarity. *Nat Rev Immunol*, *13*(7), 475-486. doi: 10.1038/nri3469.
- Zhang, S., Bakshi, R. K., Suneetha, P. V., Fytilli, P., Antunes, D. A., Vieira, G. F., . . . Cornberg, M. (2015). Frequency, private specificity, and cross-reactivity of preexisting hepatitis C virus (HCV)-specific CD8+ T Cells in HCV-seronegative individuals: implications for vaccine responses. *J Virol*, *89*(16), 8304-8317. doi: 10.1128/JVI.00539-15.
- Zhang, Z. H., Koh, J. L., Zhang, G. L., Choo, K. H., Tammi, M. T., & Tong, J. C. (2007). AllerTool: a web server for predicting allergenicity and allergic cross-reactivity in proteins. *Bioinformatics*, *23*(4), 504-506. doi: 10.1093/bioinformatics/btl621.

Anexos

Durante o Doutorado sempre colaborei com outros trabalhos na área acadêmico-científica, estando alguns já publicados e outros em fase de redação de manuscritos ou de submissão. Devido à importância destes para a minha formação, estou incluindo-os na minha Tese como Anexo. O primeiro é o de maior importância para o meu Doutorado, pois foi graças ao DockTope que surgiu o MatchTope. O segundo está situado em uma vertente mais teórica da reatividade cruzada, porém utilizando a nossa metodologia para a discussão do problema. Os outros trabalhos, apesar de serem um pouco menos impactantes para esta Tese, tiveram fundamental importância para a expansão dos meus conhecimentos, algo desejável para um aluno de Doutorado.

SCIENTIFIC REPORTS



OPEN

DockTope: a Web-based tool for automated pMHC-I modelling

Maurício Menegatti Rigo^{1,*}, Dinler Amaral Antunes^{1,2,*}, Martiela Vaz de Freitas¹, Marcus Fabiano de Almeida Mendes¹, Lindolfo Meira³, Marialva Sinigaglia¹ & Gustavo Fioravanti Vieira¹

Received: 28 June 2015

Accepted: 18 November 2015

Published: 17 December 2015

The immune system is constantly challenged, being required to protect the organism against a wide variety of infectious pathogens and, at the same time, to avoid autoimmune disorders. One of the most important molecules involved in these events is the Major Histocompatibility Complex class I (MHC-I), responsible for binding and presenting small peptides from the intracellular environment to CD8⁺ T cells. The study of peptide:MHC-I (pMHC-I) molecules at a structural level is crucial to understand the molecular mechanisms underlying immunologic responses. Unfortunately, there are few pMHC-I structures in the Protein Data Bank (PDB) (especially considering the total number of complexes that could be formed combining different peptides), and pMHC-I modelling tools are scarce. Here, we present DockTope, a free and reliable web-based tool for pMHC-I modelling, based on crystal structures from the PDB. DockTope is fully automated and allows any researcher to construct a pMHC-I complex in an efficient way. We have reproduced a dataset of 135 non-redundant pMHC-I structures from the PDB (C α RMSD below 1 Å). Modelling of pMHC-I complexes is remarkably important, contributing to the knowledge of important events such as cross-reactivity, autoimmunity, cancer therapy, transplantation and rational vaccine design.

The immune system is mainly responsible for defending the organism against a wide range of infectious pathogens, such as viruses, bacteria and fungi. At the same time, it should be able to preserve the organism, avoiding autoimmunity events, for example. This complex system is orchestrated by a set of cells and molecules involved in clearing infections and maintaining a healthy organism. One of these molecules with pivotal importance is the Major Histocompatibility Complex class I (MHC-I), which is typically capable to bind short peptides with eight to twelve amino acids in length (also called epitopes in this context). The peptide:MHC-I (pMHC-I) complex is transported through a specific endogenous pathway to the cell surface, where it can be inspected by a T Cell Receptor (TCR) of a CD8⁺ lymphocyte¹. Based on complementary structural patterns, the pMHC-I and TCR interaction can trigger an immunologic response, which will mainly depend on the peptide source²⁻⁴.

The MHC-I molecule is composed of an α domain (subdivided in $\alpha 1$, $\alpha 2$ and $\alpha 3$ regions), encoded by one of the most polymorphic regions of the genome, referred to as MHC locus, located on chromosome 6 (in humans) and chromosome 17 (in murines)^{5,6}. Additionally, a $\beta 2$ -microglobulin interacts with the MHC-I α domain providing complex stability⁷. The protein's polymorphism occur mainly in the $\alpha 1$ and $\alpha 2$ regions, which form a cleft where the peptide is bound and presented to the TCR. Each MHC-I allele encodes a specific protein, called allotype. The MHC-I allotypes are highly variable in terms of amino acids composition and, depending on the organism studied, a specific name is assigned, such as Human Leukocyte Antigen (HLA), in humans, and H-2 antigen (H-2), in murines.

The MHC-I allotype variability allows a broad array of peptides to bind inside the MHC-I cleft through a specific interaction pattern between the epitope and the MHC-I residues. Thus, the understanding of the epitope binding mode and the structural features of this protein complex is of pivotal importance to unveil the molecular basis underlying important immune responses. Unfortunately, the low number of pMHC-I structures experimentally resolved and the lack of accessible and reliable structural *in silico* modelling approaches are hindering for the evolution of this field. Currently, three-dimensional structures of pMHC-I complexes are determined through specific techniques, such as X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy, which

¹Núcleo de Bioinformática do Laboratório de Imunogenética, Departamento de Genética, Universidade Federal do Rio Grande do Sul, Postcode 91501-970, Brazil. ²Department of Computer Science, Rice University, Houston, Texas, 77005, USA. ³CESUP-Centro Nacional de Supercomputação, Universidade Federal do Rio Grande do Sul, Postcode 90035-190, Brazil. *These authors contributed equally to this work. Correspondence and requests for materials should be addressed to M.M.R. (email: mauriciomr1985@gmail.com) or G.F.V. (email: gusfioravanti@yahoo.com.br)

are costly and time-consuming. To overcome this, pMHC-I modelling represents an interesting and creative solution. Despite the availability of homology modelling techniques, each allotype presents specific particularities that cannot be simply determined through regular approaches. Thus, molecular modelling requires a careful structural study based on a solid validation process, which should take the wide MHC-I allotype variability in consideration. Immunoinformatics programs devoted to pMHC-I modelling have been developed or are under development^{8–17}, but there is currently few online programs available to the scientific community.

Here, we present DockTope, a fully automated web-server tool designed with the purpose of modelling pMHC-I complexes for two human (HLA-A*02:01 and HLA-B*27:05) and two murine (H-2-Db and H-2-Kb) MHC-I allotypes. We have validated this tool through the cross-docking reproduction of 135 non-redundant structurally resolved pMHC-I structures available in the Protein Data Bank (PDB), using C α and all atom Root Mean Square Deviation (RMSD) values for evaluation. DockTope has been fully automated using different programming languages, such as python and shell scripting. In addition, we have designed a dedicated web server providing free and easy access to any user throughout the world.

Results

The rationale behind DockTope. The DockTope tool is based on the D1-EM-D2 approach¹⁸, a pMHC-I modelling technique published by our group in 2010. The D1-EM-D2 approach is based on a protocol that employs a molecular docking step (D1), followed by an energy minimization (EM) of the pMHC-I complex and a final molecular docking round (D2). To develop DockTope, the D1-EM-D2 approach has gone through several implementations, including a new data validation improving its accuracy and reliability, and the full automation of the process. Here, we briefly describe important highlights of the technique improvement. A flowchart of the whole process is provided in Fig. 1.

Before the D1 step, the user provides the linear epitope sequence to be modelled, which is transformed into a three-dimensional structure. This is made possible by the fact that each epitope carries a specific backbone structure, depending on the MHC-I allotype where it is presented¹⁸. The epitope to be modelled is superimposed on another epitope (named here the 'Epitope pattern'), which was already determined by X-ray crystallography in the context of the target MHC-I (Table 1). Since the three-dimensional epitope structure is obtained by modelling the side chains over a constrained backbone, a brief energy minimization step is performed, allowing for a mild global relaxation of the peptide.

To perform the D1 step, all the MHC-I residues and the epitope backbone are kept rigid. Only the epitope side chains are allowed to move. During the molecular docking, the epitope can also perform rotational and translational movements, and the program AutoDock Vina¹⁹ is used to search for the best epitope conformations inside the MHC-I cleft region. One round of molecular docking provides the best epitope conformations based on a scoring function, returning a Binding Energy (BE) value in kcal/mol, which is used by the program to rank the best conformations. To improve chances to find a suitable conformation, our approach performs twenty independent docking runs, using different initial points, which ensures that the program will search a wider range of conformations. In the end, the best conformation of each docking run is retrieved, producing a total of twenty structures.

Before the pMHC-I EM, the best epitope conformation is chosen among the twenty structures generated by AutoDock Vina. This choice is based on two variables: the BE and the average RMSD of each conformation in relation to all other structures. Since the epitope conformation having the best interaction with the MHC-I presents a low BE (a basic relationship between entropy and free energy calculations of non-covalent binding²⁰), we have designed a specific shell script to calculate the BE average among the twenty epitope structures generated from D1. This way we obtain a cut-off (Co) value that is used as a first structure filter. This will exclude spurious conformations that can arise and that do not represent the binding mode of the epitope with its respective MHC-I. Additionally, the remaining epitope conformations are compared with each other, using the *g_confrms* program from the GROMACS package²¹, which returns a RMSD value for each conformation pair. The epitope conformation with the lowest RMSD mean among all outputted conformations (i.e. the average structure) is chosen as the best structure. We describe the equation for choosing the best conformation in the 'Material and Methods section'. After that, the pMHC-I complex is submitted to an EM protocol, where all the residues are kept flexible to accommodate and correct the interactions between epitope and MHC-I.

The second and final molecular docking round, D2, is performed in the same way as the D1. The D2 step is intended to refine the pMHC-I structure, which is possible because the EM has already accommodated the MHC-I side chains to the target epitope structure. The final structure with the best epitope conformation is also chosen as explained above.

Automation and validation of DockTope. Back in 2010, the D1-EM-D2 approach validation process was performed over 46 pMHC-I structures available in the PDB, including the HLA-A*02:01, H-2-Db and H-2-Kb allotypes¹⁸. Here, thanks to the DockTope automated process, a broader validation analysis is reported over 135 pMHC-I structures, encompassing the previous dataset and including the HLA-B*27:05 allotype (Table 2). This represents almost three times the number of structures previously analysed. Also, the automation process first presented here, is a crucial feature of the DockTope web-based tool; it was made possible by writing and concatenating of more than 20 shell and python scripts. After submitting a sequence, the program automatically generates the three-dimensional epitope structure, according to the MHC-I of interest, and performs the molecular docking and energy minimization steps.

Since our method uses a reference MHC-I structure to build every model of a given allotype (also referenced here as 'MHC donor'), the validation process occurred through a cross-docking scheme. Each pMHC-I structure was modelled using only the epitope linear sequence as input and, in the end, the generated pMHC-I complex (model) was compared to its respective crystal structure deposited in the PDB (target). The comparison was based on RMSD values for the epitope atoms (considering C α or all atoms) following the MHC-I chains superposition

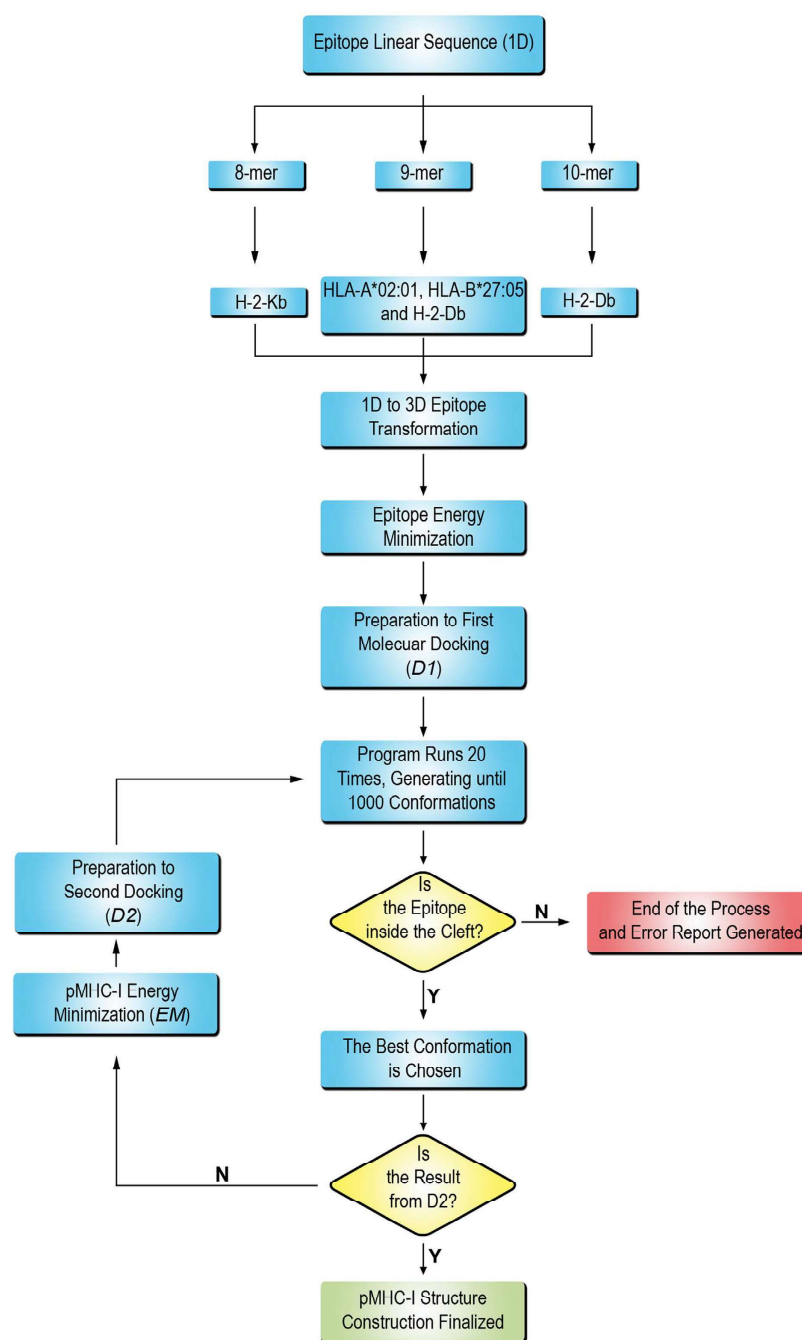


Figure 1. Flowchart showing the DockTope sequence of steps. The program starts from the linear sequence of the epitope and terminates when the pMHC-I structure is obtained. The user chooses among the allotypes HLA-A*02:01, HLA-B*27:05, H-2-Db and H-2-Kb, which depends on the epitope length (8-mer, 9-mer or 10-mer). The program prepares the files for the first docking (D1), where the best suited conformations will be saved during 20 rounds of simulation. The program checks whether the epitope is inside the cleft. In case of error, a report is written and the program stops. Otherwise, the program proceeds to the next step, where the best conformation is chosen. The program checks whether the structure generated is coming from D1. If so, the structure is energy minimized (EM) and a second docking is performed. In the end, the pMHC-I structure is generated in the PDB format.

of model and target. This way, it was possible to obtain a RMSD value taking into consideration not only conformational changes on the modelled epitope, but also translational and rotational differences inside the MHC-I cleft.

A cut-off of 2 Å or less was used to indicate the accuracy of the modelling approach. As observed in Table 1, the reproduction of all MHC-I allotypes produced average C α RMSD values below 2 Å. There were only two outliers, with C α RMSD values of 3.129 Å and 2.061 Å, respectively corresponding to the attempts of reproducing an HLA-A*02:01 (PDB ID: 2GTW) and an HLA-B*27:05 (PDB ID: 3BP4) peptide-loaded complex. When all atoms

Allotype	PDB ID (Resolution)		Reference
	MHC-I	Epitope	
HLA-A*02:01	2V2W (1.6 Å)	1T1Z (1.9 Å) (ALYNTAAAL)	53,54
HLA-B*27:05	2A83 (1.4 Å)	1JGE (2.1 Å) (GRFAAAIAK)	55,56
H-2-Db	1WBX (1.9 Å)	1JPG (2.2 Å, 9-mer epitope) (FQPQNGQFI)	57,58
H-2-Db	1WBY (2.3 Å)	1WBY (2.3 Å, 10-mer epitope) (SSELENFRAYV)	57
H-2-Kb	1LK2 (1.35 Å)	1RJY (1.9 Å) (SSIEFARL)	59,60

Table 1. PDB structures (MHC-I and epitope) used by DockTope.

MHC-I allotype	Epitope Length	Number of pMHC-I structures	RMSD (C α)		RMSD (all-atom)	
			Mean	s.d. (s.e.m)	Mean	s.d. (s.e.m)
HLA-A*02:01	9	68	0.926 Å	± 0.440 (0.053)	1.908 Å	± 0.678 (0.082)
HLA-B*27:05	9	10	1.027 Å	± 0.530 (0.167)	2.498 Å	± 1.224 (0.387)
H-2-Db	9	33	0.671 Å	± 0.331 (0.057)	1.899 Å	± 0.396 (0.069)
H-2-Db	10	5	0.439 Å	± 0.244 (0.109)	1.676 Å	± 0.590 (0.264)
H-2-Kb	8	19	1.132 Å	± 0.365 (0.083)	2.077 Å	± 0.412 (0.094)
TOTAL		135	0.882 Å	± 0.437 (0.037)	1.964 Å	± 0.655 (0.056)

Table 2. DockTope validation based on C α and all atoms RMSD average values, in angstroms. The standard deviation (s.d.) and the standard error of the mean (s.e.m.) are also provided.

were considered, the RMSD values were slightly higher in comparison to C α RMSD values. This was expected since epitope side chains could present high flexibility in the MHC-I cleft, especially the residues involved in the interaction with the T Cell Receptor (TCR)²². Considering all modelled epitopes, the overall RMSD average was 0.882 ± 0.437 Å (s.d.) and 1.964 ± 0.655 Å (s.d.) for C α and all atoms, respectively. As observed in Fig. 2, which shows the data distribution around the median with interquartile range, most of the C α RMSD measurements were grouped even below 1.5 Å, which strongly highlights the precision of the DockTope modelling approach. Of note, the median for C α /all atoms RMSD values were 0.854 Å/1.723 Å, 0.764 Å/2.578 Å, 0.629 Å/1.856 Å, 0.355 Å/1.685 Å and 1.292 Å/2.153 Å for HLA-A*02:01, HLA-B*27:05, H-2-Db (9-mer epitope), H-2-Db (10-mer epitope) and H-2-Kb, respectively. All 135 RMSD values for the evaluated structures are available in Supplementary Table S1 online.

We retrieved from the PDB the crystal resolution value (in angstrom) of each structure analysed here and calculated the average. Using a Kolmogorov-Smirnov Test, we applied tests of normality to the crystal resolution, C α RMSD and all atoms RMSD values for all allotypes. The HLA-A*02:01, HLA-B*27:05, H-2-Db (10-mer epitope) and H-2-Kb values were considered normally distributed ($p = 0.200$). Nevertheless, H-2-Db (9-mer epitope) values significantly deviated from normal distribution ($p = 0.006$). In this particular case, Kruskal-Wallis Test was performed. Each allotype was analysed individually (Fig. 3). It was observed that mean C α RMSD values of DockTope validation were significantly below crystal resolution values for HLA-A*02:01 ($p < 0.0001$), HLA-B*27:05 ($p = 0.049$), H-2-Db ($p < 0.0001$) and H-2-Kb ($p < 0.0001$). Also, in the case of H-2-Db (9-mer epitope), the all atoms RMSD mean value was also significantly below the crystal resolution value ($p = 0.0037$).

Web server. DockTope is a freely accessible tool available through the website dirac.cesup.ufgrs.br/bio/home.php, or from the CrossTope platform (<http://www.crosstope.com.br>) under the ‘Tools’ tab²³. First, the user should register an account. After that, the user receives an email with access data (login and password) to the site (Fig. 4a). To submit a new job, the user should provide a valid linear epitope sequence. The web server automatically recognizes the epitope sequence and provides a list with the possible MHC-I allotypes that can be used in the modelling (Fig. 4b). After submitting the job, the user can follow the process steps by clicking on the ‘Processing Jobs’ tab. A table is provided containing information about all jobs, such as Job ID, Job Name, Epitope sequence, MHC-I allele, Status, and Submission Date (Fig. 4c). After the job submission, a “Queued (qw)” flag is assigned. The time that the job stays with this flag will mainly depend on the demand. After that, the job proceeds to the “Running” state where the files are individually stored on our server. In case of error, the server stops the job. At the end of the process, a “Finalized” flag is assigned to the particular modelled epitope, and the pMHC-I structure file in the PDB format is sent to the registered email account provided by the user. The time spent on each job, after it enters the running state, will depend on the epitope sequence and allotype, though it should not exceed 6 hours. A 10-mer epitope of H-2-Db, for example, is expected to take longer, since the addition of one residue (in comparison with 9-mer epitopes) will increase dimensionality and require more computational time. It is also

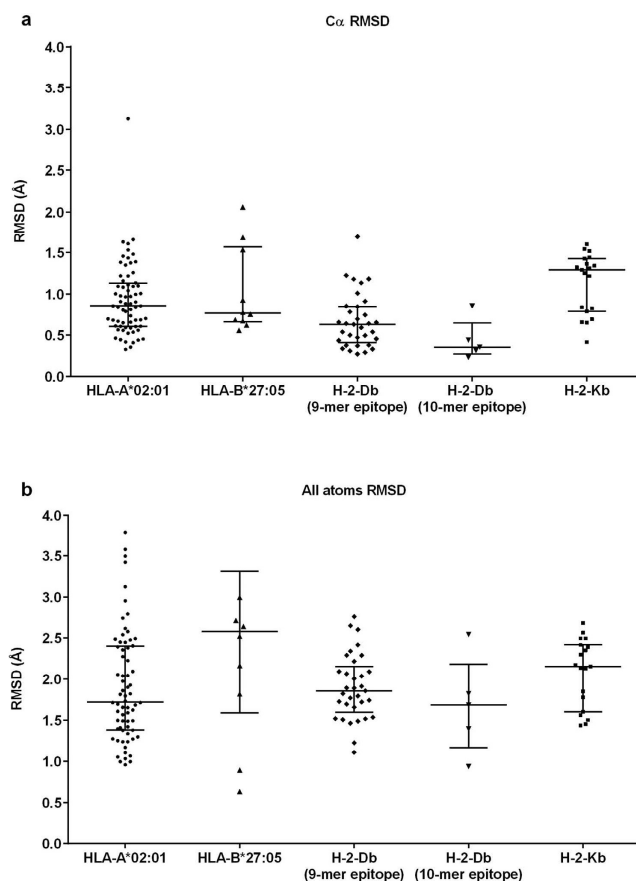


Figure 2. Scatter dot plot representing the DockTope validation values for 135 pMHC-I structures from the PDB. The validation process was performed through cross-docking, considering the C α (a) and all atoms (b) RMSD for each epitope. Each point represents the value for a reproduced structure. The statistic data are shown as a median with interquartile range (25% to 75%). On the y-axis, RMSD stands for Root Mean Square Deviation; on the x-axis, the MHC types are represented.

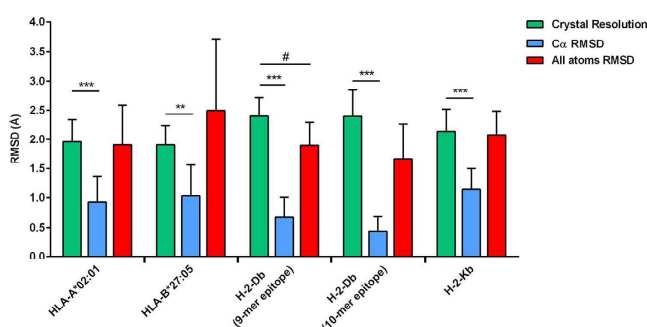


Figure 3. Graph with interleaved bars showing the mean C α (blue) and all atoms (red) RMSD values in comparison to the resolution values extracted from the PDB (green), for each MHC-I allotype. The C α RMSD mean value of all MHC-I allotypes was significantly below the crystal resolution mean values (*** $p < 0.0001$, ** $p = 0.049$). For H-2-Db (9-mer epitope), the all atoms RMSD mean value was also significantly below (# $p = 0.0037$). On y-axis, RMSD stands for Root Mean Square Deviation; on the x-axis, the MHC types are represented.

expected that epitopes with a large content of arginines, for example, take longer because of the increase in the number of side chain torsions.

The performance of the DockTope web server was assayed through the modelling of 238 immunogenic epitopes obtained from Immune Epitope Database and Analysis Resource^{24,25}. In the end, 226 epitopes were modelled by DockTope without any error and 12 were aborted along the process-six of MHC-I allotype H-2-Db (9-mer epitope) and six of H-2-Db (10-mer epitope). These results represent an accuracy of approximately 95%.

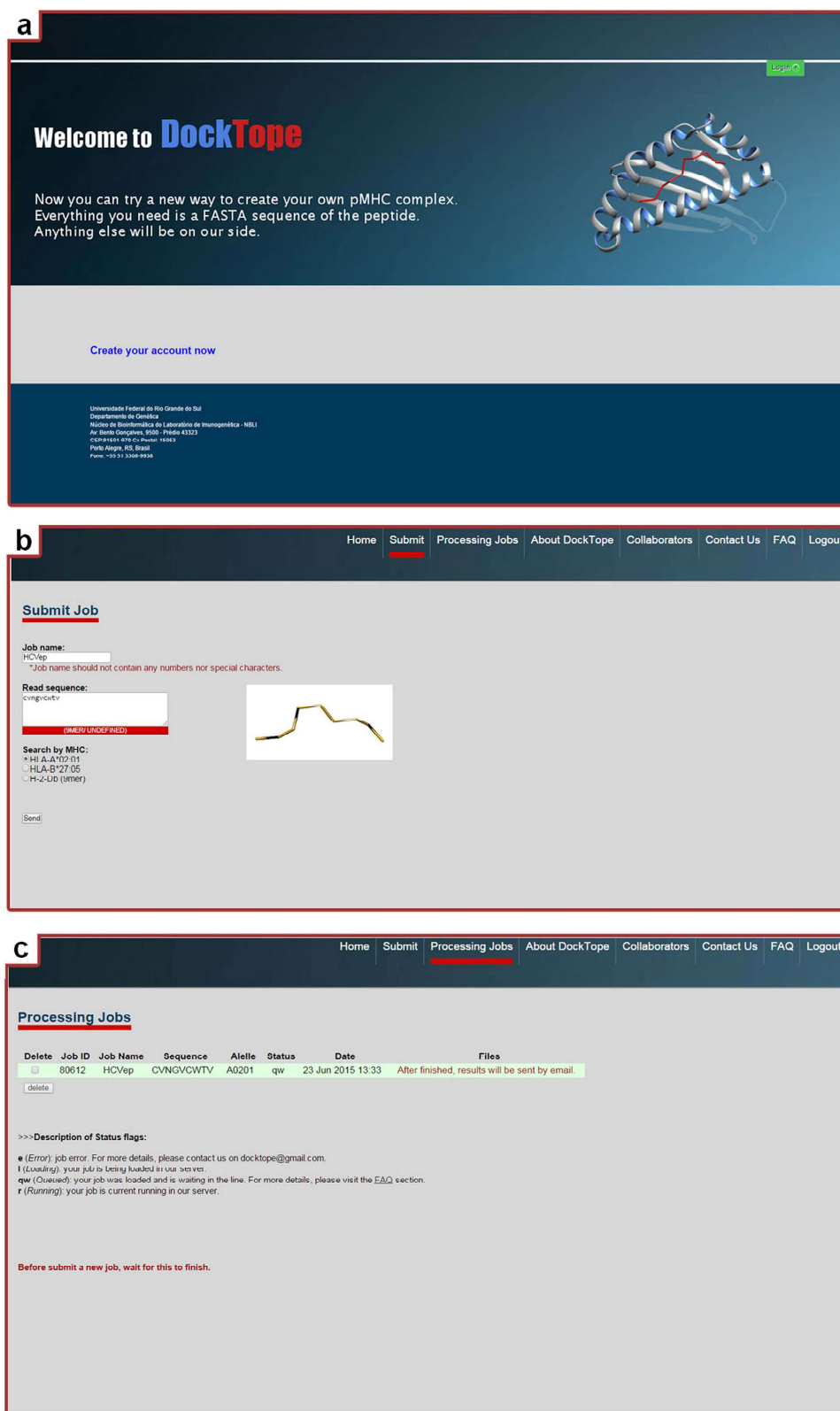


Figure 4. DockTope web interface. The first page the user will see is represented in (a), where it is possible to create a new account or directly access the tool with login and password. After the login, the user can submit a new sequence to be modeled, as represented in (b). Subsequently, the submitted job can be monitored through the “Processing jobs” tab (c).

Discussion

In this work we described a fully automated tool for the structural prediction of peptide:MHC-I (pMHC-I) complexes, DockTope, which was developed and validated for the MHC-I allotypes HLA-A*02:01, HLA-B*27:05, H-2-Kb and H-2-Db. DockTope was able to reproduce 135 crystal structures from the PDB with a RMSD mean value of 0.882 Å and 1.964 Å, for epitope C α and all atoms, respectively. The final accomplishment of this tool is (i) the complete automation, first presented here, of the established approach D1-EM-D2¹⁸, (ii) the modelling validation of all non-redundant pMHC-I structures available in the PDB at this time, and (iii) the tool availability as a web server for any researcher or user interested in pMHC-I modelling.

To automate and validate DockTope, a specific bash script was developed and executed in each step of the pMHC-I complex construction. The structure validation was performed using each pMHC-I structure in the PDB as a target to calculate the RMSD value with its respective model. Target and model were always fitted by MHC-I residues, which ensures that not only the difference between each epitope residue pair is considered, but also its displacement inside the MHC-I cleft after the molecular docking/energy minimization process. It should be noted that molecular docking programs can find unusual conformations after the searching process. To avoid this, DockTope performs a total of 20 rounds of molecular docking, generating up to 1000 conformations, which increases the probability of finding a proper epitope conformation. Still, unusual conformations can be generated (such as an inverted epitope inside the cleft or a protuberant C-terminal/N-terminal extremity pointing outside the MHC-I cleft). This phenomenon can be biologically explained, since some MHC-I allotypes do not have the capability to interact with determined epitope residues, but it can also be simply due to the fact that the docking algorithm was unable to find the correct solution. From the biological point of view, a work developed by Sidney *et al.* encompassing 945 HLA-A and HLA-B molecules reveals that some physicochemical specificities are not found in the evaluated MHC-I allotypes (considering B and F pocket residues), which in turn prevents the binding and presentation of peptides with such features²⁶. In order to avoid a misleading result, DockTope also automatically checks the epitope position and orientation (but not the binding affinity) after molecular docking, confirming its position inside the MHC-I cleft before proceeding to the search for the best pMHC-I structure. Of note, this is one of the most common sources of error reported by DockTope.

Before the implementation and automation of DockTope, the best pMHC-I structure was chosen through visual inspection only, where the most frequent conformation among the twenty generated was selected. This way, a user intervention was required, which could bias the result. Here, a new and improved algorithm is used to choose the best structure, based on the mean RMSD value between each epitope pair and on the binding energy value generated by AutoDock Vina.

Since there is a lack of pMHC-I crystal structures available in the PDB, our analysis was restricted to MHC-I allotype H-2-Kb (8-mer epitope), HLA-A*02:01 (9-mer epitope), HLA-B*27:05 (9-mer epitope), and H-2-Db (9-mer epitope and 10-mer epitope). This ensures that only experimentally-resolved protein structures are used to identify the MHC-I allotype-specific epitope pattern, which reinforces the technique specificity. Also, the low number of MHC-I allotypes available for modelling by DockTope should not be seen as a weakness, since it opens the theoretical possibility to model roughly 1.2×10^{13} pMHC-I structures, which would be unfeasible through X-ray crystallography or any other method currently available. Moreover, the importance of each one of these allotypes should be highlighted. The HLA-A*02 molecule is expressed by approximately half of the human population, and the HLA-A*02:01 allele is found in a relatively high frequency all over the world²⁷. For this reason, it is one of the most studied alleles. The HLA-B*27:05 has been associated with spondyloarthropathies disorders, such as ankylosing spondylitis^{28–30}, vaccine response^{31,32}, and HIV in elite controllers^{33,34}. The H-2-Db and H-2-Kb are widely-studied murine alleles, and recent studies have demonstrated its importance in synapse pruning of developing brain in murines^{35,36}.

Regarding the RMSD values for the DockTope validation (see Table 2 and Fig. 2), the overall RMSD average for all modelled epitopes, considering C α and all atoms, remained below 2 Å; this is considered a reference cut-off value indicating a valid crystal reproduction obtained through a cross-docking approach^{19,37–39}. In fact, the validation values are reinforced after the comparison of the C α and all atoms RMSD values by the average resolution value extracted from the PDB for all 135 structures analysed in this work (Fig. 3). The crystal resolution average value of the reproduced dataset, considering all allotypes, was 2.163 Å, which is higher than the RMSD value obtained for C α and all atoms (0.839 Å and 2.012 Å, respectively). Analysing each pMHC-I individually, we observed that all C α RMSD mean values were significantly below the crystal resolution mean. This indicates that subtle deviations between target and model are expected, especially because the epitope is not a rigid body inside the MHC-I cleft, and normal amino acid fluctuations can occur²². It came to our attention that only HLA-B*27:05-restricted complexes presented all atoms RMSD values greater than the respective crystal resolution average. This is attributed to the fact that HLA-B*27:05-restricted epitopes present a high proportion of arginine residues⁴⁰, containing long side chains, which in turn accounts for most of the RMSD deviation observed.

Most of the C α RMSD data was distributed below 1.5 Å, indicating a high precision of our technique (Fig. 2a). However, we observed an incoherent value of 3.129 Å for one of the epitopes bound to HLA-A*02:01. This value corresponds to the epitope LAGIGILTV derived from the MART-1/Melan-A protein (PDB ID: 2GTW). This epitope represents a variant of the 10-mer epitope ELAGIGILTV, which is recognized by MART-1-reactive T cells⁴¹. The interesting fact is that this 10-mer epitope presents a bulged conformation comprising the residues Gly-Ile-Gly-Ile, which is replicated by the 9-mer epitope LAGIGILTV, comprising the same residues. To adopt this conformation, the P1 leucine residue of the 9-mer epitope is inserted into the P2 pocket, exactly as it occurs with the 10-mer epitope⁴². This bulged conformation accounts for the major deviation values observed between the model and the crystal structure (Fig. 5). As discussed by Borbulevych *et al.*, this bulged conformation differs from other HLA-A*02:01 bound 9-mer epitopes from MART-1/Melan-A protein, such as ALGIGILTV (PDB ID: 2GTZ) and AAGIGILTV (PDB ID: 3QFD), which present a common extended conformation and incidentally produced better C α RMSD values here (1.435 Å and 1.377 Å, respectively). It is important to note that DockTope

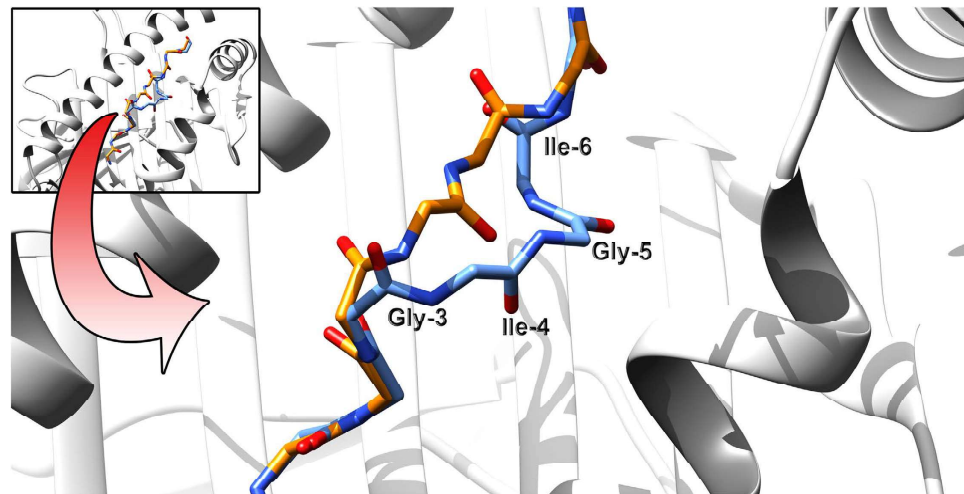


Figure 5. Epitope backbone comparison between the target (PDB ID: 2GTW), in blue, and the reproduced model, in orange, in the context of HLA-A*02:01. A top view of the pMHC-I is shown in the upper left. The arrow indicates the region (expanded at the centre) where most of the C α RMSD is observed, which accounts for a high value of 3.129 Å.

is based on a technique that uses epitope backbone patterns inside the MHC-I cleft; thus it is possible that unusual or aberrant epitope conformations will not be properly assessed.

Structures generated using DockTope can be used in several immunology fields, such as cancer research, transplantation, *in silico* stabilization assays and cross-reactivity assessment, expanding the range of possibilities to study these topics. In fact, our tool have already proven to be useful when studying cross-reactivity among different pMHC-I complexes. In a previous work, Principal Component Analysis (PCA) and Hierarchical Clustering Analysis (HCA) were employed to compare electrostatic potential data of TCR-interacting residues presented on the pMHC-I surface⁴³. A total of 28 known HCV targets (epitopes from NS3 protein) were modelled and analysed. The differences observed in PCA and HCA were evidences for structure-dependent immunogenic patterns and were in accordance with *in vitro* data of IFN- γ releasing assays⁴⁴. After that, 55 pMHC-I complexes including epitopes from different viral proteins were also modelled; this allowed us to infer other potentially cross-reactive targets with HCV-NS3₁₀₇₃, such as LMP2₃₂₉ from Epstein-Barr virus (EBV), Gag₇₇ from Human Immunodeficiency virus (HIV), and NA₂₃₁ from Influenza virus (IV). Of note, cross-reactive responses of NS3₁₀₇₃-specific CD8⁺ T cells against all of these targets were later confirmed through *in vitro* assays⁴⁵. Intriguingly, the linear sequence of the confirmed cross-reactive epitope EBV-LMP2₃₂₉ presents no similarities in amino acid sequence with the reference HCV-NS3₁₀₇₃ epitope, and shares only 33% of biochemical properties. Sequence-based analysis would most likely be unable to predict such cross-reactivity. However, an incredible resemblance is observed in a higher level of complexity, through the analysis of the TCR-interacting surface of the pMHC-I. Such analysis was made possible by modelling these pMHC-I complexes through D1-EM-D2, the approach behind DockTope (Fig. 6).

Three approaches stand out among previously published methodologies aiming at the pMHC structural prediction: (i) MHCsim¹⁰, (ii) pDOCK⁸, and (iii) a Biased-Probability Monte Carlo docking protocol published by Bordner and Abagyan⁹. MHCsim was the first automated server designed to model pMHC complexes. The server uses the input sequences (MHC and epitope) to perform a search in an internal database for the pMHC structure that is the most similar to the input sequences. Then, the template is modified at the positions where the residues differ to generate a new 3D structure. Some aspects not included in MHCsim are addressed by DockTope. The MHCsim methodology is based only on sequence similarity, which might not be sufficiently accurate to predict the 3D structure of a pMHC complex, especially when it comes to epitope conformation. This way, the provided pMHC structure is not final, but can be used for posterior refinement⁴⁶. Second, the MHCsim server allows the pMHC construction for human allotypes only, and is restricted to 9-mer epitopes. The pDOCK methodology is based mainly on ICM docking, Monte Carlo sampling and local minimization. In its validation, the authors presented C α RMSD values below 1 Å in a set of 186 pMHC-I and pMHC-II structures. However, contrary to DockTope, which used cross-docking to reproduce crystal structures, the validation process of pDOCK was performed through a re-docking approach and all atom RMSD values were not provided in the text. Also, pDOCK is currently not available as a web server, but only as an in-house protocol. The method published by Bordner and Abagyan is based on ICM docking, homology modelling and Support Vector Machine (SVM). They were able to reproduce through cross-docking a set of 14 HLA-A*02:01 epitopes and 9 H-2-Kb epitopes with epitope backbone RMSD values inferior to 1 Å. Like pDOCK, their method is not available as a web server.

DockTope emerges as a free, automated, well-validated (C α RMSD mean values below 1 Å), and user-friendly web-server tool for modelling pMHC-I complexes in a reliable way. Its usefulness was already demonstrated by previously published work. The possibility to construct pMHC-I complexes will open new avenues for structural immunoinformatics, hopefully triggering new discoveries in basic immunology and health applied sciences.

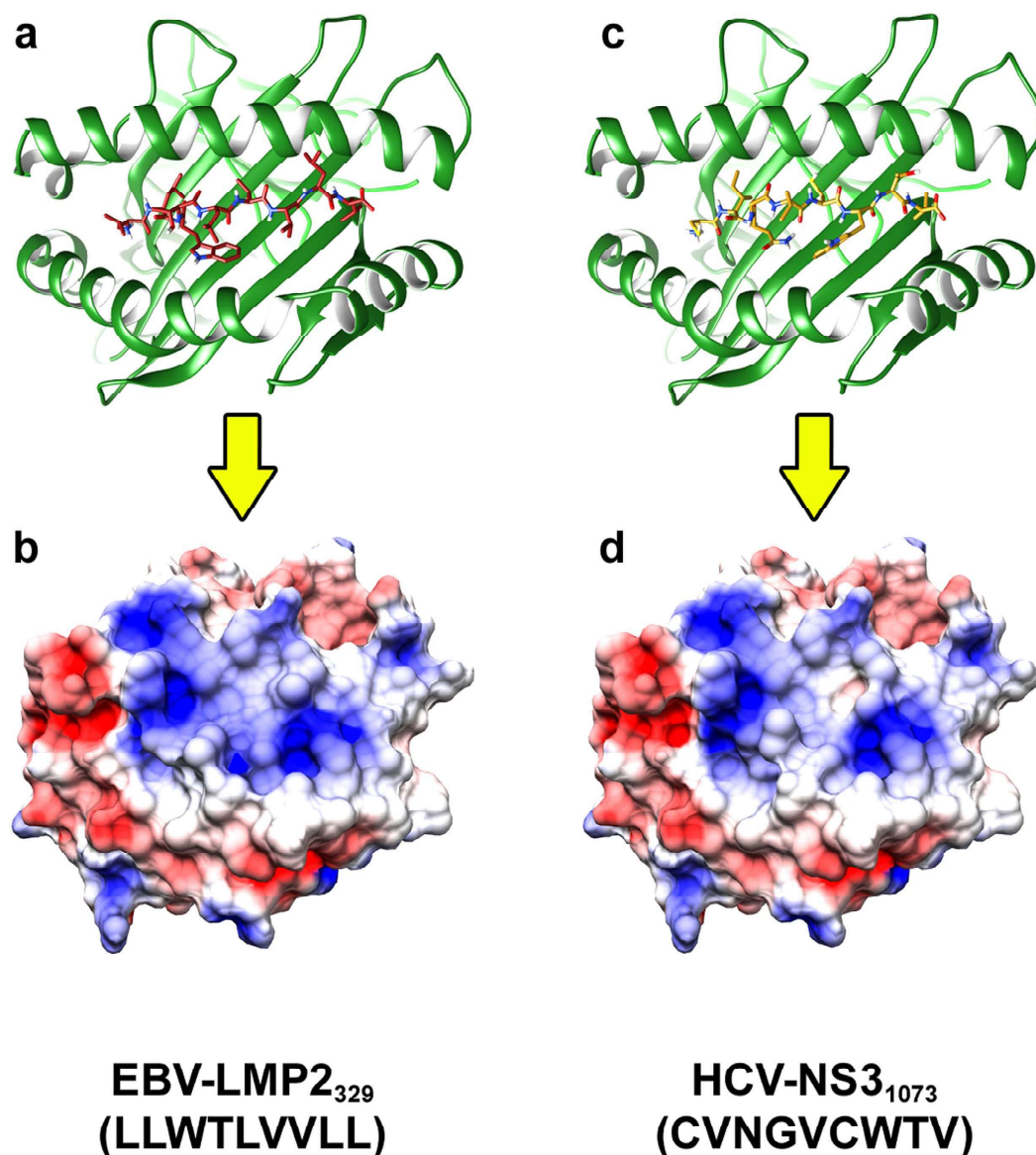


Figure 6. Two pMHC-I structures modelled using DockTope. In (a) and (c), the MHC-I (ribbon representation) and the epitope (stick representation) are depicted. In (b) and (d), the molecular surface of the TCR-interacting area was computed using UCSF Chimera package from the Computer Graphics Laboratory^{50,51} and the electrostatic potential was calculated using DelPhi⁵². The colour range (-3 kT to $+3\text{ kT}$, where k represents the Boltzmann constant and T represents the temperature) indicates the positive (blue), neutral (white) and negative (red) charges distributed on the pMHC-I surface. In (a) and (b), the epitope EBV-LMP2₃₂₉ (LLWTLVVLL) is represented and in (c) and (d) the epitope HCV-NS3₁₀₇₃ (CVNGVCWTV) is represented.

Methods

DockTope Automation. DockTope was developed as an optimized tool based on the D1-EM-D2 pMHC-I modelling approach. In order to automate the process, we employed a series of 9 shell scripts, 13 python scripts, 7 C++ executables and 2 python executables to perform the following steps: (i) Epitope structure modelling, (ii) first molecular docking (D1), (iii) choosing the best structure from D1, (iv) second molecular docking (D2), (v) choosing the best structure from D2 and (vi) writing the output. These steps are represented in the flowchart of Fig. 1 and in Supplementary Fig. S1.

Epitope structure modelling. The epitope to be modelled is provided as a linear amino acid sequence (without three-dimensional coordinates). A python script, which launches a built-in PyMOL⁴⁷ plug-in, uses the backbone of the epitope pattern to give shape to the modelled epitope. This epitope undergoes energy minimization, allowing for a mild global relaxation of the peptide.

First (D1) and Second (D2) Molecular Docking. Molecular docking is performed using the programs AutoDock Tools⁴⁸ and AutoDock Vina¹⁹; it involves three main steps. In the first step, the MHC-I molecule is prepared according to the following protocol: (i) adding all hydrogens, (ii) adding Gasteiger charges and (iii) removing non-polar hydrogens. In the second step, the modelled epitope is prepared by repeating the same protocol used for the MHC-I, but including an additional level: the torsion tree is set in a manner to maintain the epitope backbone rigid during the molecular docking process, thus only allowing the movement of side chains. In the final step, a box grid is configured to allow the search for the best epitope conformations inside the MHC-I cleft (formed by the $\alpha 1$ and $\alpha 2$ domains). The search for the best conformation is performed according to specific algorithms¹⁹ along twenty rounds (arbitrary value). In the end, the best structure is chosen according to an algorithm developed by our group.

Energy Minimization (EM). The energy minimization process is performed using the GROMACS package²¹. This process is used twice: on the modelled epitope and on the best pMHC-I complex produced by the first molecular docking, with the final goal of removing possible steric clashes and correcting distances between atoms in the system. The EM protocol is performed using a virtual cubic box filled with the protein and water (Simple Point Charge water model). Ions Na^+ and Cl^- are included to neutralize the system, maintaining a final concentration of 0.15 M/L. The GROMOS53a5 force field⁴⁹ is used to compute inter- and intramolecular interactions. The cut-off distances for the Coulomb (electrostatic and long-range attraction) and Lennard-Jones (repulsion and short-range attraction) forces are set to 1 nm. Molecular dynamics parameters include the steepest descent method of integration, with no constraints and a total of 10,000 steps, with an initial time step of 0.001 nm. The minimization converges after the maximum force is smaller than $2,000 \text{ kJ mol}^{-1} \text{ nm}^{-1}$; the lowest energy coordinates are then written to a file.

Choosing the best structure. The output of the first and second molecular docking process is composed of the best 20 epitope conformations in a set that could contain up to 1000 conformations. The best conformation, among these twenty, is chosen according to the following equations:

$$Co = \frac{\sum_{i=1}^{20} BE_i}{20} \quad (1)$$

$$BE_i < Co \rightarrow BE_i = n_i \quad (2)$$

$$M = n_i \times n_i \quad (3)$$

$$\bigcup_{k=1}^n k_{ie} \quad (4)$$

$$\frac{\sum_{k=1}^n RMSD_k}{n} = RMSD_j \quad (5)$$

$$\bigcup_{j=1}^n RMSD_j \quad (6)$$

$$Best = \min(RMSD_j) \quad (7)$$

Where, BE represents the binding energy value provided by AutoDock Vina for each epitope conformation ($i = 1, i = 2, \dots, i = 20$); Co represents a cut-off value based on the average of the twenty BE ; " n " represents the structure conformation chosen based on equation (2) and that will be used in the next steps of the calculation; " M " represents a matrix used to combine the selected data from equation (2); " k_{ie} " is the resultant file from the union (" \cup ") of all data from equation (3); " $RMSD_j$ " represents the file containing the RMSD average from the data contained in the k_{ie} files; and " $Best$ " is the final structure containing the three-dimensional coordinates of the chosen epitope.

DockTope validation. For each pMHC-I structure downloaded from the PDB, the epitope and MHC-I parts were separated. Next, the epitope linear sequence and the MHC-I allele name (according to Table 1) were used as input for a cross-docking process, using DockTope. At the end of the process, the modelled pMHC-I structure was compared to its respective structure available in the PDB; quantitative data was obtained through the RMSD analysis of the two structures, considering the $C\alpha$ and the all atoms RMSD displacement of the epitope. For the analysis, the structures available in the PDB were refined to contain only the pMHC-I structure, without TCR and possible ligands interfering with the peptide:MHC-I interaction. In the end, a total of 135 pMHC-I structures encompassing the MHC-I allotypes HLA-A*02:01, HLA-B*27:05, H-2-Db and H-2-Kb were evaluated (Table 2). The performance of DockTope was also evaluated through the modelling of 238 epitopes downloaded from the Immune Epitope Database and Analysis Resource (IEDB). The IEDB parameters for epitope search were set to contain only linear epitopes, from any disease, and confirmed by T cell assays (positive).

Web server. DockTope can be accessed through the CrossTope website (<http://www.crosstope.com.br>), in the “Tools” tab, or directly from dirac.cesup.ufrgs.br/bio/home.php. To use this tool, the user should sign in providing basic information such as name, email address, institution and academic degree. After logging in, the user will find the following tabs: Home, Submit, Processing Jobs, About DockTope, Collaborators, Contact Us and Frequent Asked Questions (FAQ). The web server includes two interfaces: user-tool and tool-server. The user-tool interface uses more than one programming language to better integrate all the modules. The visual module (web interface) was developed in PHP and jQuery Ajax, which are based on an HTML structure. All internal actions of the web interface are controlled and executed through JavaScript, especially processes validation, such as login validation, for example. The interface management and integration service available to the user, as well as the non-visible part (such as the execution of .js files) were obtained through the XAMPP server, which includes the APACHE, MySQL and PHP packages. The tool-server interface works exclusively through JavaScript (connection, submission and receipt of submitted jobs). A verification module works constantly over each created page to ensure the database connection, allowing access to the user. All the jobs, after being submitted, enter in a queue until the server checks and allows them to run.

Statistical analysis. The statistical analyses were performed using SPSS Software (IBM SPSS Statistics for Windows, Version 16.0. Armonk, NY: IBM Corp) and GraphPad Prism version 6.05 for Windows (GraphPad Software, La Jolla California USA, www.graphpad.com). We checked the normality of the data with the Kolmogorov-Smirnov Test, considering a level of significance of 0.05 ($p < 0.05$). We used one-way analysis of variance (one-way ANOVA) to perform multiple comparisons of the averages. The statistic of normal distribution data was analysed with Tukey’s post-hoc test. Data without normal distribution was analysed with the Kruskal-Wallis test.

References

1. Yewdell, J. W. & Bennink, J. R. Immunodominance in major histocompatibility complex class I-restricted T lymphocyte responses. *Annu Rev Immunol* **17**, 51–88 (1999).
2. Jorgensen, J. L., Esser, U., Fazekas de St Groth, B., Reay, P. A. & Davis, M. M. Mapping T-cell receptor-peptide contacts by variant peptide immunization of single-chain transgenics. *Nature* **355**, 224–230 (1992).
3. Wooldridge, L. *et al.* A single autoimmune T cell receptor recognizes more than a million different peptides. *J Biol Chem* **287**, 1168–1177 (2012).
4. He, L. *et al.* Integrated assessment of predicted MHC binding and cross-conservation with self reveals patterns of viral camouflage. *BMC Bioinformatics* **15**, (Suppl 4):S1 (2014).
5. Mungall, A. J. *et al.* The DNA sequence and analysis of human chromosome 6. *Nature* **425**, 805–811 (2003).
6. Park, H. J., Kim, J. Y., Jung, K. I. & Kim, T. J. Characterization of a Novel Gene in the Extended MHC Region of Mouse, NG29/Cd320, a Homolog of the Human CD320. *Immune Netw* **9**, 138–146 (2009).
7. Berko, D. *et al.* Membrane-anchored beta 2-microglobulin stabilizes a highly receptive state of MHC class I molecules. *J Immunol* **174**, 2116–2123 (2005).
8. Khan, J. M. & Ranganathan, S. pDOCK: a new technique for rapid and accurate docking of peptide ligands to Major Histocompatibility Complexes. *Immunome Research* **6**, (Suppl 1):S2 (2010).
9. Bordner, A. J. & Abagyan, R. Ab initio prediction of peptide-MHC binding geometry for diverse class I MHC allotypes *Proteins* **63**, 512–526 (2006).
10. Todman, S. J. *et al.* Toward the atomistic simulation of T cell epitopes automated construction of MHC: peptide structures for free energy calculations. *J Mol Graph Model* **26**, 957–961 (2008).
11. Schaffroth, H. D. & Floudas, C. A. Predicting peptide binding to MHC pockets via molecular modeling, implicit solvation, and global optimization. *Proteins* **54**, 534–556 (2004).
12. Tong, J. C., Tan, T. W. & Ranganathan, S. Modeling the structure of bound peptide ligands to major histocompatibility complex. *Protein Sci* **13**, 2523–2532 (2004).
13. Rognan, D., Lauemoller, S. L., Holm, A., Buus, S. & Tschinke, V. Predicting binding affinities of protein ligands from three-dimensional models: application to peptide binding to class I major histocompatibility proteins. *J Med Chem* **42**, 4650–4658 (1999).
14. Sezerman, U., Vajda, S. & DeLisi, C. Free energy mapping of class I MHC molecules and structural determination of bound peptides. *Protein Sci* **5**, 1272–1281 (1996).
15. Rosenfeld, R., Zheng, Q., Vajda, S. & DeLisi, C. Computing the structure of bound peptides. Application to antigen recognition by class I major histocompatibility complex receptors. *J Mol Biol* **234**, 515–521 (1993).
16. Antes, I. DynaDock: A new molecular dynamics-based algorithm for protein-peptide docking including receptor flexibility. *Proteins* **78**, 1084–1104 (2009).
17. Antes, I., Siu, S. W. & Lengauer, T. DynaPred: a structure and sequence based method for the prediction of MHC class I binding peptide sequences and conformations. *Bioinformatics* **22**, e16–24 (2006).
18. Antunes, D. A. *et al.* Structural allele-specific patterns adopted by epitopes in the MHC-I cleft and reconstruction of MHC:peptide complexes to cross-reactivity assessment. *PLoS One* **5**, e10353 (2010).
19. Trott, O. & Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry* **31**, 455–461 (2009).
20. Zhou, H. X. & Gilson, M. K. Theory of free energy and entropy in noncovalent binding. *Chem Rev* **109**, 4092–4107 (2009).
21. Van Der Spoel, D. *et al.* GROMACS: fast, flexible, and free. *Journal of Computational Chemistry* **26**, 1701–1718 (2005).
22. Reboul, C. F., Meyer, G. R., Porebski, B. T., Borg, N. A. & Buckle, A. M. Epitope flexibility and dynamic footprint revealed by molecular dynamics of a pMHC-TCR complex. *PLoS Comput Biol* **8**, e1002404 (2012).
23. Sinigaglia, M., Antunes, D. A., Rigo, M. M., Chies, J. A. & Vieira, G. F. CrossTope: a curate repository of 3D structures of immunogenic peptide: MHC complexes. *Database (Oxford)* **2013**, bat002 (2013).
24. Vita, R. *et al.* *Immune Epitope Database and Analysis Resource*. (2015) Available at: www.iedb.org. (Accessed: 6th October 2015)
25. Vita, R. *et al.* The immune epitope database (IEDB) 3.0. *Nucleic Acids Res* **43**, D405–412 (2014).
26. Sidney, J., Peters, B., Frahm, N., Brander, C. & Sette, A. HLA class I supertypes: a revised and updated classification. *BMC Immunology* **9**, 1 (2008).
27. Choo, J. A., Liu, J., Toh, X., Grotenbreg, G. M. & Ren, E. C. The immunodominant influenza A virus M158-66 cytotoxic T lymphocyte epitope exhibits degenerate class I major histocompatibility complex restriction in humans. *J Virol* **88**, 10613–10623 (2014).
28. Nasution, A. R. *et al.* HLA-B27 subtypes positively and negatively associated with spondyloarthritis. *J Rheumatol* **24**, 1111–1114 (1997).

29. Abualrous, E. T. *et al.* F pocket flexibility influences the tapasin dependence of two differentially disease-associated MHC Class I proteins. *Eur J Immunol* **45**, 1248–1257 (2015).
30. Powis, S. J., Santos, S. G. & Antoniou, A. N. Biochemical features of HLA-B27 and antigen processing. *Adv Exp Med Biol* **649**, 210–216 (2009).
31. Posteraro, B. *et al.* The link between genetic variation and variability in vaccine responses: systematic review and meta-analyses. *Vaccine* **32**, 1661–1669 (2014).
32. Ovsyannikova, I. G., Pankratz, V. S., Larrabee, B. R., Jacobson, R. M. & Poland, G. A. HLA genotypes and rubella vaccine immune response: additional evidence. *Vaccine* **32**, 4206–4213 (2014).
33. Loffredo, J. T. *et al.* Two MHC class I molecules associated with elite control of immunodeficiency virus replication, Mamu-B*08 and HLA-B*2705, bind peptides with sequence similarity. *J Immunol* **182**, 7763–7775 (2009).
34. Kaslow, R. A. *et al.* Influence of combinations of human major histocompatibility complex genes on the course of HIV-1 infection. *Nat Med* **2**, 405–411 (1996).
35. Adelson, J. D. *et al.* Developmental Sculpting of Intracortical Circuits by MHC Class I H2-Db and H2-Kb. *Cereb Cortex*, 1–11 (2014).
36. Lee, H. *et al.* Synapse elimination and learning rules co-regulated by MHC class I H2-Db. *Nature* **509**, 195–200 (2014).
37. Bergeron, B. *Bioinformatics Computing 1st edn* (Prentice Hall, 2002).
38. Bagaria, A., Jaravine, V., Huang, Y. J., Montelione, G. T. & Guntert, P. Protein structure validation by generalized linear model root-mean-square deviation prediction. *Protein Sci* **21**, 229–238 (2011).
39. Madurga, S., Belda, I., Llorca, X. & Giralt, E. Design of enhanced agonists through the use of a new virtual screening method: application to peptides that bind class I major histocompatibility complex (MHC) molecules. *Protein Sci* **14**, 2069–2079 (2005).
40. Madden, D. R., Gorga, J. C., Strominger, J. L. & Wiley, D. C. The structure of HLA-B27 reveals nonamer self-peptides bound in an extended conformation. *Nature* **353**, 321–325 (1991).
41. Kawakami, Y. *et al.* Identification of the immunodominant peptides of the MART-1 human melanoma antigen recognized by the majority of HLA-A2-restricted tumor infiltrating lymphocytes. *J Exp Med* **180**, 347–352 (1994).
42. Borbulevych, O. Y. *et al.* Structures of MART-126/27-35 Peptide/HLA-A2 complexes reveal a remarkable disconnect between antigen structural homology and T cell recognition. *J Mol Biol* **372**, 1123–1136 (2007).
43. Antunes, D. A. *et al.* Structural in silico analysis of cross-genotype-reactivity among naturally occurring HCV NS3-1073-variants in the context of HLA-A*02:01 allele. *Molecular Immunology* **48**, 1461–1467 (2011).
44. Fyttili, P. *et al.* Cross-genotype-reactivity of the immunodominant HCV CD8 T-cell epitope NS3-1073. *Vaccine* **26**, 3818–3826 (2008).
45. Zhang, S. *et al.* Frequency, private specificity and cross-reactivity of pre-existing HCV-specific CD8 + T cells in HCV seronegative individuals: implication for vaccine responses. *J Virol* **89**, 8304–8317 (2015).
46. Vivona, S. *et al.* Computer-aided biotechnology: from immuno-informatics to reverse vaccinology. *Trends Biotechnol* **26**, 190–200 (2008).
47. Schrodinger, L. L. C. The PyMOL Molecular Graphics System, Version 1.3r1 (2010).
48. Morris, G. M. *et al.* AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *Journal of Computational Chemistry* **30**, 2785–2791 (2009).
49. Oostenbrink, C., Villa, A., Mark, A. E. & van Gunsteren, W. F. A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6. *Journal of Computational Chemistry* **25**, 1656–1676 (2004).
50. Pettersen, E. F. *et al.* UCSF Chimera—a visualization system for exploratory research and analysis. *J Comput Chem* **25**, 1605–1612 (2004).
51. Sanner, M. F., Olson, A. J. & Spehner, J. C. Reduced surface: an efficient way to compute molecular surfaces. *Biopolymers* **38**, 305–320 (1996).
52. Li, L. *et al.* DelPhi: a comprehensive suite for DelPhi software and associated resources. *BMC Biophys* **5**, 9 (2012).
53. Lee, J. K. *et al.* T cell cross-reactivity and conformational changes during TCR engagement. *J Exp Med* **200**, 1455–1466 (2004).
54. Martinez-Hackert, E. *et al.* Structural basis for degenerate recognition of natural HIV peptide variants by cytotoxic lymphocytes. *J Biol Chem* **281**, 20205–20212 (2006).
55. Ruckert, C. *et al.* Conformational dimorphism of self-peptides and molecular mimicry in a disease-associated HLA-B27 subtype. *J Biol Chem* **281**, 2306–2316 (2006).
56. Hulsmeyer, M. *et al.* HLA-B27 subtypes differentially associated with disease exhibit subtle structural alterations. *J Biol Chem* **277**, 47844–47853 (2002).
57. Meijers, R. *et al.* Crystal structures of murine MHC Class I H-2 D(b) and K(b) molecules in complex with CTL epitopes from influenza A virus: implications for TCR repertoire selection and immunodominance. *J Mol Biol* **345**, 1099–1110 (2005).
58. Ciatto, C. *et al.* Zooming in on the hydrophobic ridge of H-2D(b): implications for the conformational variability of bound peptides. *J Mol Biol* **312**, 1059–1071 (2001).
59. Rudolph, M. G. *et al.* A peptide that antagonizes TCR-mediated reactions with both syngeneic and allogeneic agonists: functional and structural aspects. *J Immunol* **172**, 2994–3002 (2004).
60. Miley, M. J. *et al.* Structural basis for the restoration of TCR recognition of an MHC allelic variant by peptide secondary anchor substitution. *J Exp Med* **200**, 1445–1454 (2004).

Acknowledgements

This work was supported by *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq), *Coordenação De Aperfeiçoamento De Pessoal De Nível Superior* (CAPES), and *Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul* (FAPERGS). We would like to thank the *Centro de Supercomputação da UFRGS* (CESUP-UFRGS) for providing its facilities and support. The authors would also like to thank Caio Diniz de Farias, Marina Roberta Scheid, Renata Fioravanti Tarabini and Marcelo Alves Bragatte de Souza for their involvement in this project, as well as Dr. José Artur Bogo Chies, for the support and immunological discussions. Finally, we thank Dr. Didier Devaurs for his helpful comments on the final manuscript.

Author Contributions

M.M.R., D.A.A., M.S. and G.F.V. conceived the study. M.M.R. wrote the paper. M.M.R. and D.A.A. developed the scripts for the DockTope automation. M.M.R., D.A.A., M.F.A.M., M.V.F. conducted experiments. M.M.R., D.A.A., M.F.A.M., M.V.F., L.M., M.S. and G.F.V. analysed and interpreted the data. M.F. and L.M. were responsible for web-server development and the database integration with CESUP-UFRGS. M.M.R., D.A.A., M.F.A.M., M.V.F., M.S., L.M. and G.F.V. revised the manuscript for intellectual content and approved the final version to be published.

Additional Information

Supplementary information accompanies this paper at <http://www.nature.com/srep>

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Menegatti Rigo, M. *et al.* DockTope: a Web-based tool for automated pMHC-I modelling. *Sci. Rep.* **5**, 18413; doi: 10.1038/srep18413 (2015).



This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>



Interpreting T-Cell Cross-reactivity through Structure: Implications for TCR-Based Cancer Immunotherapy

Dinler A. Antunes^{1,2}, Maurício M. Rigo^{1,3}, Martiela V. Freitas¹, Marcus F. A. Mendes¹, Marialva Sinigaglia¹, Gregory Lizée⁴, Lydia E. Kavraki², Liisa K. Selin⁵, Markus Cornberg^{6,7} and Gustavo F. Vieira^{1,8*}

¹Núcleo de Bioinformática do Laboratório de Imunogenética (NBLI), Department of Genetics, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, Brazil, ²Kavraki Lab, Department of Computer Science, Rice University, Houston, TX, United States, ³Laboratório de Imunologia Celular e Molecular, Instituto de Pesquisas Biomédicas (IPB), Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS), Porto Alegre, Brazil, ⁴Lizée Lab, Department of Melanoma Medical Oncology – Research, The University of Texas M. D. Anderson Cancer Center, Houston, TX, United States, ⁵Selin Lab, Department of Pathology, University of Massachusetts Medical School, Worcester, MA, United States, ⁶Cornberg Lab, Department of Gastroenterology, Hepatology and Endocrinology, Hannover Medical School, Hannover, Germany, ⁷German Center for Infection Research (DZIF), Partner-Site Hannover-Braunschweig, Hannover, Germany, ⁸Programa de Pós-Graduação em Saúde e Desenvolvimento Humano, Universidade La Salle, Porto Alegre, Brazil

OPEN ACCESS

Edited by:

Cyrille J. Cohen,
Bar-Ilan University, Israel

Reviewed by:

Alessandro Poggi,
Ospedale Policlinico
San Martino, Italy
Zsolt Sebestyén,
University Medical Center
Utrecht, Netherlands

*Correspondence:

Gustavo F. Vieira
gustavo.vieira@unilasalle.edu.br

Specialty section:

This article was submitted
to Cancer Immunity
and Immunotherapy,
a section of the journal
Frontiers in Immunology

Received: 24 July 2017

Accepted: 12 September 2017

Published: 04 October 2017

Citation:

Antunes DA, Rigo MM, Freitas MV,
Mendes MFA, Sinigaglia M, Lizée G,
Kavraki LE, Selin LK, Cornberg M
and Vieira GF (2017) Interpreting
T-Cell Cross-reactivity through
Structure: Implications for TCR-
Based Cancer Immunotherapy.
Front. Immunol. 8:1210.
doi: 10.3389/fimmu.2017.01210

Immunotherapy has become one of the most promising avenues for cancer treatment, making use of the patient's own immune system to eliminate cancer cells. Clinical trials with T-cell-based immunotherapies have shown dramatic tumor regressions, being effective in multiple cancer types and for many different patients. Unfortunately, this progress was tempered by reports of serious (even fatal) side effects. Such therapies rely on the use of cytotoxic T-cell lymphocytes, an essential part of the adaptive immune system. Cytotoxic T-cells are regularly involved in surveillance and are capable of both eliminating diseased cells and generating protective immunological memory. The specificity of a given T-cell is determined through the structural interaction between the T-cell receptor (TCR) and a peptide-loaded major histocompatibility complex (MHC); i.e., an intracellular peptide–ligand displayed at the cell surface by an MHC molecule. However, a given TCR can recognize different peptide–MHC (pMHC) complexes, which can sometimes trigger an unwanted response that is referred to as T-cell cross-reactivity. This has become a major safety issue in TCR-based immunotherapies, following reports of melanoma-specific T-cells causing cytotoxic damage to healthy tissues (e.g., heart and nervous system). T-cell cross-reactivity has been extensively studied in the context of viral immunology and tissue transplantation. Growing evidence suggests that it is largely driven by structural similarities of seemingly unrelated pMHC complexes. Here, we review recent reports about the existence of pMHC “hot-spots” for cross-reactivity and propose the existence of a TCR interaction profile (i.e., a refinement of a more general TCR footprint in which some amino acid residues are more important than others in triggering T-cell cross-reactivity). We also make use of available structural data and pMHC models to interpret previously reported cross-reactivity patterns among virus-derived peptides. Our study provides further evidence that structural analyses of pMHC complexes can be used to assess the intrinsic likelihood of cross-reactivity among peptide-targets. Furthermore,

we hypothesize that some apparent inconsistencies in reported cross-reactivities, such as a preferential directionality, might also be driven by particular structural features of the targeted pMHC complex. Finally, we explain why TCR-based immunotherapy provides a special context in which meaningful T-cell cross-reactivity predictions can be made.

Keywords: T-cell cross-reactivity, peptide–MHC complex, cross-reactivity hot-spots, TCR-interacting surface, hierarchical clustering, TCR/pMHC, cancer immunotherapy

1. HYPOTHESIS AND THEORY

1.1. Cellular Immunity, Private Specificity, and T-Cell Cross-reactivity

Cellular immunity relies on T-cell lymphocytes and their ability to produce unique T-cell receptors (TCRs), while humoral immunity relies on B-cell lymphocytes and their ability to produce antibodies (also referred to as B-cell receptors) (1, 2). Combined, these two branches compose the adaptive immunity, a major “upgrade” in the evolution of the immune system, first seen in jawed vertebrates (1, 2). Different from more ancestral mechanisms of innate immunity, adaptive immunity allows creating specific immune responses to virtually any new pathogen encountered by the host organism. It also allows generating immunological memory, protecting the host against future encounters with the same pathogen (3). This new system was essential in facing the threat of viruses, which are incredibly diverse and evolve at an amazing rate (4). While antibodies can neutralize circulating viruses, cytotoxic T-cells can find and eliminate infected cells (i.e., the “hijacked factories” producing new viral particles). In fact, coevolution with viruses is a major factor shaping the complexity and diversity of the mechanisms involved in cellular immunity (5–7).

The key players in this system are the major histocompatibility complex (MHC) molecules, a diverse set of protein receptors capable of binding peptides derived from intracellular proteins and displaying them at the cell surface (5). This allows circulating cytotoxic T-cells to interact directly with these peptide–MHC (pMHC) complexes, using their TCRs. After a complex selection process in early stages of their development (8, 9), T-cells are able to recognize “non-self” pMHC complexes. For instance, a virus-infected cell displays at its surface MHC molecules loaded with virus-derived peptides. These non-self pMHC complexes can trigger a T-cell response that, in turn, eliminates the infected cell. Moreover, the recognition of these non-self pMHCs can generate immunological memory against this particular virus strain (3).

The efficiency of antiviral immunity, however, depends on the ability of an individual to produce and store a pool of memory T-cells (i.e., a T-cell *repertoire*) able to specifically recognize most of the hugely variable pMHC complexes displayed by cells in different tissues. It actually is quite a puzzling task, if one considers (i) the diversity of MHC allotypes of the host (i.e., the number of MHC protein variants in the human population), (ii) the genetic variability of viruses (i.e., peptide diversity), and (iii) the frequency of viral infections. The solution to this puzzle involves a combination of two important features of cellular immunity: (i) *somatic recombination* of TCR-encoding genes and (ii) *T-cell cross-reactivity*. Somatic recombination allows for a potential

combinatorial diversity of TCRs which exceeds 10^{20} (10, 11). Cross-reactivity allows optimizing the repertoire of T-cells for the recognition of most possible targets, despite the limited number of T-cells that can exist in a given individual, at a given time ($\approx 10^{11}$ in humans) (10, 12). Each newly generated T-cell has a unique TCR and is added to the diverse repertoire of circulating T-cells. If activated by a given pMHC, one T-cell generates an entire pool of clone cells (referred to as a *T-cell line*). All these clones display essentially the same TCR and, therefore, are specific to the same (*cognate*) pMHC. However, after being added to the memory pool, some of these T-cells can be recruited in an initial response to a different *heterologous* pMHC (e.g., the same MHC displaying the peptide of a different virus).

T-cell cross-reactivity is defined as the ability of a given T-cell to be activated by two or more heterologous pMHCs (12). This cross-reactivity can even mediate *heterologous immunity*, when a contact with one pathogen generates a partial immunity against a second (heterologous) pathogen (13). Heterologous immunity is a double-edged sword: it can be protective and desired for wide spectrum vaccine development (14, 15), but it can also mediate impaired cellular response, chronic infection and immunopathology (15–18). The stochastic nature of TCR specificity generation entails that each individual has a unique set of TCRs (referred to as *private specificity*) (13). In addition, given the size limit of the T-cell repertoire and the constant challenges with a variety of pathogens, the memory pool of an individual is ever changing (e.g., some T-cell lines expand, others are lost) (19, 20). In time, cross-reactive cells represent an important part of our memory repertoire, and our immunity against every new challenge is directly influenced by our *immunological history* (12, 19, 21–23). Note that there exist some known biases in the somatic recombination process, producing some TCR sequence combinations with higher frequency in a population (24). This phenomenon is referred to as *public TCR usage* and will be discussed later (see section 1.4).

Recent studies are corroborating the idea that T-cell cross-reactivity is the rule, rather than the exception (19, 22, 25, 26), and that structural features involved in specific TCR/pMHC interactions are the main features driving cross-reactive responses against heterologous targets (25, 27–29). Despite all the evidence accumulated in the context of viral immunity and tissue transplantation, integration of T-cell cross-reactivity into other fields of immunology and human health has been rather slow. This delay can be partially explained by the complexity of the mechanisms involved, as well as concerns about the reproducibility of experimental results characterizing T-cell cross-reactivity (26).

In a pioneering study, Wedemeyer and colleagues were able to collect T-cells recognizing a peptide derived from hepatitis

C virus (HCV), from the blood of healthy donors (30) who had no history of infection by HCV. This implied that these HCV-specific T-cells were probably cross-reactive memory cells previously triggered by a heterologous pathogen. In fact, the authors were able to identify a peptide from influenza A virus (IAV) having 77% of sequence similarity with the HCV-derived peptide used to expand the T-cells. They also showed that these cells were able to recognize both peptides, and that T-cells with the same specificity were generated in response to IAV infection. However, a later study by Kasprowicz et al. (31) suggested that cross-reactivity between these heterologous peptides was rather weak and had a preferential *directionality* from HCV to IAV (i.e., T-cells primed with the HCV-derived peptide also recognize the IAV-derived peptide, but the opposite was usually not true) (31). More recent studies help clarify situations like this, showing that heterologous immunity between viruses is greatly influenced by private specificities and immunological history (19, 23, 32). Therefore, observed results are not solely determined by peptide sequence similarity, but also dependent on the particular T-cells dominating the response (*in vivo*), or the T-cell line selected for the experiments (*in vitro* or *ex vivo*) (24).

The rebirth of T-cell cross-reactivity as a major interest for human health, however, is coming from cancer research. For decades, immunologists have suggested that the same mechanisms involved in antiviral surveillance were also involved in detecting and eliminating cancer cells, which can display MHCs loaded with tumor-specific peptides (33). More recently, the field of cancer immunotherapy has grown as one of the most promising paths for cancer treatment, relying on the mechanisms of cellular immunity to provide personalized therapies that can eliminate tumors in different tissues and even generate protective memory (33–36). A number of TCR-based therapies were put forward, making use of the latest molecular biology technologies to enhance TCR affinity against tumor-specific peptides (37). Unfortunately, the excitement was tempered by safety concerns. These supposedly tumor-specific T-cells can present unexpected T-cell cross-reactivities in some individuals, attacking healthy tissues (38). In fact, off-target toxicity effects have been observed in recent clinical trials, with at least 5 deadly cases reported (39–41). Two of these cases were clearly linked to T-cell cross-reactivity between the targeted tumor-specific peptide (the melanoma-associated antigen MAGE-A3) and a Titin-derived peptide expressed in healthy cardiac cells (39, 42). The peptides involved have only 55% of sequence similarity, exemplifying the great challenge faced by current preclinical screenings. Later analysis using X-ray crystallography confirmed the structural similarity of the corresponding pMHC complexes as the molecular basis for the observed T-cell cross-reactivity (43).

In response to this critical need, new computational approaches are being developed and tested to improve our capacity to screen for potentially dangerous cross-reactivities. Some of these methods involve assessing peptide sequence similarity, while also accounting for protein tissue expression and MHC binding (44, 45). Others are based on pMHC structural similarity (46–48) or some combination of previously mentioned features (49, 50). Despite the incredible challenge at hand and the current limitations of these computational methods, encouraging results are

being reported. For instance, some of these methods can predict the previously mentioned cross-reactivity between the peptides derived from MAGE-A3 and Titin. A better understanding of the mechanisms underlying T-cell cross-reactivity, as well as the relationship between structural features of pMHC complexes and the activation of T-cell clones, is of utmost importance to further improve these computational methods. In turn, such progress will allow us to provide useful predictions that can be directly translated to the clinic.

In the following sections we attempt to connect the dots between the current understanding of pMHC structure and the goal of making safer TCR-based immunotherapies. First, we review structural aspects of the TCR/pMHC interaction and introduce the idea of structural clustering of pMHC complexes (section 1.2). Then we apply clustering methods to both available crystallographic data and modeled pMHC complexes, providing further evidence that pMHC structural information is essential to understand T-cell cross-reactivity (section 1.3). Next, we review how structural features of the pMHC complex can actually shape the TCR repertoire (section 1.4). Going one step further, we hypothesize how the same features might be shaping different patterns of cross-reactivity: they can be responsible for weak cross-reactivity among similar peptide-targets (section 1.5), or, conversely, drive cross-reactive responses among completely unrelated peptide-targets (section 1.6). Finally, we consider the implications of our work for T-cell cross-reactivity prediction and discuss why cancer immunotherapy provides a special context in which meaningful progress can be made (section 1.7).

1.2. Structural Analyses Can Uncover Key Features for T-Cell Activation

For simplicity, we usually talk about cross-reactivity of TCRs that recognize different peptides, but it is important to keep in mind that the TCR does not recognize the peptide itself; it recognizes the combined surface of the pMHC complex (51). Therefore, observed cross-reactivities between peptides are linked to their presentation in the “context” of a particular MHC. Even if two different MHCs are capable of binding the same peptide, which is not common, the resulting pMHC complexes will most likely be different (52). In fact, this is one of the causes for rejection in (allogeneic) tissue transplantation (26, 53). In this study, we focus on cross-reactivity between peptides presented by the same class I MHC. However, cross-reactivity involving different MHCs has also been reported (53, 54), and the discussion presented here can also be extended to that context.

Studies using X-ray crystallography have greatly contributed to the current understanding of the TCR/pMHC interaction, which was recently reviewed by Degauque et al. (26). The TCR structure contains flexible loops that can come in contact with the *TCR-interacting surface* of the pMHC (i.e., the “face” of the pMHC complex exposed to TCR interaction; see Figures S1A–C in Supplementary Material). These loops include the complementarity-determining regions (CDRs), which are the most variable regions of the TCR structure and the result of the previously mentioned somatic recombination. Despite the structural flexibility

of these loops and the possibility of local conformational changes (25, 55), there is a conserved binding mode for the TCR/pMHC complex. Most times, the CDRs corresponding to the α chain of the TCR will interact with the amino-terminal portion of the peptide, while the β chain CDRs will interact with the carboxy-terminal of the peptide, at a particular angle (26, 51) (Figure S1D in Supplementary Material). Note that the general docking mode of a TCR to its cognate pMHC is referred to as the *TCR footprint* (51). Although the mechanisms are still open for debate, recent studies suggest that the orientation of the TCR footprint is guided by genetically imprinted biases (on the TCR) to recognize conserved MHC amino acid residues (i.e., *germline bias*) (26, 29). However, with the accumulation of crystal structures and evidence from new experimental approaches, one can also see that different TCRs establish different interaction networks, and that some interactions on the pMHC surface seem more important than others to trigger recognition by a particular T-cell (24, 29). These special contacts have been previously referred to as *hot-spots* for T-cell cross-reactivity (25, 29, 56).

In previous work, our group described an *in silico* approach to evaluate the structural similarity of pMHC complexes (46, 48). We used hierarchical clustering as a tool to group pMHC complexes according to the similarity of their TCR-interacting surfaces. We also used available crystal structures as a reference to implement a method to model pMHC complexes for which no structural data were available (52, 57). Combining these methods, we were able to reproduce experimentally observed cross-reactivity patterns for a dataset of 28 naturally occurring variants of an HCV-derived peptide used for vaccine development (CINGVCWTV) (46). We also applied these methods to predict potential cross-reactivities between this HCV vaccine peptide and a dataset of non-related virus-derived peptides, in the context of a particular human MHC (HLA-A*02:01) (46). Our predictions were later confirmed by *in vitro* and *ex vivo* experiments (47), highlighting the prospecting potential of our methods. One of the detected cross-reactive peptides, derived from Epstein–Barr virus (LLWTLVLL), shared no sequence similarity with the vaccine peptide. Notwithstanding, both peptides show remarkably similar TCR-interacting surfaces when bound to HLA-A*02:01 (46, 47).

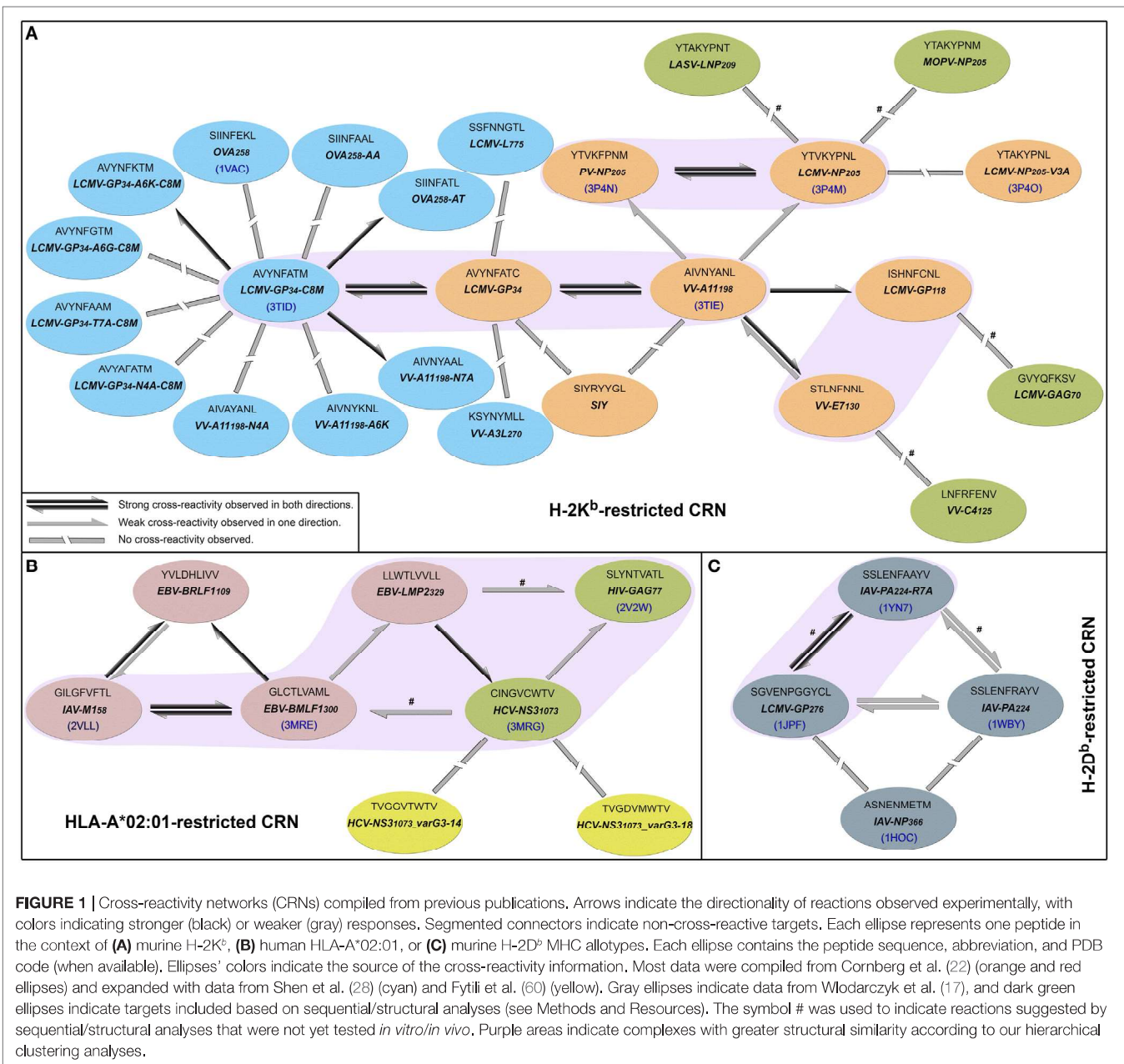
1.3. Structural Similarity of pMHC Complexes Can Reveal Their Likelihood for T-Cell Cross-reactivity

In 2010, Cornberg et al. (22) described *cross-reactivity networks* involving virus-derived peptides, within both human and murine memory T-cell pools (CD8⁺/CD44^{hi}). They used as a reference a peptide derived from vaccinia virus (VV), corresponding to a 9-mer sequence starting at position 198 of the A11 protein (hereafter denoted by VV-A11₁₉₈). Using this VV-derived peptide, which is displayed by the murine MHC H-2K^b, the authors were able to activate three different memory T-cell populations that also recognized peptides from lymphocytic choriomeningitis virus (LCMV-GP₃₄, LCMV-GP₁₁₈, and LCMV-NP₂₀₅). Therefore, VV-A11₁₉₈ could be seen as a cross-reactivity “hub,” connected to all these LCMV-derived peptides (Figure 1A). The concept of

cross-reactivity networks is interesting in highlighting how broad these T-cell cross-reactivities can be (25), sometimes involving completely unrelated targets. In this sense, graphical representations of such networks have been used in previous works (13, 50, 58, 59). However, it is extremely important to keep in mind that despite providing a nice way to visually summarize cross-reactivity relationships, the topology of these networks might not correspond to the cross-reactivities observed for a particular T-cell line. In other words, the “real” topology of the network in terms of T-cell activation depends on which T-cell is used to test these peptide-targets. In this study, we use cross-reactivity networks to summarize the information from previous studies, as a reference to analyze structural data and discuss cross-reactivity patterns (Figure 1). In our representation, each node describes a given peptide, and only peptides displayed by the same MHC are included in a given network (i.e., MHC-restricted network). Note that this is a schematic representation of the known relationships among peptides that are relevant to our discussion, and not a complete picture of known cross-reactivities; it is not expected to reflect the patterns observed in any particular T-cell assay. Additional information on all peptides included in our analysis can be found in Table S1 in Supplementary Material.

In their original study, Cornberg et al. (22) suggested that observed cross-reactivity patterns present a within-individual variation driven by private specificities and immunological history. For instance, the authors were able to collect VV-A11₁₉₈-specific T-cells from mice previously immunized with LCMV (i.e., LCMV-immune mice). Note that if the donor had no previous contact with VV-derived peptides, these VV-A11₁₉₈-specific T-cells should be cross-reactive cells primarily expanded *in vivo* by recognizing some LCMV-derived target. These cells were further expanded *in vitro* with the cognate (VV-A11₁₉₈) peptide and challenged with different peptides derived from LCMV, VV and pichinde virus (PV). Interestingly, these VV-A11₁₉₈-specific T-cells presented cross-reactivity with LCMV-GP₃₄, LCMV-GP₁₁₈, LCMV-NP₂₀₅, and PV-NP₂₀₅ (22) (Figure 2A). However, cross-reactivity against another VV-derived peptide (VV-E7₁₃₀) was not observed. On the other hand, a very different pattern was observed when the authors performed a similar experiment, but expanding VV-A11₁₉₈-specific T-cells from VV-immune mice instead of LCMV-immune mice (Figure 2B). In this case, cross-reactivity with VV-E7₁₃₀ and LCMV-GP₃₄ was observed, but no cross-reactivity was observed with LCMV-GP₁₁₈, LCMV-NP₂₀₅, and PV-NP₂₀₅. These contrasting results suggest the use of a different T-cell population with a different specificity (22). They also suggest a greater structural similarity between VV-A11₁₉₈ and LCMV-GP₃₄, since this cross-reactivity was observed for both LCMV-immune and VV-immune background. In fact, structural similarity between these targets was later confirmed by Shen et al. (28), which solved the crystal structures of VV-A11₁₉₈ and LCMV-GP₃₄-C8M bound to H-2K^b (PDB codes 3TIE and 3TID, respectively).

Out of the 25 pMHCs included in our H-2K^b-restricted network (Figure 1A), at the time of our analysis, only 6 had their structure determined by experimental methods. Using our previously described structure-based approach (46), we performed a hierarchical clustering of these 6 crystallographic structures



(Figure S2 in Supplementary Material). Supported by multiscale bootstrap resampling with the R package *pvclust* (61), the clustering agreed with experimental data. The cross-reactive targets VV-A11₁₉₈ and LCMV-GP₃₄ fall in the same cluster; the same is observed for the highly cross-reactive targets LCMV-NP₂₀₅ and PV-NP₂₀₅. These four targets are closer to one another than to the non-cross-reactive target OVA₂₅₈. Finally, the most different structure in this analysis contained the non-cross-reactive escape variant LCMV-NP₂₀₅-V3A (21, 62).

To expand our analysis, we used the pMHC modeling method implemented in DockTope (52, 57), obtaining the structures of other complexes previously tested by Cornberg et al. (22) (Figure 1A). We also included in this analysis two unrelated

peptides, VV-C4₁₂₅ and LCMV-GAG₇₀, as putative non-cross-reactive controls (Table S1 in Supplementary Material). Our expanded hierarchical clustering reflects the greater structural similarity between VV-A11₁₉₈ and LCMV-GP₃₄, since both complexes fall in the same cluster, with the edge presenting the lowest height and the highest *p*-values (Figure 3). Peptides LCMV-GP₁₁₈ and VV-E7₁₃₀, which are cross-reactive with VV-A11₁₉₈, fall in the next branch, followed by a cluster with other cross-reactive targets (LCMV-NP₂₀₅ and PV-NP₂₀₅). All these cross-reactive targets were grouped into a bigger cluster (see edge 5 in Figure 3), apart from all the non-cross-reactive targets. As discussed by Cornberg et al. (22), these cross-reactivities could not be easily predicted with peptide sequence

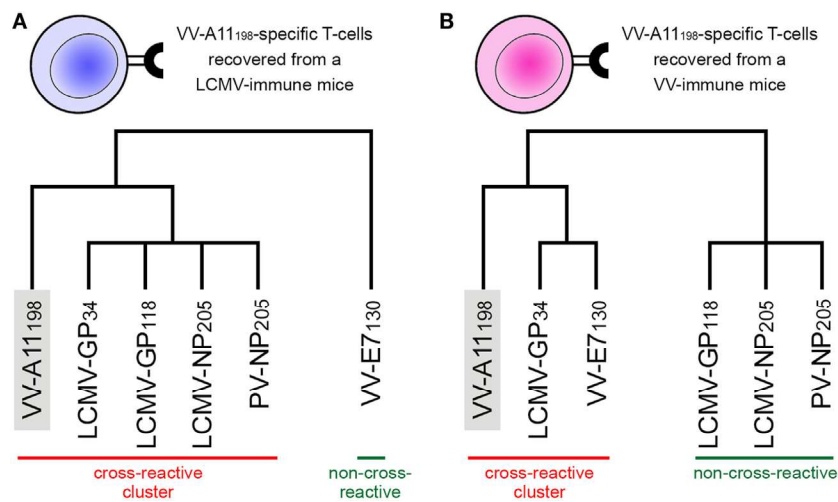


FIGURE 2 | Schematic representation of experimentally observed cross-reactivity patterns. Two alternative dendrograms were drawn to represent alternative outcomes observed in experiments previously performed by Cornberg et al. (22). **(A)** VV-A11₁₉₈-specific T-cells recovered from mice previously immunized with lymphocytic choriomeningitis virus (LCMV) recognize the cognate peptide (indicated by the gray box) as well as three other peptides derived from LCMV and one derived from pichinde virus (PV). We can represent these connections as a “cross-reactivity-cluster” in our dendrogram, as indicated in red. Another peptide derived from vaccinia virus (VV-E7₁₃₀), however, is not recognized. **(B)** VV-A11₁₉₈-specific T-cells recovered from mice previously immunized with vaccinia virus (VV) recognize the cognate peptide (gray box) as well as the other VV-derived peptide (VV-E7₁₃₀) and one LCMV-derived peptide (LCMV-GP₃₄). However, in this experiment, no cross-reactivity was observed against peptides LCMV-GP₁₁₈, LCMV-NP₂₀₅, and PV-NP₂₀₅ (indicated by the green bar). Although targeting the same VV-derived peptide, the alternative cross-reactivity patterns described in panels **(A,B)** reflect the use of different T-cell lines in each experiment (indicated as a blue or pink T-cell). Note that cross-reactivity between VV-A11₁₉₈ and LCMV-GP₃₄ was observed in both experiments, suggesting higher structural similarity of these peptides when displayed by H-2K^b. All peptides involved in these experiments are restricted to the murine MHC H-2K^b. This is a schematic representation, and the heights of the edges in the dendrogram do not capture the actual “distances” among the peptide-targets. Additional information on the presented peptides can be found in Table S1 in Supplementary Material.

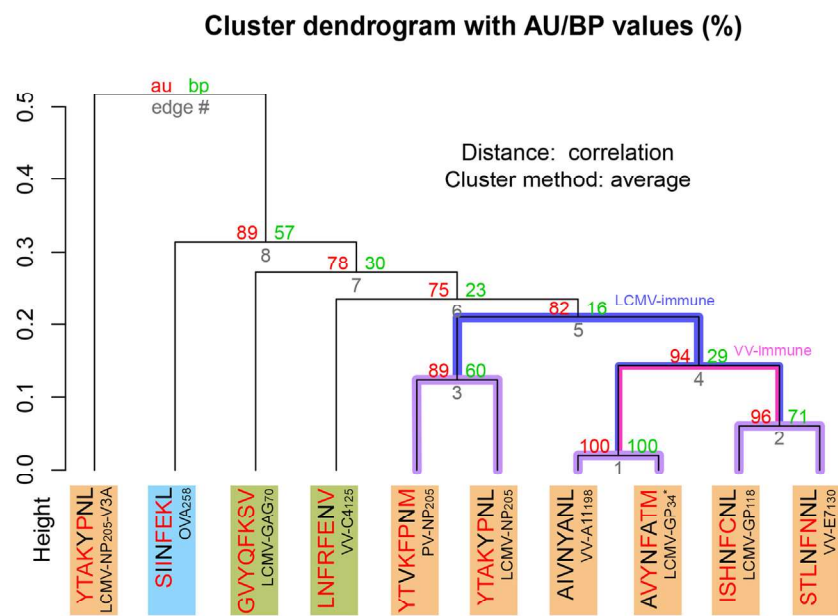


FIGURE 3 | Extended H-2K^b-restricted clustering. Structure-based hierarchical clustering performed with pvclust (61). Each putative cluster is represented by a specific edge (gray numbers), in order of increasing heights (y axis). Cluster confidence is measured with two p-values, approximately unbiased (AU), and bootstrap probabilities (BP). Lines highlighted in purple indicate structures with greater structural similarity (as represented in **Figure 1**). Lines highlighted in blue and pink indicate putative cross-reactivity thresholds for different memory T-cells (see **Figure 2**). Each peptide target is colored according to **Figure 1**. Peptide abbreviation and sequence are provided, with red amino acids indicating changes in relation to VV-A11₁₉₈. *Crystal structure 3TID was used to represent LCMV-GP₃₄, despite presenting a C8M exchange, as indicated by its sequence (see Methods and Resources).

similarity, since all these peptides share less than 50% of their amino acid residues. For instance, sequence similarity between VV-A11₁₉₈ and LCMV-GP₃₄ is only 37.5%, the same as between VV-A11₁₉₈ and the non-cross-reactive target OVA₂₅₈. In spite of that, our results show that this cluster of cross-reactivity involving peptides from three different viruses could be predicted by an *in silico* analysis of the corresponding pMHC structures (see edge 5 in **Figure 3**).

Cross-reactivity was indeed observed among these 6 peptides in the context of H-2K^b (22, 28). These pMHC complexes also present structural similarities, being clustered together in our structure-based hierarchical clustering. However, there was no experimental evidence of one T-cell population able to recognize all six peptides (22). As already discussed, cross-reactivity patterns depend on the specific T-cell population tested. Assuming our clustering correctly captures the relationships among these pMHCs, in terms of structural similarity, we can make some inferences about the T-cells used in the aforementioned experiments. We can say that T-cells from LCMV-immune mice are more cross-reactive, and we can visually represent them with a higher threshold in our clustering analysis (defining the blue cluster in **Figure 3**). Such threshold would correctly predict most of the observed cross-reactivities, with the exception of VV-E7₁₃₀ (which was not recognized). On the other hand, T-cells from VV-immune mice can be represented with a lower threshold (defining the pink cluster in **Figure 3**), since VV-A11₁₉₈-specific T-cells recognize neither NP₂₀₅ peptides. The exception in this case, would be LCMV-GP₁₁₈. These exceptions cannot be predicted considering the information provided by the pMHC structures, since they are most likely driven by TCR variability and private specificities. In spite of that, our data suggest a correlation between pMHC structural similarity and the probability to find cross-reactivity among pMHC targets; that is, the higher the similarity, the higher the likelihood of observing cross-reactive responses. Although cross-reactivity between LCMV-GP₁₁₈ and VV-E7₁₃₀ was not observed using the VV-A11₁₉₈-specific or VV-E7₁₃₀-specific T-cells (22), the similarity of these pMHC complexes (Figure S3 in Supplementary Material) suggests that this cross-reactivity should be observed using another T-cell population; maybe with LCMV-GP₁₁₈-specific T-cells.

1.4. Structural Features of the pMHC Can Shape the TCR Repertoire

More than a decade ago, Turner and colleagues (63) described differences in the T-cell population stimulated by a featureless peptide (referred to as a “vanilla” peptide), and a peptide having a prominent feature exposed to the TCR (hereafter referred to as a “spicy” peptide). The authors used a peptide derived from the polymerase acidic protein of influenza A virus as an example of spicy peptide (IAV-PA₂₂₄, see Table S1 in Supplementary Material). This peptide has an arginine at position 7 (P7), which becomes an exposed feature when displayed by the murine MHC molecule H-2D^b (Figure S4 in Supplementary Material). Immunization with this peptide triggered the expansion of a very diverse pool of T-cells, including cells with high affinity to the target pMHC.

Comparing the response across different animals, the authors noticed great variability in TCR usage. In other words, in each animal the response was dominated by TCRs with unique CDR sequences (i.e., shaped by private specificity).

Surprisingly, opposite results were observed when using a vanilla peptide. Immunization with a mutated version of IAV-PA₂₂₄, replacing the arginine at P7 with an alanine (IAV-PA₂₂₄-R7A), triggered the expansion of a much less diverse T-cell population. In this case, similar CDR sequences were observed for different individuals (i.e., public TCR usage). The same results were observed with a wild-type vanilla peptide (IAV-NP₃₆₆). Therefore, structural features of the pMHC complex can shape the composition of the TCR repertoire during a cellular immune response. A pMHC displaying a vanilla peptide has a TCR-interacting surface dominated by the (self) MHC; given the negative selection of T-cells, very few available TCRs can recognize this complex. This could explain the observation of a less diverse population and the use of public TCRs, sharing a germline bias to interact with the MHC. In addition, we could expect such TCRs to be more cross-reactive, since they rely mostly on (self) MHC features for the recognition. On the other hand, a spicy peptide offers a more evident discerning feature that various TCRs can recognize (in slightly different ways). Given their “focus” on this outstanding feature, we could expect such TCRs to be intrinsically less cross-reactive and they should be incapable (or impaired) to recognize pMHCs lacking such feature.

It is easier to understand this analogy of the spicy feature having in mind some prominent structure that is specific to the peptide, as the examples mentioned earlier and in the next section. However, the TCR/pMHC interaction can be influenced by more subtle features, as recently described by Song et al. (24). They performed a comprehensive evaluation of the T-cell response to the peptide IAV-M1₅₈, displayed by HLA-A*02:01, using the next-generation sequencing of TCRs. In addition, they resolved the crystal structures of two selected TCR/pMHC complexes. IAV-M1₅₈ has been described as a vanilla peptide, since most of its side chains are buried when displayed by HLA-A*02:01. In turn, it was suggested that the lack of recognizable peptide features would lead to a very narrow T-cell response (i.e., lack of TCR diversity among stimulated T-cells). However, Song et al. (24) observed that the IAV-M1₅₈:HLA-A*0201 complex can actually be recognized by a broad range of TCRs; most of them sharing the same V β domain. They were also able to identify a conserved structural feature that seemed to be required for the recognition of this peptide. Interestingly, it was not something “prominent,” and it was not exactly a feature of the peptide alone. In fact, the authors describe a unique exposed pocket between the peptide and the MHC, with which very different TCRs are able to interact. In other words, this particular pocket is a recognizable structural feature that is specific to the IAV-M1₅₈:HLA-A*0201 complex. As a result, in the context of our discussion, we can describe IAV-M1₅₈:HLA-A*0201 as a spicy complex. The lack of a prominent peptide feature might facilitate the selection of some public TCRs, as indeed observed experimentally (24). But the pMHC-specific pocket allows the selection of a broad TCR repertoire, in the same way as for spicy peptides. Once again, these findings highlight the fact that in most cases we

cannot discuss T-cell activation or T-cell cross-reactivity only in terms of peptide-targets, since the key features for recognition might come from the unique combined structure of the pMHC complex.

1.5. Local Structural Differences among pMHC Complexes Can Account for Limited Cross-reactivity and Lack of Reciprocity

In a recent study, Wlodarczyk et al. (17) described a weak cross-reactivity between IAV-PA₂₂₄:H-2D^b and a heterologous complex displaying a peptide derived from lymphocytic choriomeningitis virus (LCMV-GP₂₇₆:H-2D^b, see Table S1 in Supplementary Material). Since crystal structures are available for both complexes, we can visually compare their TCR-interacting surfaces (Figures 4A–C). Notably, LCMV-GP₂₇₆:H-2D^b differs from IAV-PA₂₂₄:H-2D^b by not having the featured arginine at P7. As expected, using our structure-based hierarchical clustering, we can see greater proximity (i.e., structural similarity) between LCMV-GP₂₇₆:H-2D^b and IAV-PA₂₂₄-R7A:H-2D^b, than between these complexes and the wild-type (IAV-PA₂₂₄:H-2D^b) or the non-cross-reactive complex IAV-NP₃₆₆:H-2D^b (Figure S5 in Supplementary Material). As described by Wlodarczyk et al. (17), cross-reactivity between GP₂₇₆:H-2D^b and IAV-PA₂₂₄:H-2D^b was weak and showed a preferential directionality. From the pool of T-cells recognizing GP₂₇₆:H-2D^b (primer) it was possible to extract T-cells that also recognize IAV-PA₂₂₄:H-2D^b (i.e., heterologous challenge). However, the reverse experiment was not successful.

Taken together, these results allow us to postulate that immunization with IAV-PA₂₂₄ stimulates a pool of T-cells dominated by clones with high specificity to the spicy feature (in this case, a peptide feature: the R at P7). By challenging with a heterologous peptide that lacks this prominent feature, we would most likely fail to find a T-cell clone that can also recognize the heterologous vanilla peptide-target (e.g., LCMV-GP₂₇₆). However, by using the vanilla peptide as a primer, we would start from a population of T-cells that is less diverse (i.e., dominated by public TCRs) but more cross-reactive. These TCRs are primarily engaging with (self) MHC structural features; some of these clones might also recognize the heterologous spicy peptide (IAV-PA₂₂₄), regardless of the prominent amino acid residue at P7. We believe this recognition might involve some adjustment of the CDR loops around the center of the peptide, as recently discussed by Adams et al. (29). Naturally, some TCRs will not be able to undergo such adjustment and will not show cross-reactivity. We also hypothesize that the “stronger” the spicy feature (or the combination of diverging features), the stronger the directionality and the lower the likelihood of cross-reactivity. Conversely, we believe stronger cross-reactivity should be observed between very similar pMHC complexes, regardless of directionality. For instance, stronger cross-reactivity should be observed between LCMV-GP₂₇₆:H-2D^b and the mutated IAV-PA₂₂₄-R7A:H-2D^b, than with the wild-type (Figure 1C).

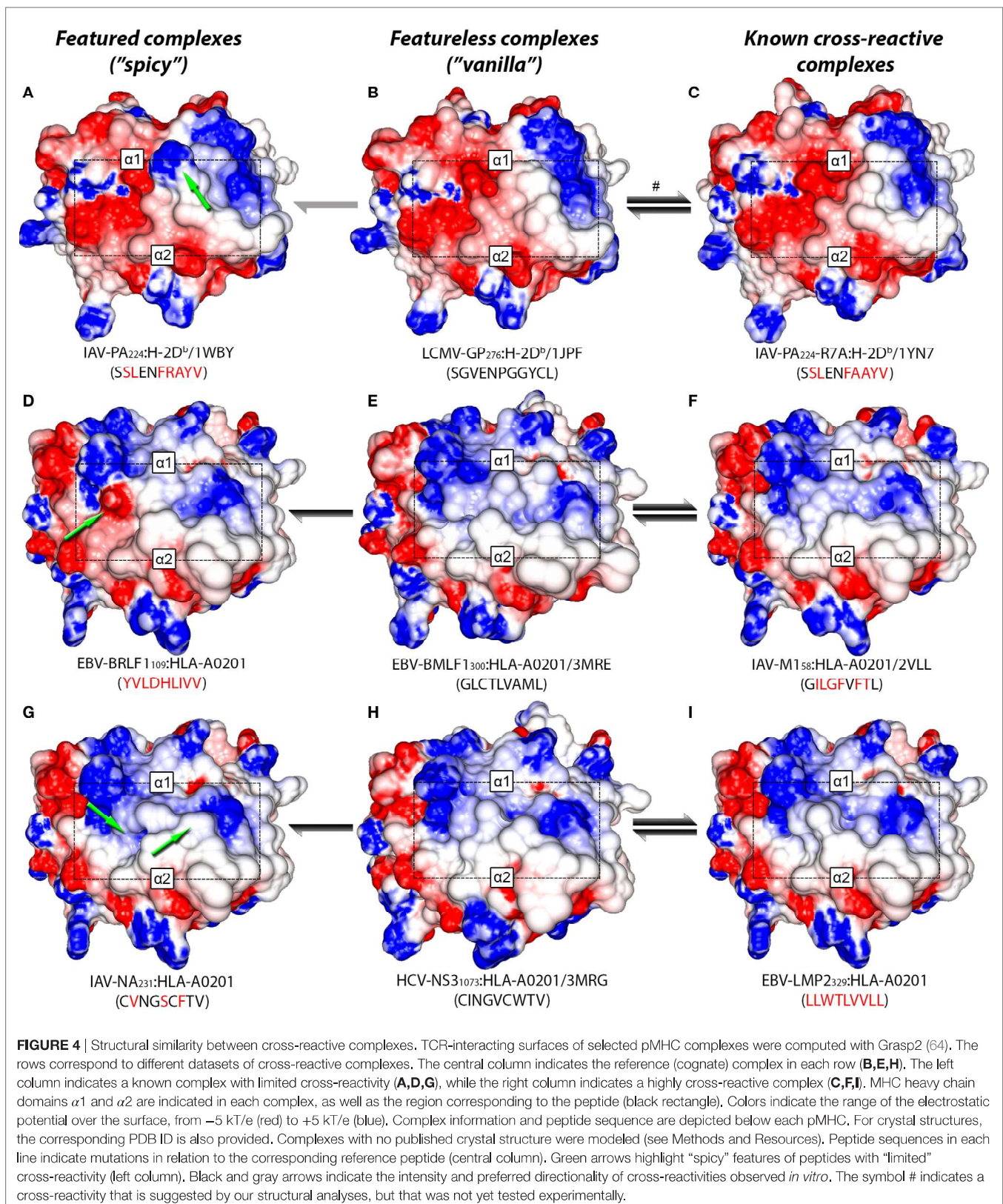
Additional examples supporting this theory can also be found in the context of human MHCs. By the time of our analysis, out of the 9 virus-derived peptides included in our

HLA-A*02:01-restricted network (Figure 1B), only 4 had available crystal structures. We modeled the remaining complexes and performed a hierarchical clustering (Figure S6 in Supplementary Material). As expected, the cross-reactive peptide-targets EBV-BMLF1₃₀₀, IAV-M1₅₈, HCV-NS3₁₀₇₃, HIV-GAG₇₇, and EBV-LMP2₃₂₉ were clustered together (see edge 5 in Figure S6 in Supplementary Material). These last two structures were actually the most similar pair of structures inside this cluster, in agreement with previous clustering results from our group (46).

Two non-cross-reactive variants of HCV-NS3₁₀₇₃ derived from HCV genotype 3, previously referred to as G3-14 and G3-18 (46, 60), fell in separate branches. Despite being the outermost branch of the main cluster (see edge 6 in Figure S6 in Supplementary Material), the small distance between G3-14 and the cross-reactive targets suggest that cross-reactivity with this HCV-derived escape variant might be observed depending on the T-cell population tested. Interestingly, the complex presenting EBV-BRLF1₁₀₉ falls in the same branch as G3-18, which is far from its cross-reactive target (EBV-BMLF1₃₀₀). This HCA result was due to a negatively charged spot in the surface of the EBV-BRLF1₁₀₉:HLA-A*0201 complex, which was not seen in its cross-reactive counterparts (Figures 4D–F). If we remove from our analysis this negatively charged spot, EBV-BRLF1₁₀₉ is clustered with EBV-BMLF1₃₀₀ (Figure 5). Note that we have had access to a yet unpublished crystal structure of EBV-BRLF1₁₀₉:HLA-A*0201, recently resolved by the team of Dr. Lawrence Stern (UMass Medical School, MA, USA), which confirms the existence of the outstanding negatively charged spot observed in our model (Song I, personal communication, June 2017).

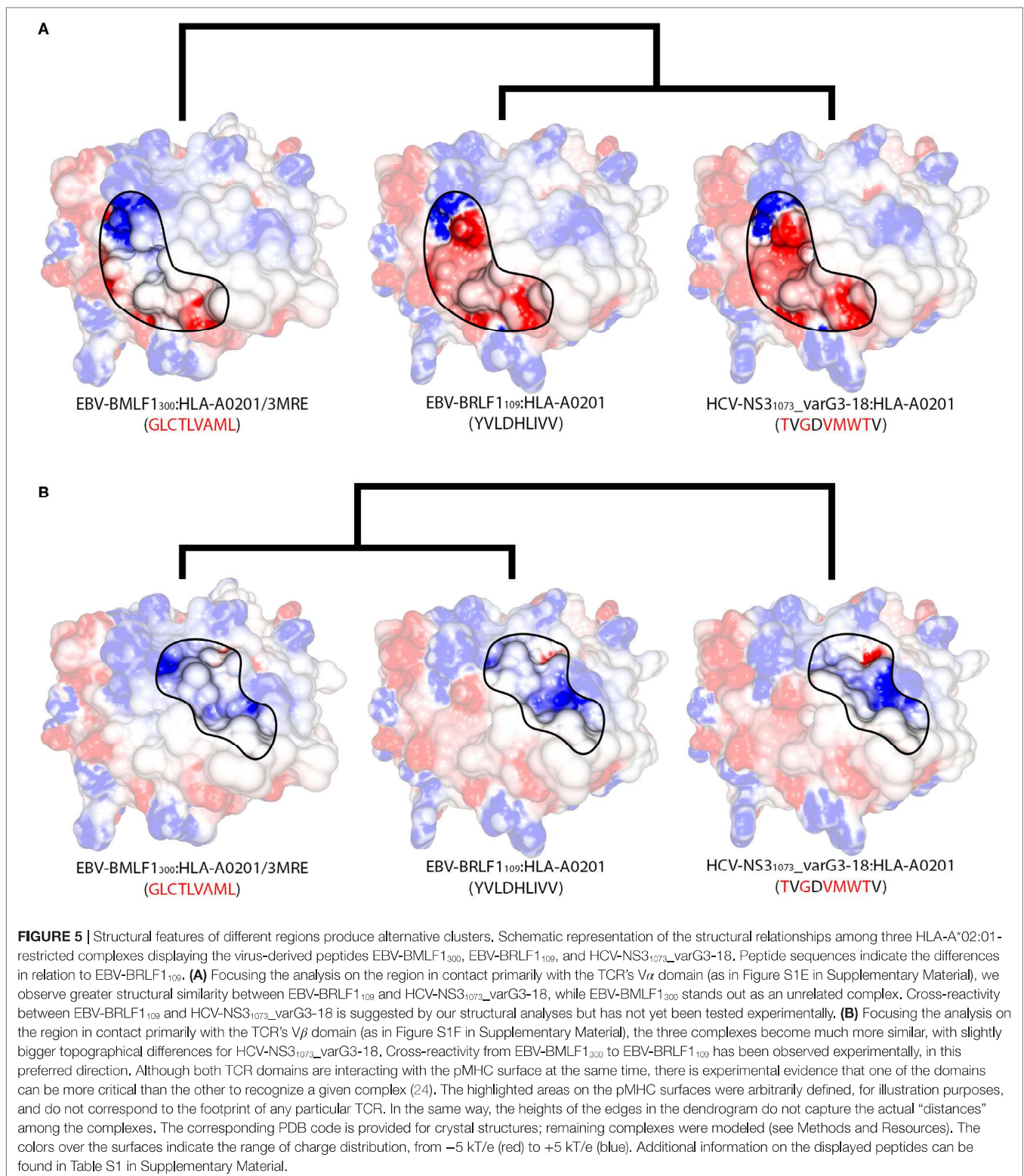
Similar to the situation described for IAV-PA₂₂₄, cross-reactivities involving EBV-BRLF1₁₀₉ feature several peculiarities. For instance, they are not observed for most T-cell populations and normally respect a given directionality, from EBV-BMLF1₃₀₀ to EBV-BRLF1₁₀₉ (22). EBV-BRLF1₁₀₉-specific T-cells recovered from EBV-immune individuals and expanded *in vitro* in the presence of the cognate peptide present higher affinity/avidity in TCR/pMHC interaction. Note that these cells are not cross-reactive with EBV-BMLF1₃₀₀. On the other hand, EBV-BMLF1₃₀₀-specific T-cells expanded *in vitro* in the presence of the cognate peptide might also recognize EBV-BRLF1₁₀₉ (22). Further expansion of this population with the heterologous peptide (i.e., EBV-BRLF1₁₀₉) produces (cross-reactive) EBV-BRLF1₁₀₉-specific T-cells with lower affinity/avidity in TCR/pMHC interaction (data not shown).

It is known that TCRs usually interact with pMHCs using a “canonical” binding mode (25, 51, 65, 66), but it was shown that a given TCR can preferentially use distinct amino acid residues to come in contact with different complexes (67) or even modify its CDR loops to accommodate different peptides (68). Therefore, it can be argued that immunization with a spicy peptide (such as IAV-PA₂₂₄ or EBV-BRLF1₁₀₉) will trigger a highly polyclonal T-cell response, with a broad spectrum of TCR specificities. Some of these are less specific to the homologous target, and more cross-reactive with other peptides, probably by establishing an interaction “focused” on surface regions that are shared among these targets (Figure 4). On the other hand, some of these cells present higher affinity/avidity with this homologous peptide, by establishing an



interaction "focused" on unique features of its surface (**Figure 5**). In turn, cross-reactivity between a spicy and a vanilla peptide depends on which T-cell populations are being tested.

We here hypothesize that, despite different TCRs can share a similar TCR footprint or even interact with the same pMHC amino acid residues, each TCR has a specific "interaction profile."



That is, some TCR/pMHC interactions are more important than others for triggering the T-cell response, and this interaction profile is specific to each TCR (Figures S1E,F in Supplementary Material). Knowing the specific hot-spots of a cognate pMHC,

i.e., the aforementioned "focus" of the TCR, would be key to predict cross-reactivity against heterologous pMHC targets. Moreover, although we tend to think of these hot-spots as pMHC amino acid residues, we need to expand this concept to

account for more subtle features of the TCR/pMHC interaction (e.g., pockets, hydrogen bonds, van der Waals contacts, and coordination of water molecules) (24).

1.6. T-Cell Cross-reactivity Can Be Triggered by High-Affinity Interactions with Specific Structural Features of the pMHC Complex

We previously suggested that T-cells expanded in response to a vanilla peptide should be intrinsically more cross-reactive, since they are focused on patterns shared across different pMHC complexes. Conversely, T-cells expanded in response to a spicy peptide are expected to be less cross-reactive in general, since most heterologous peptides would lack the spicy feature that is the focus of the response. However, these cells should still be cross-reactive with peptides having the spicy feature, in some cases regardless of other evident differences.

In fact, studies in cancer immunotherapy show that mutations leading to increased affinity of a given TCR-peptide interaction can actually increase cross-reactivity (38–40, 69). We hypothesize that although not changing the overall TCR footprint, such mutations can change the interaction profile of the TCR. In other words, the enhanced peptide-specific interaction becomes much more important for T-cell activation than the additional pMHC interactions, and any heterologous pMHC sharing the structural feature recognized by this enhanced TCR can become a cross-reactive target.

Further evidence for this hypothesis comes from a recent publication by Adams et al. (29). Using a carefully designed experimental approach, the authors investigated cross-reactive peptides showing limited sequence identity with the reference cognate peptide (restricted to H-2K^d). Despite apparent sequence diversity among peptides recognized by the probe TCR, closer analysis revealed a repeated focus on structurally and chemically similar elements of the peptides. For instance, the authors describe a preferred interaction with hydrophobic amino acid residues at P7; particularly phenylalanine. The authors refer to this amino acid residue as a peptide hot-spot for cross-reactivity, which in combination with some germline-mediated interactions greatly constrains the actual pool of potential cross-reactive pMHC targets (for the probe TCR). They also relate this description of the TCR/pMHC interaction with a more general feature of protein-protein interactions: a few energetically important contacts (usually in the center), surrounded by weaker and more diverse peripheral interactions. In the context of our discussion, we could see the phenylalanine at P7 as a spicy feature of the cognate peptide and the most important contact in the interaction profile of the probe TCR.

As mentioned earlier, we have previously described cross-reactivity between peptides with no sequence similarity, but with remarkably similar TCR-interacting surfaces (Figures 4G–I). The results described by Song et al. (24) provide an interesting example in which even greater variability can be anticipated. If the main feature for TCR recognition is a pocket defined by the peptide in the MHC cleft (e.g., IAV-M1₅₈:HLA-A*0201), we can expect that such “pocket-specific” T-cells will be cross-reactive

to other pMHC complexes having a similar pocket, maybe regardless of other differences in the TCR-interacting surface. For instance, it is possible for a completely unrelated pMHC (with a different peptide sequence and/or MHC allotype) to have a very similar pocket and, therefore, be a cross-reactive target for IAV-M1₅₈-specific T-cells.

As also discussed by Adams et al. (29), the implications of such “hot-spots” for cross-reactivity prediction are clear. A superficial look at the sequence diversity of cross-reactive peptides might suggest a completely promiscuous recognition, even considering a single TCR. The picture becomes even more complex if on top of that we start considering different pools of T-cells or the *in vivo* response of different individuals, which adds variability given to private specificity and immunological history. This complex picture helps understand the challenge of comparing results from different studies and drawing general conclusions about T-cell cross-reactivity. On the other hand, the characterization of cross-reactivity hot-spots and TCR-specific interaction profiles should allow us to focus our research and make progress for meaningful cross-reactivity predictions.

In fact, Arber et al. (56) published a study that goes in this very direction. They combined T-cell assays and computational analysis to evaluate T-cell cross-reactivity of different clones in the context of cancer immunotherapy. Based on IFN- γ production against a panel of alanine-exchanged variants of the cognate peptide, they defined T-cell-specific sequence motifs. These motifs were meant to capture T-cell-specific cross-reactivity hot-spots; they were later used for a sequence-based screening of potential cross-reactive targets in the human proteome. A number of positive hits were selected and tested experimentally, confirming that one T-cell line was much safer (i.e., less cross-reactive) than the other. The scope of this screening was still limited, not accounting for structural information of the pMHC or other potentially relevant features (14, 24, 70). Nevertheless, it provides us with an example of the type of framework that would be required for T-cell-specific prediction of potential cross-reactive targets.

1.7. Conclusions and Implications for Cancer Immunotherapy

Several immunotherapy trials are currently underway in a number of different tumor types to target tumor-associated peptides (71), including the melanoma-associated antigens MAGE-A3 and MART-1. These tumor antigens are expressed by multiple tumor types (39) but are not expressed by most normal tissues. Since MART-1 is highly expressed in both melanoma and normal melanocytes, MART-1 TCR-based therapies have led to antitumor responses concurrent with vitiligo and melanocyte destruction in the eye and inner ear, side effects that could be relieved with steroid administration (72). However, more severe safety issues with other TCR-based therapies have raised major concerns about this approach (33, 73, 74). As mentioned earlier, fatal adverse events were reported following adoptive transfer of TCR-transduced T-cells targeting complexes displaying the MAGE-A3 peptide (39–42). In two of these patients, unexpectedly severe cardiac toxicity was attributed to recognition of a completely unrelated peptide. This heterologous peptide-target was derived from the

self protein Titin and displayed by HLA-A*01:01 at the surface of healthy cardiac cells (43). As discussed by Stone et al. (38), T-cell cross-reactivity becomes specially relevant in the context of affinity-enhanced TCRs. Approaches like this are becoming more popular through the use of chimeric antigen receptors (CARs) (71). However, as reported by van den Berg et al. (41), severe off-target reactions can occur even without TCR-affinity enhancement. And this adds a layer of concern on top of toxicity and autoimmunity that might occur even with the use of autologous tumor infiltrating T-cells (72, 75). Moreover, as highlighted in our review, T-cell cross-reactivity seems to be rather the rule than the exception. Therefore, despite all mechanisms of central and peripheral tolerance (76), off-target toxicity mediated by T-cell cross-reactivity must be a concern in any TCR-based immunotherapy. However, the risk for off-target toxicity will differ depending on which specific form of therapy is being used.

Our study corroborates the idea that structural similarity among pMHC complexes is one of the main features driving the likelihood of cross-reactive T-cell responses. Cross-reactivity is very likely to be observed between two structurally identical complexes, for most T-cell lines recognizing one of the complexes, and in both directions. On the other hand, finding a T-cell line capable of recognizing two completely different pMHC complexes is highly unlikely. However, in most cases, two complexes will have common features but also different ones. In this situation, cross-reactivity can only be assessed by the level of pMHC structural similarity, as an intrinsic likelihood. However, its occurrence, intensity and directionality will be driven by the specific T-cell population stimulated by the first target and selectively expanded after heterologous challenges.

In the context of polyclonal T-cell populations, this outcome is mostly a consequence of private specificity and immunological history (19, 20, 32). Therefore, predicting patient cross-reactivity in response to immunization, infection or tissue transplantation is very challenging. Even knowing the peptide-targets and the MHC alleles of the patient, and having the perfect tools to estimate intrinsic cross-reactivity probabilities, we would still lack information on the available T-cell repertoire and the interaction profile of the dominating T-cell line. On the other hand, some problems in cancer immunotherapy offer a much more constrained scenario. In the context of TCR-based immunotherapies, researchers know which TCR is being used to recognize the tumor-derived peptide-target and can ensure that this will be the dominating population during treatment. By narrowing our analysis to a particular therapeutic T-cell line, we can limit the scope of cross-reactivity to structural features of the targeted pMHC; more specifically, to hot-spots that are the focus of the therapeutic TCR.

Therefore, we advocate that an important goal of structural analyses in the field of immunotherapy should be the characterization of the TCR-specific recognition profile. This profile should be a refinement of a more general TCR footprint, highlighting which pMHC structural features are more important for triggering this particular T-cell. In turn, this information can be used to guide large-scale *in silico* screenings, based on a combination of structural and sequential information. Currently, no tool can perform such screenings in a personalized fashion, especially

when considering the diversity of MHC alleles in the human population (5). However, T-cell cross-reactivity prediction will soon be enabled by advances in both pMHC structural modeling and TCR sequence analyses.

On the pMHC side, the combination of new modeling methods (57, 77) and structural clustering approaches (48, 78, 79) will allow considering structural information for larger datasets, regardless of whether experimental data are available. On the TCR side, recent reports have shown exciting results in the identification of conserved CDR motifs that can be directly linked to TCR specificity (11, 80). In time, we should be able to define T-cell-specific interaction profiles based on the sequence of the CDR regions of the TCR of interest.

Finally, better understanding of all subtle structural features relevant to TCR/pMHC engagement (11, 24) and their contributions to TCR binding affinity (37, 81–83) will also facilitate efforts toward TCR engineering and rational design (37, 84–86). The TCR-specific interaction profile can inform computer-aided efforts to increase TCR affinity to tumor-specific peptides, while reducing the risk for off-target toxicity. Hopefully, the combined use of these new technologies will soon allow researchers to predict and validate potentially dangerous cross-reactivities in the early stages of therapy development, guiding additional procedures to achieve safer TCR-based immunotherapies. Despite the overall complexity of the subject, urgent needs in cancer immunotherapy are pushing the discussion forward and should pave the way for many additional contributions to other areas of human health.

2. METHODS AND RESOURCES

2.1. Experimental Data on Cross-reactivity Networks

Cross-reactivity networks depicted in **Figure 1** were compiled from previously published experiments. Most data were made available by Cornberg et al. (22), who first presented these networks. The authors also described an escape variant of LCMV-NP₂₀₅ with a V3A substitution (21), suggested its sequence similarity with peptides from old world arenaviruses (MOPV-NP₂₀₅ and LASV-LNP₂₀₉) and finally solved its 3D crystal structure in the context of H-2K^b (62). This study with murine cross-reactivities was further explored by Shen et al. (28). The murine H-2D^b-restricted network was depicted with data from Włodarczyk et al. (17).

Cornberg et al. (22) also described a human HLA-A*02:01-restricted network. We expanded this network by including a cross-reactive target prospected through structural *in silico* analysis (46) and already confirmed experimentally (47), as well as two non-cross-reactive targets described by Fytily et al. (60). These tested non-cross-reactive targets were included both in human and murine cross-reactivity networks to provide further experimental information to guide our structure-based analysis.

A careful verification of peptides' information was performed to determine the correct protein name and peptide position, providing an updated reference for future studies (Table S1 in Supplementary Material). Curated information from Uniprot (87) was used as the main reference, and GenBank (88) was also

consulted. References to the Immune Epitope Database (IEDB) (89), the Protein Data Bank (PDB) (90), and the CrossTope Database (91) were also provided, when available.

2.2. Crystal Structures

Crystal structures were obtained from the Protein Data Bank (PDB) (90) and revised as needed using the PyMOL Viewer (92). The resulting pMHC structure was submitted to a short energy minimization with the Gromacs 4.5.1 package (93).

Note that 3TID is referred to as the crystal structure of LCMV-GP₃₄:H-2K^b complex, despite presenting an amino acid exchange at P8 (LCMV-GP₃₄-C8M). According to the authors who described the structure (28), this exchange has no significant impact on TCR/pMHC interactions and this C8M variant was used in previous studies as an “equivalent” to the wild-type sequence. Here, sequence divergence between LCMV-GP₃₄ and LCMV-GP₃₄-C8M is indicated in **Figure 3**, but 3TID was considered as the crystal structure of LCMV-GP₃₄ for all structure-based analyses.

2.3. Modeled Structures

Peptide–MHC complexes without published crystal structures were predicted using the DockTope webserver (57). Briefly, a reference crystal structure of the MHC allotype of interest (without its ligand) was used as a receptor (“MHC_donor”) for a molecular docking with Autodock Vina 1.1.2 (94). The input ligand structure was produced by mutating a peptide structure obtained in the context of the same MHC allotype (“Peptide_pattern”). The resulting pMHC structure was then refined through a full atom energy minimization step with the Gromacs 4.5.1 package (93). A new docking search was performed with only the peptide side chains being flexible. This automated approach for pMHC structure prediction was largely validated against available crystal structures (57).

2.4. Electrostatic Potential Calculation and Image Analysis

Electrostatic potential over the TCR-interacting surface of pMHCs (for both crystals and models) was calculated using Delphi (95), through the molecular viewer software GRASP2 (64). Automated scripts were used to prepare the structures for this analysis, allowing all pMHCs to be observed in the same fixed orientation. Images of the TCR-interacting surfaces were saved and imported to the ImageJ 1.46r software (National Institute of Health, USA, <https://imagej.nih.gov/ij/>). Using preexisting classes from ImageJ, our team adapted a plugin to import RGB values from predetermined regions over the pMHC surface (as in Figure S1D in Supplementary Material), following a previously described protocol (46, 47). Values were exported as “csv” tables and used as input for hierarchical cluster analysis.

2.5. Hierarchical Cluster Analysis

In this study, hierarchical clustering was used as a tool to assess structure-based similarity among pMHC complexes. Input values were extracted from the images of the TCR-interacting surfaces (see section 2.4). Hierarchical clustering was performed

with pvclust (61), an R package for assessing the uncertainty in hierarchical clustering. The “average” linkage method was used with “correlation” distance, and the number of bootstrap replications was set to 10,000. Results were plotted as dendrograms with approximately unbiased (AU) and bootstrap probabilities (BP) p-values. BP values are calculated by normal bootstrap resampling, and AU values are computed through multiscale bootstrap resampling, which is considered a better approximation to unbiased p-value (61). SEs for AU p-values were obtained with seplot, presenting values lower than 0.01 for all clusterings performed.

AUTHOR CONTRIBUTIONS

DA, GV, MC, and LS suggested the initial idea behind this work. DA, MR, MS, and GV conceived the experiments. DA selected the dataset and MS curated the information on selected peptides. DA, MR, and MM conducted the modeling and clustering experiments. MF adapted the ImageJ plugin and helped with the extraction of the values for clustering. LK revised clustering experiments and algorithmic choices. DA, MR, GV, MS, MC, and LS analyzed and interpreted the results. GL contributed with the applications to immunotherapy and the review of related literature. DA wrote the manuscript. All the authors reviewed and approved the final manuscript.

ACKNOWLEDGMENTS

The authors thank the *Centro Nacional de Supercomputação* (CESUP/UFRGS) for allowing access to its computational resources. The authors also thank Inyoung Song and Dr. Lawrence Stern, from the University of Massachusetts Medical School (Worcester, MA, USA), for sharing the crystal structure of EBV-BRLF1₁₀₉:HLA-A*0201 before publication. Finally, the authors thank Dr. Didier Devaux for his helpful comments on the final manuscript.

FUNDING

This work has been supported in part by *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq/Brazil) and *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* (CAPES/Brazil). This work was also partially supported by the Cancer Prevention & Research Institute of Texas (CPRIT), under award number RP170508.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at <http://journal.frontiersin.org/article/10.3389/fimmu.2017.01210/full#supplementary-material>.

TABLE S1 | List of all studied complexes. Identification and source information for each peptide and MHC, as well as corresponding access codes to relevant databases.

FIGURE S1 | The TCR-interacting surface and the proposed TCR interaction profiles. **(A)** Top view of a pMHC complex depicting the MHC-receptor as *cartoon* (gray) and the peptide–ligand as *sticks* (pink). **(B)** Top view of the same

pMHC complex, depicting the exposed surface of the MHC (gray) and the exposed surface of the peptide (pink). **(C)** The combined surface of the pMHC complex, with the colors indicating the range of charge distribution over the surface from -5 kT/e (red) to $+5$ kT/e (blue). This is the “face” of the pMHC exposed for TCR recognition, referred to as the *TCR-interacting surface*. **(D)** The TCR binds to the pMHC in a conserved orientation: the TCR’s variant domain $V\alpha$ primarily interacts with the N-terminal portion of the peptide, while the $V\beta$ domain primarily interacts with the C-terminal portion of the peptide. This area of TCR/pMHC interaction, in a particular docking angle, is referred to as the *TCR footprint*. **(E)** Schematic representation of a *TCR-specific interaction profile* over the pMHC surface. Colored boxes indicate “hot-spots” for cross-reactivity (green) and secondary contacts that also contribute to TCR binding affinity (yellow). **(F)** Schematic representation of a different interaction profile, displayed by a different TCR that still shares the same general TCR footprint. Both depicted profiles are simplified schematic representations and do not represent known interactions of a any particular TCR.

FIGURE S2 | Crystal-based H-2K^b-restricted clustering. Structure-based hierarchical clustering performed with pvclust (61). Each putative cluster is represented by a specific edge (gray numbers), in order of increasing heights (y axis). Cluster confidence is measured with two p-values, approximately unbiased (AU) and bootstrap probabilities (BP). Lines highlighted in purple indicate structures with greater structural similarity (as represented in **Figure 1**). Peptide abbreviation and corresponding PDB code for each crystal structure (in blue) are provided. *Crystal structure 3TID was used to represent LCMV-GP₃₄, despite presenting a C8M exchange (see Methods and Resources).

FIGURE S3 | TCR-interacting surfaces of predicted cross-reactive targets. Regions with positive (blue) and negative (red) charges are represented with a scale from -5 to $+5$ kT/e. Information on the corresponding peptide and MHC restriction is provided below each complex. Amino acid exchanges in relation to LCMV-GP₁₁₈ are indicated. Great structural similarity is observed between these

two complexes, both in terms of topography and electrostatic potential over the TCR-interacting surface. Note that other subtle structural differences might exist but are not well captured by this representation of the complexes.

FIGURE S4 | Specific interaction with a prominent peptide amino acid. **(A)** Surface of the IAV-PA₂₂₄:H-2D^b complex (spicy peptide) according to a crystal structure obtained in the absence of the TCR (PDB code 1WBY). **(B)** Surface of the same complex according to a crystal structure obtained in the presence of the TCR (PDB code 3PQY). **(C)** Cartoon depiction of the 3PQY structure highlighting TCR amino acid residues that interact directly with a prominent arginine at the peptide (R7), forming a negatively charged cavity. Side chain of amino acid R7 is depicted in *ball and stick*. TCR and MHC domains are indicated, and green arrows highlight the location of amino acid residue R7. Electrostatic potentials were computed with Grasp2 (64).

FIGURE S5 | Crystal-based H-2D^b-restricted clustering. Structure-based hierarchical clustering performed with pvclust. Each putative cluster is represented by a specific edge (gray numbers), in order of increasing heights (y axis). Cluster confidence is measured with two p-values, approximately unbiased (AU) and bootstrap probabilities (BP). Peptide abbreviation and the respective PDB code for each crystal structure (in blue) are provided. Lines highlighted in purple indicate structures with greater structural similarity (as represented in **Figure 1**).

FIGURE S6 | Extended HLA-A*02:01-restricted clustering. Structure-based hierarchical clustering performed with pvclust. Each putative cluster is represented by a specific edge (gray numbers), in order of increasing heights (y axis). Cluster confidence is measured with two p-values, approximately unbiased (AU) and bootstrap probabilities (BP). Abbreviation of crystal structures includes their PDB code (in blue), while “Mod” indicates modeled structures. Lines highlighted in purple indicate structures with greater structural similarity (as represented in **Figure 1**).

REFERENCES

- Dzik JM. The ancestry and cumulative evolution of immune reactions. *Acta Biochim Pol* (2010) 57(4):443–66.
- Hirano M. Evolution of vertebrate adaptive immunity: immune cells and tissues, and AID/APOBEC cytidine deaminases. *Bioessays* (2015) 37(8):877–87. doi:10.1002/bies.201400178
- Welsh RM, Selin LK, Szomlanyi-Tsuda E. Immunological memory to viral infections. *Annu Rev Immunol* (2004) 22:711–43. doi:10.1146/annurev.immunol.22.012703.104527
- Lauring AS, Frydman J, Andino R. The role of mutational robustness in RNA virus evolution. *Nat Rev Microbiol* (2013) 11(5):327–36. doi:10.1038/nrmicro3003
- Vandiedonck C, Knight JC. The human Major Histocompatibility Complex as a paradigm in genomics research. *Brief Funct Genomic Proteomic* (2009) 8(5):379–94. doi:10.1093/bfpp/elp010
- Paterson S, Vogwill T, Buckling A, Benmayor R, Spiers AJ, Thomson NR, et al. Antagonistic coevolution accelerates molecular evolution. *Nature* (2010) 464(7286):275–8. doi:10.1038/nature08798
- Kubinak JL, Ruff JS, Hyzer CW, Slev PR, Potts WK. Experimental viral evolution to specific host MHC genotypes reveals fitness and virulence trade-offs in alternative MHC types. *Proc Natl Acad Sci U S A* (2012) 109(9):3422–7. doi:10.1073/pnas.1112633109
- Sohn SJ, Thompson J, Winoto A. Apoptosis during negative selection of autoreactive thymocytes. *Curr Opin Immunol* (2007) 19(5):510–5. doi:10.1016/j.coi.2007.06.001
- Elhanati Y, Murugan A, Callan CG, Mora T, Walczak AM. Quantifying selection in immune receptor repertoires. *Proc Natl Acad Sci U S A* (2014) 111(27):9875–80. doi:10.1073/pnas.1409572111
- Zarnitsyna VI, Evavold BD, Schoettle LN, Blattman JN, Antia R. Estimating the diversity, completeness, and cross-reactivity of the T cell repertoire. *Front Immunol* (2013) 4:485. doi:10.3389/fimmu.2013.00485
- Dash P, Fiore-Gartland AJ, Hertz T, Wang GC, Sharma S, Souquette A, et al. Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature* (2017) 547(7661):89–93. doi:10.1038/nature22383
- Welsh RM, Selin LK. No one is naive: the significance of heterologous T-cell immunity. *Nat Rev Immunol* (2002) 2(6):417–26. doi:10.1038/nri820
- Welsh RM, Che JW, Brehm MA, Selin LK. Heterologous immunity between viruses. *Immunol Rev* (2010) 235(1):244–66. doi:10.1111/j.0105-2896.2010.00897.x
- Vieira GF, Chies JA. Immunodominant viral peptides as determinants of cross-reactivity in the immune system – can we develop wide spectrum viral vaccines? *Med Hypotheses* (2005) 65(5):873–9. doi:10.1016/j.mehy.2005.05.041
- Welsh RM, Fujinami RS. Pathogenic epitopes, heterologous immunity and vaccine design. *Nat Rev Microbiol* (2007) 5(7):555–63. doi:10.1038/nrmicro1709
- Selin LK, Cornberg M, Brehm MA, Kim SK, Calcagno C, Ghersi D, et al. CD8 memory T cells: cross-reactivity and heterologous immunity. *Semin Immunol* (2004) 16(5):335–47. doi:10.1016/j.smim.2004.08.014
- Wlodarczyk MF, Kraft AR, Chen HD, Kenney LL, Selin LK. Anti-IFN- γ and peptide-tolerization therapies inhibit acute lung injury induced by cross-reactive Influenza A-specific memory T cells. *J Immunol* (2013) 190(6):2736–46. doi:10.4049/jimmunol.1201936
- Cornberg M, Kenney LL, Chen AT, Waggoner SN, Kim SK, Dienes HP, et al. Clonal exhaustion as a mechanism to protect against severe immunopathology and death from an overwhelming CD8 T cell response. *Front Immunol* (2013) 4:475. doi:10.3389/fimmu.2013.00475
- Cornberg M, Wedemeyer H. Hepatitis C virus infection from the perspective of heterologous immunity. *Curr Opin Virol* (2016) 16:41–8. doi:10.1016/j.coviro.2016.01.005
- Gil A, Yassai MB, Naumov YN, Selin LK. Narrowing of human Influenza A virus-specific T cell receptor α and β repertoires with increasing age. *J Virol* (2015) 89(8):4102–16. doi:10.1128/JVI.03020-14
- Cornberg M, Chen AT, Wilkinson LA, Brehm MA, Kim SK, Calcagno C, et al. Narrowed TCR repertoire and viral escape as a consequence of heterologous immunity. *J Clin Invest* (2006) 116(5):1443–56. doi:10.1172/JCI27804
- Cornberg M, Clute SC, Watkin LB, Saccoccio FM, Kim SK, Naumov YN, et al. CD8 T cell cross-reactivity networks mediate heterologous immunity in human EBV and murine vaccinia virus infections. *J Immunol* (2010) 184(6):2825–38. doi:10.4049/jimmunol.0902168

23. Gil A, Kenney LL, Mishra R, Watkin LB, Aslan N, Selin LK. Vaccination and heterologous immunity: educating the immune system. *Trans R Soc Trop Med Hyg* (2015) 109(1):62–9. doi:10.1093/trstmh/tru198
24. Song I, Gil A, Mishra R, Ghersi D, Selin LK, Stern LJ. Broad TCR repertoire and diverse structural solutions for recognition of an immunodominant CD8(+) T cell epitope. *Nat Struct Mol Biol* (2017) 24(4):395–406. doi:10.1038/nsmb.3383
25. Birnbaum ME, Mendoza JL, Sethi DK, Dong S, Glanville J, Dobbins J, et al. Deconstructing the peptide-MHC specificity of T cell recognition. *Cell* (2014) 157(5):1073–87. doi:10.1016/j.cell.2014.03.047
26. Degauque N, Brouard S, Souillou JP. Cross-reactivity of TCR repertoire: current concepts, challenges, and implication for allotransplantation. *Front Immunol* (2016) 7:89. doi:10.3389/fimmu.2016.00089
27. Yin Y, Li Y, Mariuzza RA. Structural basis for self-recognition by autoimmune T-cell receptors. *Immunol Rev* (2012) 250(1):32–48. doi:10.1111/imr.12002
28. Shen ZT, Nguyen TT, Daniels KA, Welsh RM, Stern LJ. Disparate epitopes mediating protective heterologous immunity to unrelated viruses share peptide-MHC structural features recognized by cross-reactive T cells. *J Immunol* (2013) 191(10):5139–52. doi:10.4049/jimmunol.1300852
29. Adams JJ, Narayanan S, Birnbaum ME, Sidhu SS, Blevins SJ, Gee MH, et al. Structural interplay between germline interactions and adaptive recognition determines the bandwidth of TCR-peptide-MHC cross-reactivity. *Nat Immunol* (2016) 17(1):87–94. doi:10.1038/ni.3310
30. Wedemeyer H, Mizukoshi E, Davis AR, Bennink JR, Rehmann B. Cross-reactivity between hepatitis C virus and Influenza A virus determinant-specific cytotoxic T cells. *J Virol* (2001) 75(23):11392–400. doi:10.1128/JVI.75.23.11392-11400.2001
31. Kasprovicz V, Ward SM, Turner A, Grammatikos A, Nolan BE, Lewis-Ximenez L, et al. Defining the directionality and quality of influenza virus-specific CD8+ T cell cross-reactivity in individuals infected with hepatitis C virus. *J Clin Invest* (2008) 118(3):1143–53. doi:10.1172/JCI33082
32. Che JW, Selin LK, Welsh RM. Evaluation of non-reciprocal heterologous immunity between unrelated viruses. *Virology* (2015) 482:89–97. doi:10.1016/j.virol.2015.03.002
33. Lizée G, Overwijk WW, Radvanyi L, Gao J, Sharma P, Hwu P. Harnessing the power of the immune system to target cancer. *Annu Rev Med* (2013) 64:71–90. doi:10.1146/annurev-med-112311-083918
34. Yee C, Lizée G, Schueneman AJ. Endogenous T-cell therapy: clinical experience. *Cancer J* (2015) 21(6):492–500. doi:10.1097/PPO.0000000000000158
35. Rouce RH, Sharma S, Huynh M, Heslop HE. Recent advances in T-cell immunotherapy for haematological malignancies. *Br J Haematol* (2017) 176(5):688–704. doi:10.1111/bjh.14470
36. Menon S, Shin S, Dy G. Advances in cancer immunotherapy in solid tumors. *Cancers (Basel)* (2016) 8(12):1–21. doi:10.3390/cancers8120106
37. Pierce BG, Hellman LM, Hossain M, Singh NK, Vander Kooi CW, Weng Z, et al. Computational design of the affinity and specificity of a therapeutic T cell receptor. *PLoS Comput Biol* (2014) 10(2):e1003478. doi:10.1371/journal.pcbi.1003478
38. Stone JD, Harris DT, Kranz DM. TCR affinity for p/MHC formed by tumor antigens that are self-proteins: impact on efficacy and toxicity. *Curr Opin Immunol* (2015) 33:16–22. doi:10.1016/j.coi.2015.01.003
39. Linette GP, Stadtmauer EA, Maus MV, Rapoport AP, Levine BL, Emery L, et al. Cardiovascular toxicity and titin cross-reactivity of affinity-enhanced T cells in myeloma and melanoma. *Blood* (2013) 122(6):863–71. doi:10.1182/blood-2013-03-490565
40. Morgan RA, Chinnsamy N, Abate-Daga D, Gros A, Robbins PF, Zheng Z, et al. Cancer regression and neurological toxicity following anti-MAGE-A3 TCR gene therapy. *J Immunother* (2013) 36(2):133–51. doi:10.1097/CJI.0b013e3182829903
41. van den Berg JH, Gomez-Eerland R, van de Wiel B, Hulshoff L, van den Broek D, Bins A, et al. Case report of a fatal serious adverse event upon administration of T cells transduced with a MART-1-specific T-cell receptor. *Mol Ther* (2015) 23(9):1541–50. doi:10.1038/mt.2015.60
42. Cameron BJ, Gerry AB, Dukes J, Harper JV, Kannan V, Bianchi FC, et al. Identification of a Titin-derived HLA-A1-presented peptide as a cross-reactive target for engineered MAGE A3-directed T cells. *Sci Transl Med* (2013) 5(197):197ra103. doi:10.1126/scitranslmed.3006034
43. Raman MC, Rizkallah PJ, Simmons R, Donnellan Z, Dukes J, Bossi G, et al. Direct molecular mimicry enables off-target cardiovascular toxicity by an enhanced affinity TCR designed for cancer immunotherapy. *Sci Rep* (2016) 6:18851. doi:10.1038/srep18851
44. Haase K, Raffegerst S, Schendel DJ, Frishman D. Expitope: a web server for epitope expression. *Bioinformatics* (2015) 31(11):1854–6. doi:10.1093/bioinformatics/btv068
45. Jaravine V, Raffegerst S, Schendel DJ, Frishman D. Assessment of cancer and virus antigens for cross-reactivity in human tissues. *Bioinformatics* (2017) 33(1):104–11. doi:10.1093/bioinformatics/btw567
46. Antunes DA, Rigo MM, Silva JP, Cibulski SP, Sinigaglia M, Chies JA, et al. Structural in silico analysis of cross-genotype-reactivity among naturally occurring HCV NS3-1073-variants in the context of HLA-A*02:01 allele. *Mol Immunol* (2011) 48(12–13):1461–7. doi:10.1016/j.molimm.2011.03.019
47. Zhang S, Bakshi RK, Suneetha PV, Fytily P, Antunes DA, Vieira GF, et al. Frequency, private specificity, and cross-reactivity of preexisting hepatitis C virus (HCV)-specific CD8+ T cells in HCV-seronegative individuals: implications for vaccine responses. *J Virol* (2015) 89(16):8304–17. doi:10.1128/JVI.00539-15
48. Mendes MF, Antunes DA, Rigo MM, Sinigaglia M, Vieira GF. Improved structural method for T-cell cross-reactivity prediction. *Mol Immunol* (2015) 67(2 Pt B):303–10. doi:10.1016/j.molimm.2015.06.017
49. Dhanik A, Kirshner JR, MacDonald D, Thurston G, Lin HC, Murphy AJ, et al. In-silico discovery of cancer-specific peptide-HLA complexes for targeted therapy. *BMC Bioinformatics* (2016) 17:286. doi:10.1186/s12859-016-1150-2
50. Moise L, Gutierrez AH, Bailey-Kellogg C, Terry F, Leng Q, Abdel Hady KM, et al. The two-faced T cell epitope: examining the host-microbe interface with JanusMatrix. *Hum Vaccin Immunother* (2013) 9(7):1577–86. doi:10.4161/hv.24615
51. Adams JJ, Narayanan S, Liu B, Birnbaum ME, Kruse AC, Bowerman NA, et al. T cell receptor signaling is limited by docking geometry to peptide-major histocompatibility complex. *Immunity* (2011) 35(5):681–93. doi:10.1016/j.immuni.2011.09.013
52. Antunes DA, Vieira GF, Rigo MM, Cibulski SP, Sinigaglia M, Chies JA. Structural allele-specific patterns adopted by epitopes in the MHC-I cleft and reconstruction of MHC:peptide complexes to cross-reactivity assessment. *PLoS One* (2010) 5(4):e10353. doi:10.1371/journal.pone.0010353
53. D'Orsogna LJ, Roelen DL, Doxiadis II, Claas FH. TCR cross-reactivity and allorecognition: new insights into the immunogenetics of allorecognition. *Immunogenetics* (2012) 64(2):77–85. doi:10.1007/s00251-011-0590-0
54. van den Heuvel H, Heutinck KM, van der Meer-Prins EMW, Yong SL, van Miert PPMC, Anholts JDH, et al. Allo-HLA cross-reactivities of Cytomegalovirus-, Influenza-, and Varicella Zoster virus-specific memory T cells are shared by different healthy individuals. *Am J Transplant* (2017) 17(8):2033–44. doi:10.1111/ajt.14279
55. Ayres CM, Scott DR, Corcelli SA, Baker BM. Differential utilization of binding loop flexibility in T cell receptor ligand selection and cross-reactivity. *Sci Rep* (2016) 6:25070. doi:10.1038/srep25070
56. Arber C, Feng X, Abhyankar H, Romero E, Wu MF, Heslop HE, et al. Survivin-specific T cell receptor targets tumor but not T cells. *J Clin Invest* (2015) 125(1):157–68. doi:10.1172/JCI75876
57. Rigo MM, Antunes DA, Vaz de Freitas M, Fabiano de Almeida Mendes M, Meira L, Sinigaglia M, et al. DockTope: a web-based tool for automated pMHC-I modelling. *Sci Rep* (2015) 5:18413. doi:10.1038/srep18413
58. Selin LK, Wlodarczyk MF, Kraft AR, Nie S, Kenney LL, Puzone R, et al. Heterologous immunity: immunopathology, autoimmunity and protection during viral infections. *Autoimmunity* (2011) 44(4):328–47. doi:10.3109/08916934.2011.523277
59. Petrova GV, Naumova EN, Gorski J. The polyclonal CD8 T cell response to influenza M158-66 generates a fully connected network of cross-reactive clonotypes to structurally related peptides: a paradigm for memory repertoire coverage of novel epitopes or escape mutants. *J Immunol* (2011) 186(11):6390–7. doi:10.4049/jimmunol.1004031
60. Fytily P, Dalekos GN, Schlaphoff V, Suneetha PV, Sarrazin C, Zauner W, et al. Cross-genotype-reactivity of the immunodominant HCV CD8 T-cell epitope NS3-1073. *Vaccine* (2008) 26(31):3818–26. doi:10.1016/j.vaccine.2008.05.045

61. Suzuki R, Shimodaira H. PvcLust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* (2006) 22(12):1540–2. doi:10.1093/bioinformatics/btl117
62. Chen AT, Cornberg M, Gras S, Guillonnet C, Rossjohn J, Trees A, et al. Loss of anti-viral immunity by infection with a virus encoding a cross-reactive pathogenic epitope. *PLoS Pathog* (2012) 8(4):e1002633. doi:10.1371/journal.ppat.1002633
63. Turner SJ, Kedzierska K, Komodromou H, La Gruta NL, Dunstone MA, Webb AI, et al. Lack of prominent peptide-major histocompatibility complex features limits repertoire diversity in virus-specific CD8+ T cell populations. *Nat Immunol* (2005) 6(4):382–9. doi:10.1038/ni1175
64. Petrey D, Honig B. GRASP2: visualization, surface properties, and electrostatics of macromolecular structures and sequences. *Methods Enzymol* (2003) 374:492–509. doi:10.1016/S0076-6879(03)74021-X
65. Garcia KC, Adams JJ, Feng D, Ely LK. The molecular basis of TCR germline bias for MHC is surprisingly simple. *Nat Immunol* (2009) 10(2):143–7. doi:10.1038/ni.f.219
66. Gras S, Burrows SR, Turner SJ, Sewell AK, McCluskey J, Rossjohn J. A structural voyage toward an understanding of the MHC-I-restricted immune response: lessons learned and much to be learned. *Immunol Rev* (2012) 250(1):61–81. doi:10.1111/j.1600-065X.2012.01159.x
67. Santori FR, Holmberg K, Ostrov D, Gascoigne NR, Vukmanovic S. Distinct footprints of TCR engagement with highly homologous ligands. *J Immunol* (2004) 172(12):7466–75. doi:10.4049/jimmunol.172.12.7466
68. Mazza C, Auphan-Anezin N, Gregoire C, Guimezanes A, Kellenberger C, Roussel A, et al. How much can a T-cell antigen receptor adapt to structurally distinct antigenic peptides? *EMBO J* (2007) 26(7):1972–83. doi:10.1038/sj.emboj.7601605
69. Holler PD, Chlewicki LK, Kranz DM. TCRs with high affinity for foreign pMHC show self-reactivity. *Nat Immunol* (2003) 4(1):55–62. doi:10.1038/ni863
70. Frankild S, de Boer RJ, Lund O, Nielsen M, Kesmir C. Amino acid similarity accounts for T cell cross-reactivity and for “holes” in the T cell repertoire. *PLoS One* (2008) 3(3):e1831. doi:10.1371/journal.pone.0001831
71. Wang RF, Wang HY. Immune targets and neoantigens for cancer immunotherapy and precision medicine. *Cell Res* (2017) 27(1):11–37. doi:10.1038/cr.2016.155
72. Dudley ME, Wunderlich JR, Robbins PF, Yang JC, Hwu P, Schwartzentruber DJ, et al. Cancer regression and autoimmunity in patients after clonal repopulation with antitumor lymphocytes. *Science* (2002) 298(5594):850–4. doi:10.1126/science.1076514
73. Tan MP, Gerry AB, Brewer JE, Melchiori L, Bridgeman JS, Bennett AD, et al. T cell receptor binding affinity governs the functional profile of cancer-specific CD8+ T cells. *Clin Exp Immunol* (2015) 180(2):255–70. doi:10.1111/cei.12570
74. Perica K, Varela JC, Oelke M, Schneck J. Adoptive T cell immunotherapy for cancer. *Rambam Maimonides Med J* (2015) 6(1):e0004. doi:10.5041/RMMJ.10179
75. Weber JS, Yang JC, Atkins MB, Disis ML. Toxicities of immunotherapy for the practitioner. *J Clin Oncol* (2015) 33(18):2092–9. doi:10.1200/JCO.2014.60.0379
76. Xing Y, Hogquist KA. T-cell tolerance: central and peripheral. *Cold Spring Harb Perspect Biol* (2012) 4(6):1–15. doi:10.1101/cshperspect.a006957
77. Khan JM, Ranganathan S. pDOCK: a new technique for rapid and accurate docking of peptide ligands to Major Histocompatibility Complexes. *Immune Res* (2010) 6:S2. doi:10.1186/1745-7580-6-S1-S2
78. Richter S, Wenzel A, Stein M, Gabdoulline RR, Wade RC. webPIPSA: a web server for the comparison of protein interaction properties. *Nucleic Acids Res* (2008) 36(Web Server issue):W276–80. doi:10.1093/nar/gkn181
79. Bryant DH, Moll M, Chen BY, Fofanov VY, Kavraki LE. Analysis of substructural variation in families of enzymatic proteins with applications to protein function prediction. *BMC Bioinformatics* (2010) 11:242. doi:10.1186/1471-2105-11-242
80. Glanville J, Huang H, Nau A, Hatton O, Wagar LE, Rubelt F, et al. Identifying specificity groups in the T cell receptor repertoire. *Nature* (2017) 547(7661):94–8. doi:10.1038/nature22976
81. Borrmann T, Cimons J, Cosiano M, Purcaro M, Pierce BG, Baker BM, et al. ATLAS: a database linking binding affinities with structures for wild-type and mutant TCR-pMHC complexes. *Proteins* (2017) 85(5):908–16. doi:10.1002/prot.25260
82. Hoffmann T, Marion A, Antes I. DynaDom: structure-based prediction of T cell receptor inter-domain and T cell receptor-peptide-MHC (class I) association angles. *BMC Struct Biol* (2017) 17(1):2. doi:10.1186/s12900-016-0071-7
83. Riley TP, Singh NK, Pierce BG, Weng Z, Baker BM. Computational modeling of T cell receptor complexes. *Methods Mol Biol* (2016) 1414:319–40. doi:10.1007/978-1-4939-3569-7_19
84. Zoete V, Irving M, Ferber M, Cuendet MA, Michielin O. Structure-based, rational design of T cell receptors. *Front Immunol* (2013) 4:268. doi:10.3389/fimmu.2013.00268
85. Malecek K, Grigoryan A, Zhong S, Gu WJ, Johnson LA, Rosenberg SA, et al. Specific increase in potency via structure-based design of a TCR. *J Immunol* (2014) 193(5):2587–99. doi:10.4049/jimmunol.1302344
86. Sharma P, Kranz DM. Recent advances in T-cell engineering for use in immunotherapy. *F1000Res* (2016) 5:1–12. doi:10.12688/f1000research.9073.1
87. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* (2017) 45(D1):D158–69. doi:10.1093/nar/gkw1099
88. Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. *Nucleic Acids Res* (2016) 44(D1):67–72. doi:10.1093/nar/gkv1276
89. Vita R, Overton JA, Greenbaum JA, Ponomarenko J, Clark JD, Cantrell JR, et al. The immune epitope database (IEDB) 3.0. *Nucleic Acids Res* (2015) 43(Database issue):D405–12. doi:10.1093/nar/gku938
90. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. *Nucleic Acids Res* (2000) 28(1):235–42. doi:10.1093/nar/28.1.235
91. Sinigaglia M, Antunes DA, Rigo MM, Chies JA, Vieira GF. CrossTope: a curate repository of 3D structures of immunogenic peptide: MHC complexes. *Database (Oxford)* (2013) 2013:bat002. doi:10.1093/database/bat002
92. Schrödinger, LLC. *The PyMOL Molecular Graphics System, Version 1.8*. (2015). Available from: <https://pymol.org/citing>
93. Pronk S, Pall S, Schulz R, Larsson P, Bjelkmar P, Apostolov R, et al. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* (2013) 29(7):845–54. doi:10.1093/bioinformatics/btt055
94. Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem* (2010) 31(2):455–61. doi:10.1002/jcc.21334
95. Li L, Li C, Sarkar S, Zhang J, Witham S, Zhang Z, et al. DelPhi: a comprehensive suite for DelPhi software and associated resources. *BMC Biophys* (2012) 5:9. doi:10.1186/2046-1682-5-9

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2017 Antunes, Rigo, Freitas, Mendes, Sinigaglia, Lizée, Kavraki, Selin, Cornberg and Vieira. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Immunological Landscape of esophageal squamous cell carcinoma

IMMUNE LANDSCAPE IN ESOPHAGEAL SQUAMOUS CELL CARCINOMA

Author information:

LUCIANA RODRIGUES CARVALHO BARROS¹, PAULO THIAGO SANTOS¹, MARCO ANTONIO MARQUES PRETTI¹, GUSTAVO FIORAVANTI VIEIRA^{5,6}, MARCELO ALVES DE SOUZA BRAGATTE⁶, MARCUS FABIANO DE ALMEIDA MENDES⁶, MARTIELA VAZ DE FREITAS⁶, NICOLE DE MIRANDA SCHERER², IVANIR MARTINS⁷, TATIANA DE ALMEIDA SIMÃO⁸, MARIANA BORONI², SHEILA COELHO SOARES LIMA¹, LUIS FELIPE RIBEIRO PINTO^{1,3}, MARTIN HERNAN BONAMINO^{1,4}

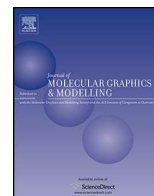
Affiliations:

1. Programa de Carcinogênese Molecular, Instituto Nacional de Câncer - INCA, Rio de Janeiro, Brazil.
2. Laboratório de Bioinformática e Biologia Computacional, Instituto Nacional de Câncer, INCA
3. Universidade do Estado do Rio de Janeiro, UERJ.
4. Vice-Presidência de Pesquisa e Coleções Biológicas (VPPCB) - Fundação Oswaldo Cruz, FIOCRUZ.
5. Programa de Pós-Graduação em Saúde e Desenvolvimento Humano/Universidade La Salle-Canoas
6. Programa de Pós-Graduação em Genética e Biologia Molecular/UFRGS
7. Divisão de Patologia, Instituto Nacional de Câncer, Rio de Janeiro, Brazil.
8. Departamento de Bioquímica, IBRAG, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, Brazil.



Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/JMGMLessons from molecular modeling human α -L-iduronidase

Danieli Forgiarini Figueiredo^{a,d,1}, Dinler A. Antunes^{a,d,1}, Maurício M. Rigo^{a,d},
 Marcus F.A. Mendes^{a,d}, Jader P. Silva^a, Fabiana Q. Mayer^b, Ursula Matte^{c,d},
 Roberto Giugliani^{c,d}, Gustavo F. Vieira^{a,d}, Marialva Sinigaglia^{a,d,*}

^a NBLI – Núcleo de Bioinformática do Laboratório de Imunogenética, Departamento de Genética, Universidade Federal do Rio Grande do Sul, Porto Alegre, Rio Grande do Sul, Brazil

^b Molecular Biology Laboratory, Instituto de Pesquisas Veterinárias Desidério Finamor, Fundação Estadual de Pesquisa Agropecuária Porto Alegre, Rio Grande do Sul, Brazil

^c Gene Therapy Center, Experimental Research Center, Hospital de Clínicas de Porto Alegre, Porto Alegre, Rio Grande do Sul, Brazil

^d Programa de Pós-Graduação em Genética e Biologia Molecular (PPGBM), Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, Rio Grande do Sul, Brazil

ARTICLE INFO

Article history:

Accepted 8 October 2014

Available online 18 October 2014

Keywords:

Homology modeling

Low identity template

 α -L-Iduronidase (IDUA)

Model evaluation tools

Molecular dynamics

Secondary structure assessment

ABSTRACT

Human α -L-iduronidase (IDUA) is a member of glycoside hydrolase family and is involved in the catabolism of glycosaminoglycans (GAGs), heparan sulfate (HS) and dermatan sulfate (DS). Mutations in this enzyme are responsible for mucopolysaccharidosis I (MPS I), an inherited lysosomal storage disorder. Despite great interest in determining and studying this enzyme structure, the lack of a high identity to templates and other technical issues have challenged both bioinformaticians and crystallographers, until the recent publication of an IDUA crystal structure (PDB: 4JXP). In the present work, four alternative IDUA models, generated and evaluated prior to crystallographic determination, were compared to the 4JXP structure. A combined analysis using several viability assessment tools and molecular dynamics simulations highlights the strengths and limitations of different comparative modeling protocols, all of which are based on the same low identity template (only 22%). Incorrect alignment between the target and template was confirmed to be a major bottleneck in homology modeling, regardless of the modeling software used. Moreover, secondary structure analysis during a 50 ns simulation seems to be useful for indicating alignment errors and structural instabilities. The best model was achieved through the combined use of Phyre 2 and Modeller, suggesting the use of this protocol for the modeling of other proteins that still lack high identity templates.

© 2014 Elsevier Inc. All rights reserved.

1. Introduction

Classical studies with globins provided the first example of a “molecular disease” [1,2], the first clues regarding the impact of amino acid exchanges on protein structure/function, and their consequences to human health [3–5]. The globin family also provided evidence that protein structures evolve more conservatively than protein sequences [6], and the tertiary structure conservancy observed among homologues enables comparative structure predictions. Since then, accurate prediction of the 3D structure of

proteins related to human diseases has been a major goal in structural bioinformatics [7].

Great progress has occurred over the past few decades, with improvements in the algorithms and computational resources. The Critical Assessment of protein Structure Prediction (CASP) competitions, which biannually challenge researchers to provide accurate models of an unreleased crystal structure, provide a wealth of information to this field [8,9]. These competitions also highlight the evolution of structure prediction and the persistent limitations and bottlenecks for comparative modeling [10,11].

In the present work, we focus on the homology modeling of the human α -L-iduronidase (IDUA, E.C: 3.2.1.76), a member of the glycoside hydrolase family [12]. Mutations in this enzyme are responsible for an inherited lysosomal storage disorder, called MPS I (Mucopolysaccharidosis I, OMIM #607014, #607015, #607016) [13,14]. The enzyme has 653 amino acids and is synthesized on the rough endoplasmic reticulum (ER), presenting a signal peptide with

* Corresponding author at: Universidade Federal do Rio Grande do Sul (UFRGS), Av. Bento Gonçalves 9500, Building 43323, Room 225, Brazil. Tel.: +555133089938

E-mail address: msinigaglia@gmail.com (M. Sinigaglia).

¹ These authors contributed equally to this work.

26 residues in its N-terminal portion. There are six glycosylation sites in its sequence, which are important for enzyme trafficking to the lysosome [15]. The enzyme architecture corresponds to a ($\alpha\beta$)₈ domain, which is a conserved structure that is also known as a TIM barrel fold [12].

Determining the 3D structure of this important enzyme would be pivotal for understanding phenotypic variations presented by MPS I [13], for predicting the impact of mutations [16] and for devising new pharmacological treatments [14]. However, technical issues postponed by years the determination of a crystal structure and the absence of a high identity template has been a major challenge for accurate modeling [12,17,18]. The recent publication of a human α -L-iduronidase crystal structure [19] allowed us to directly assess the efficacy and limitations of different comparative modeling protocols, all performed with a low identity template (only 22%). This convenient experiment is able to provide important lessons on comparative modeling and structural analysis, which can be applied for other challenging molecular targets involved in human diseases.

2. Materials and methods

2.1. Molecular modeling of α -L-iduronidase (IDUA)

The human α -L-iduronidase FASTA sequence was recovered from UNIPROT [20] (code P35475), and the first 26 amino acids, which belong to the signal peptide, were removed. Three different models of IDUA were then generated using alternative approaches. In each case, one model was selected among several, after an evaluation with the proper tools.

2.1.1. Model IDUA I-TASSER

The I-TASSER server was used to produce five models, which were ranked by the C-score and TM-score [21]. After the server is provided with the amino acid sequence of the target, it retrieves template proteins of similar folds from the PDB by LOMETS (Local Meta-Threading-Server). Afterwards, contiguous fragments recovered from the PDB are reassembled into full-length models, filling the missing regions using *ab initio* modeling. Additional steps are performed to remove the steric clash and to refine the global topology of the models (more information at <http://zhanglab.ccmb.med.umich.edu/I-TASSER/>).

For the IDUA sequence, I-TASSER identified 1uhvA as the best template for modeling (id1=0.21, id2=0.18, cov 0.75 and z-score=2.14). It should be noted that the models produced by this approach lack a C-terminal region that was not modeled, and modeled structures correspond to residues 27–635 of the wild-type IDUA. The best model presented a C-score of -1.58 and TM-score of 0.52 ± 0.15 , which are indicative of a reliable model with a correct global topology.

2.1.2. Model IDUA Rempel et al./Modeller (IDUA-RM)

Modeller 9v9 [22] was applied to produce a model using the crystal structure of beta-D-xylosidase from *Thermoanaerobacterium saccharolyticum* (PDB: 1UHV) as a template. The template structure was obtained after a combined search using psiBLAST and BLASTp from NCBI as well as HHPRED from the Max Plank Institute (<http://toolkit.tuebingen.mpg.de/hhpred>). The sequence alignment used for this approach was the same as previously published by Rempel and colleagues [17] (PDB: 1Y24). One hundred models were generated that covered residues from 36 to 504 of the wild-type IDUA, and the best model was selected using the DOPE score [22] and Procheck [23].

2.1.3. Model IDUA Phyre 2.0/Modeller (IDUA-PM)

In this approach, Modeller 9v9 [22] was applied using the same template (PDB: 1UHV) but with a different alignment. Sequence alignment covered residues from 27 to 542 of the wild-type IDUA (Figure S1), presenting 22% identity with the template sequence, and this was performed by folding recognition with Phyre 2 (<http://www.sbg.bio.ic.ac.uk/phyre2>) [24]. The alignment was further checked manually and adjusted, considering the location of insertion/deletion in loops and the correct positioning of crucial residues from the catalytic domain. One hundred models were generated, and the best model was selected using the DOPE score [22] and Procheck [23].

Supplementary Figure S1 related to this article can be found, in the online version, at <http://dx.doi.org/10.1016/j.jmngm.2014.10.004>.

2.1.4. Parameters for homology modeling with Modeller

Homology modeling with Modeller 9v9 was performed in a semi-automated fashion through the use of python scripts previously developed by our group. The modeling protocol followed the default optimization and refinement protocol, as described in the Modeller online manual (available at <http://salilab.org/modeller/9.13/manual/node19.html>). Briefly, each model is optimized with the variable target function method (VTFM) with conjugate gradients (CG), and later refined using molecular dynamics (MD) with simulated annealing (SA).

2.2. Comparison of generated models with IDUA crystal structure

The best model for each approach was evaluated using Procheck [23], Verify 3D (http://nihserver.mbi.ucla.edu/Verify_3D/) [25,26] and ModFOLD (<http://www.reading.ac.uk/bioinf/ModFOLD/>) [27], and the results were compared with those obtained for a crystal structure of IDUA (PDB: 4JXP). The IDUA model previously published by Rempel and colleagues [17] (PDB: 1Y24) was also included in this analysis.

2.3. Refinement of an IDUA crystal structure for molecular dynamics

The recently published crystal structure of human α -L-iduronidase (PDB: 4JXP) had two missing sites (residues 55–61 and 103–106) [19]. These sites had to be modeled before submitting this structure to a molecular dynamics (MD) simulation. Using the full sequence of human IDUA from UNIPROT [20] (code P35475), a new crystal structure containing missing loops (IDUA-Crystal) was generated with Modeller 9v9 [22].

Of note, the 4JXP crystal structure was later exchanged at PDB by the 4MJ4 structure. There are no relevant structural differences between 4JXP and 4MJ4, and both present the same sequence gaps. However, the 4MJ4 sequence presents three amino acid exchanges with respect to 4JXP and to the human α -L-iduronidase sequence recovered from UNIPROT [20] (code P35475). Because this work was performed before this exchange at PDB and because 4JXP presented the same reference sequence used for all models, we did not exchange our reference structure for 4MJ4.

2.4. Molecular dynamics simulations

IDUA models (IDUA-ITASSER, IDUA-RM, IDUA-PM and 1Y24) and the IDUA crystal structure were subjected to 50 ns of a molecular dynamics simulation using the GROMACS v4.5.1 package [28] on a Linux platform using the GROMOS96 (53a6) force field. An appropriate number of sodium (Na⁺) and chloride (Cl⁻) counterions was added to neutralize the system, with a final concentration

of 0.15 mol/L. A cubic box was defined with at least 15 Å of solvation layer around the protein, using a SPC water model and periodic boundary conditions. The *v-rescale* ($\tau_{\text{v}}=0.1$ ps) and *parrinello-rahman* ($\tau_{\text{p}}=2$ ps) algorithms were used for temperature and pressure coupling, respectively. Cutoff values of 1.2 nm were used both for van der Waals and Coulomb interactions, with *Fast Particle-Mesh Ewald* electrostatics (PME).

Our MD simulations were divided into four main stages: energy minimization (EM), solvation, thermalization and production. The EM stage was subdivided into three steps. First, the *steepest-descent* algorithm with position restraints for protein heavy atoms ($5000 \text{ kJ}^{-1} \text{ mol}^{-1} \text{ nm}^{-1}$) was applied, allowing only the solvent to relax. Afterwards, an EM with the same algorithm and no restraints was performed, allowing relaxation of the entire system. Finally, an EM using a *conjugate gradient* (CG) algorithm with no restraints was performed. The solvation stage was divided into several steps. First, an MD simulation with an *md* integrator algorithm and position restraints for all protein heavy atoms ($5000 \text{ kJ}^{-1} \text{ mol}^{-1} \text{ nm}^{-1}$) was performed at a temperature of 300 K for a period of 500 ps to allow for solvation layers formation. Then, temperature was reduced to 20 K (2 steps, total of 20 ps), after which position restraints were gradually reduced to $0.2 \text{ kJ}^{-1} \text{ mol}^{-1} \text{ nm}^{-1}$ (11 steps, total of 130 ps). During thermalization, the system was gradually heated from 20 to 300 K (with no restraints), increasing by approximately 50 K at each 320 ps step. Together, these equilibrium stages complete 2500 ps of simulation. This is the initial time for the production stage, in which the system was held at a constant temperature (300 K) and had no restraints up to 50 ns.

Simulation plots were generated with the respective programs from GROMACS v4.5.1 package [29] and visualized with xmgrace, the full-featured GUI-based version of Grace (<http://plasma-gate.weizmann.ac.il/Grace/>). Visual inspection of the MD trajectories was performed with VMD 1.9.1 [30], PyMOL 1.0 [31] and UCSF Chimera [32].

2.5. RMSD weighted by reference structure dynamics

Protein frames of each simulation were recovered (each 5 ns) and used to calculate the RMSD (root-mean-square deviation) against the reference structure (IDUA-Crystal). This was obtained following the rationale: $(Smod_{t(x)} - Scryst_{t(0)}) - (Scryst_{t(x)} - Scryst_{t(0)})$, where “ $Smod_{t(x)} - Scryst_{t(0)}$ ” is the RMSD of the modeled structure at a time x , in relation to the crystal structure (time = 0). Only sections of β -sheets and the α -helix of the TIM barrel (excluding loops) were considered because the great variability of the loops was expected to be even for the reference structure. The limits of each secondary structure were defined by considering the crystal structure sequence (PDB: 4JXP). An alternative weighted RMSD was calculated by discounting the variability presented by the IDUA-Crystal at each time point of the simulation. In both cases, the RMSD was calculated for all atoms (backbone and side chains).

3. Results and discussion

3.1. Previous models of human α -L-iduronidase (IDUA)

The human α -L-iduronidase is an enzyme from the glycoside hydrolases group and is involved in the catabolism of glycosaminoglycans (GAGs), heparan sulfate (HS) and dermatan sulfate (DS) [12,13]. Accurate modeling of α -L-iduronidase was a major goal for researchers because several mutations are related to a broad variety of clinical manifestations of mucopolysaccharidosis type I (MPS I) [16,33]. However, lack of high identity templates and difficulties in producing a crystal structure of the enzyme have challenged

researchers. Rempel and colleagues published the first model for the IDUA enzyme in 2005 [17], which was deposited at PDB under the code 1Y24. This model was obtained with the SWISS-MODEL online server [34], which performs homology modeling in an automated fashion, and used the beta-D-xylosidase of *T. saccharolyticum* (PDB: 1UHV, 1PX8) as a template. This model was downloaded and included in our analysis to evaluate the impact of alternative alignments and molecular modeling tools.

Recently, a second model of IDUA was published by Chandar & Mahalingam using the automated tool Schrödinger PRIME [18]. The structure, however, was not made available.

The IDUA crystal structure was finally published in November 2013 by Bie and colleagues [19] and provided insightful information on the complete structure of this important enzyme. This structure was initially made available under the PDB code 4JXP, which was afterwards replaced by the 4MJ4 structure (see Section 2). It is important to consider that all models from the present study were generated before the publication of the IDUA crystal structure and were therefore predicted and evaluated without any influence from these experimental data.

3.2. Evaluation of generated models and comparison with a reference crystal structure

A Ramachandran plot is a well-known evaluation tool to assess the stereochemical quality of a given model through the analysis of *phi* and *psi* angles for all protein residues [23]. This analysis was performed with Procheck software [23] for the best models produced by three alternative approaches (Table 1) as well as for the previously published model (PDB: 1Y24) and for the reference crystal structure (PDB: 4JXP). Considering this analysis, the best models were IDUA-RM and IDUA-PM, which had 87.3% and 85% of residues in the most favored regions, respectively. Both models were produced with Modeller [22], which indicates the ability of this software to assign stereochemical properties. The percentage of residues in disallowed regions was also lower for these two models when compared to 1Y24 and IDUA-ITASSER, which was the worst model in this analysis (Table 1).

The overall quality of all models was also evaluated using other two well-known pieces of software, ModFOLD [27] and Verify3D [26] (Table 2). Models 1Y24 and IDUA-RM had a global quality score lower than the others at ModFOLD and have presented a low percentage of its residues (<60%) with an averaged 3D-1D score >0.2 at Verify 3D, which configures that this model failed in this evaluation. IDUA-PM and IDUA-ITASSER presented high confidence at ModFold (probability of an incorrect model at lower than 1/100) and were approved at Verify 3D. Moreover, IDUA-PM presented the highest value of global quality (0.3734). Of note, values above 0.4 indicate high similarity with the native structure. Considering these three independent methods for viability assessment, IDUA-PM was considered to be the best IDUA model generated in this work.

As a crystal structure of IDUA was recently made available (PDB: 4JXP), it is possible to directly assess which model was closest to the native structure (Fig. 1). Visual inspection indicates a similar arrangement of the catalytic domain (TIM barrel) for all models. However, important differences were observed among the models when considering the specific amino acids sequence that composes the structure of each TIM barrel. These differences are highlighted by performing the root mean square deviation (RMSD) for the crystal structure backbone with respect to the same residues in each proposed model (Table 3). Again, the closest model to the native structure was IDUA-PM, which presented an RMSD of 5.78 Å for the catalytic domain and 6.17 Å for the whole modeled structure. Although this value was high, it is 2-fold smaller than that of the previous model (1Y24), which had an RMSD of 12.24 Å for

Table 1
Percentage (%) of protein residues in each region of the Ramachandran Plot.

Regions	4JXP (crystal)	1Y24 (model)	IDUA-RM (model)	IDUA-PM (model)	IDUA-ITASSER (model)
Most favored	86.5	72.5	87.3	85.0	76.7
Additional allowed	12.9	23.4	11.0	12.3	16.5
Generously allowed	0.6	2.4	1.0	1.6	4.1
Disallowed	0.0	1.7	0.7	1.1	2.7

All values are given as %.

Table 2
Additional evaluation tools for model viability assessment.

Models	ModFOLD			Verify 3D	
	Global model quality score	P-value	Confidence	Residues with an averaged 3D-1D > 0.2 ^a	Results
1Y24	0.2081	2.663E-2	Intermediate	56.29	Reproved
IDUA-RM	0.2160	2.288E-2	Intermediate	56.08	Reproved
IDUA-PM	0.3734	1.898E-3	High	86.05	Approved
IDUA-ITASSER	0.2961	5.747E-3	High	80.79	Approved
4JXP	0.8555	1.239E-4	Cert	97.07	Approved

^a Values given in %.

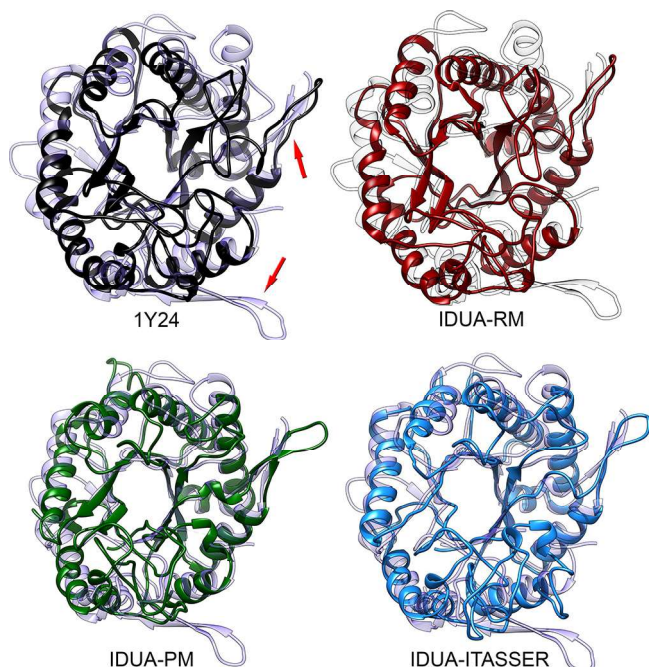


Fig. 1. Catalytic domain of α -L-iduronidase (TIM Barrel). Superposition of different models over IDUA crystal structure (4JXP), which is depicted in light gray. Red arrows indicate two β -hairpins which protract from TIM barrel domain. Observe that one of these β -hairpins was not properly folded in any of the models, and the other was completely folded only in IDUA-PM.

Table 3
RMSD against crystal structure.

Model	RMSD ^a (TIM barrel)	RMSD ^a (protein)
1Y24	12.24 Å	16.65 Å
IDUA-RM	11.84 Å	16.7 Å
IDUA-PM	5.78 Å	6.17 Å
IDUA-ITASSER	6.35 Å	9.12 Å

^a Backbone Root Mean Square Deviation (RMSD). Values correspond to the RMSD of specific amino acids which compose each model TIM barrel, against the same residues in IDUA crystal structure (PDB: 4JXP).

the TIM barrel and 16.65 Å for the entire structure. The second model, which had better results in this analysis, is IDUA-ITASSER, which presented a slightly larger RMSD for the catalytic domain (6.35 Å).

The TIM barrel domain was already present at the structure of beta-D-xylosidase of *T. saccharolyticum* (PDB: 1UHV), which was used as the template for three of the models evaluated in our study (1Y24, IDUA-RM and IDUA-PM). Comparison between this template and the reference crystal structure for human α -L-iduronidase (PDB: 4JXP) reveals some differences that contributed to the high RMSD values observed for the models (Figure S2). In spite of the great structural similarity between the two catalytic domains, it is important to note the presence of two β -hairpins that protract from the IDUA barrel. One of these structures was not present in the template and, consequently, was not properly folded in any of the models (Fig. 1). Moreover, it is also possible to observe the absence of the fibronectin domain in the beta-D-xylosidase structure (a domain that was not included in our modes) as well as small differences at the β -sandwich domain.

Supplementary Figure S2 related to this article can be found, in the online version, at <http://dx.doi.org/10.1016/j.jmglm.2014.10.004>.

The choice of the best model and the sequence alignment are key points for molecular homology modeling [35,36]. When sequence identity between the target and template is higher than 90%, highly accurate models are obtained. On the other hand, sequence alignment represents the major bottleneck for homology modeling in cases where the sequence identity is lower than 25%, which may produce crucial errors in the generated models [37,38]. In our work, IDUA has sequence identity of only 22% with the chosen template, and sequence alignment was decisive for obtaining a more reliable model. Our IDUA-PM model, which was generated with Modeller [22] using an alignment from Phyre 2 [24], was the closest model in relation to the crystal structure. The most dramatic errors were observed for the previously published model, 1Y24, and for IDUA-RM, which used the same alignment, regardless of the use of different software for modeling (Swiss PDBViewer [34] and Modeller [22], respectively).

Differences among models become clearer when comparing the maps of secondary structures (Figure S3). A shorter sequence was used to model the catalytic domain of the 1Y24 and IDUA-RM models (the region between the red boxes). Moreover, great divergence of secondary structures is observed when comparing these models with the crystal structure, especially in the region between the α 7 and α 8 helices. These models also lack a β 1 sheet and two β -hairpins (indicated with "C" and "D"). Although present in these models, the β 6 and β 8 sheets were formed in an incorrect region of the sequence. Finally, there are also problems with the size and location of the α 3, α 6 and α 8 helices. On the other hand, IDUA-PM

and IDUA-ITASSER models presented the TIM barrel in the same region as the crystal structure, with extensive agreement in the alignment of their secondary structures. The IDUA-PM model lacks $\beta 1$ and $\beta 8$ sheets, but all helices are in place. Moreover, one of the β -hairpins was correctly folded (Figure S3).

Supplementary Figure S3 related to this article can be found, in the online version, at <http://dx.doi.org/10.1016/j.jmglm.2014.10.004>.

3.3. Molecular dynamics of the modeled structures

Molecular dynamics simulations were also used to evaluate the models and to compare the dynamics to the reference structure (IDUA-Crystal). This powerful approach allows us to assess structure dynamics in solution, which provides insightful information on protein stability, and has been successfully applied to evaluate structural models [39,40]. In some cases, it has also been used to refine models [41,42]. However, a recent study presenting simulation results for 24 proteins selected from CASP showed that, in most cases, proteins drifted away from crystal structure [43]. According to the authors, force field accuracy seems to be a limiting factor for this purpose.

In the present work, great stability was observed for the IDUA-Crystal regarding the maintenance of secondary structures during a 50 ns simulation (Figure S3). Greater variation was observed for the proposed models, especially considering the catalytic domain of the 1Y24 and IDUA-RM models. The alignment error becomes evident when considering the secondary structures of these models. The end of the TIM barrel domain is defined by $\alpha 8$ and $\beta 8$ structures, which in these models were located before the correct position. This error was highlighted by molecular dynamics, which indicated great instability at this region.

The IDUA-PM and IDUA-ITASSER models had an overall stability similar to the reference structure. The IDUA-PM model presented partial alterations in some helices and the unfolding of the $\beta 7$ sheet, which was also observed for the other models. Of note, the correctly predicted β -hairpin was stable throughout the simulation. The IDUA-ITASSER model had similar results. In this case, it is interesting to observe the formation of the $\beta 8$ sheet after 10 ns of simulation. This structure was not predicted in the input model.

Considering the RMSD of the catalytic domain, the IDUA-Crystal presented stability after 16 ns of simulation, with lower values when compared to the models (Fig. 2). Models 1Y24, IDUA-RM and IDUA-ITASSER presented a similar RMSD, achieving stability after only 30 ns. Greater instability was observed in this analysis for the IDUA-PM model, which achieved stability after only 35 ns, with higher RMSD values. This analysis refers to the deviation of catalytic domain (including loops) for each of the models in relation to its own conformation at the beginning of the simulation. The main idea was to observe the overall variability of initial structures throughout the simulation.

Using as the input the structural data from molecular dynamics, an alternative validation method was performed. First, the RMSD against the reference structure (IDUA-Crystal) was calculated for ten frames (each 5 ns) that were recovered from each simulated model (Figure S4). Afterwards, aiming to highlight the dynamic similarities in respect to the reference structure, a “weighted RMSD” was calculated by discounting the IDUA-Crystal variability at each time of simulation (Fig. 3). This analysis indicates a tendency of all models to converge with the reference structure throughout the simulation (reduction of RMSD values). Moreover, considering the maintenance of the 3D scaffold of the catalytic domain, IDUA-ITASSER and IDUA-PM were the two models that most closely matched the dynamic behavior of the reference structure. In this analysis, the average RMSD for IDUA-ITASSER and

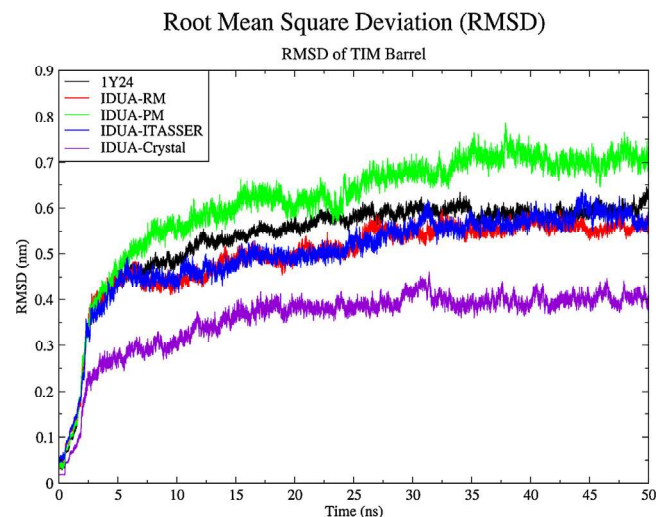


Fig. 2. TIM Barrel Root Mean Square Deviation (RMSD) of each simulated model and crystal structure (Model-4JXP), throughout a 50 ns molecular dynamics simulation. Each line indicates the divergence of a given structure in relation to its own initial conformation.

IDUA-PM was 3.66 and 4.28 Å, respectively (for all atoms). Thus, despite the fact that the RMSD against initial structures (Fig. 2) indicates greater stability for 1Y24 and IDUA-RM, it actually reflects an alternative (incorrect) folding of the catalytic domain in these models and a higher value of RMSD against the reference structure (Fig. 3).

Supplementary Figure S4 related to this article can be found, in the online version, at <http://dx.doi.org/10.1016/j.jmglm.2014.10.004>.

The higher RMSD values observed for IDUA-PM in Fig. 2 reflect a small increase in diameter of the catalytic domain during the simulated time (Figure S5), without domain unfolding, which might be influenced by the absence of the other domains of the enzyme. In spite of that, IDUA-PM presented a more precise modeling of the TIM barrel, which is sustained throughout the simulation (Fig. 3 and Figure S3). Regarding the radius of gyration (RoG) analysis, a better

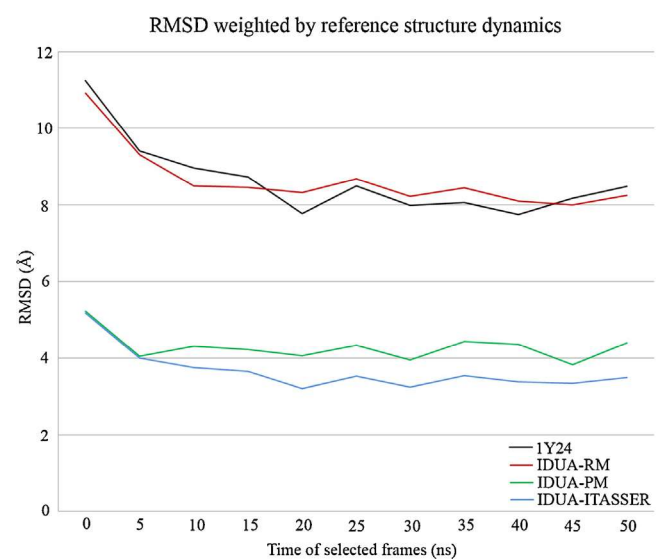


Fig. 3. TIM Barrel Root Mean Square Deviation (RMSD) weighted by reference structure dynamics. Each line indicates the divergence of a given structure in relation to the reference structure (Model-4JXP), discounting the divergence observed during Model-4JXP simulation. Direct RMSD against Model-4JXP (without discounts) can be seen in Figure S3.

fit can also be observed of IDUA-PM and IDUA-ITASSER values with those from the reference structure at the first half of simulation (Figure S5). The IDUA-RM and 1Y24 models presented lower values of RoG, which was also a consequence of using a shorter sequence for catalytic domain folding (wrong alignment). Additionally, the IDUA structure is highly glycosylated *in vivo*, and the absence of glycans in our models can also contribute to the observed instability. Recent structural analyses revealed the importance of N372 glycosylation for enzyme activity, indicating a direct interaction between an N-glycan mannose residue and the substrate [19,44].

Supplementary Figure S5 related to this article can be found, in the online version, at <http://dx.doi.org/10.1016/j.jmglm.2014.10.004>.

Finally, a root mean square fluctuation (RMSF) analysis was also performed for the TIM barrel residues of each model during the simulation. Little overall fluctuation was observed for catalytic domain residues of the IDUA-Crystal, presenting only six peaks of higher flexibility (Figure S6). These peaks are related to coil regions (near residues 60, 187 and 350) and two β -hairpins (near residues 105 and 373). One exception is made for the third peak (near residue 165), which corresponds to the final portion of α 3, a region that was unfolded during simulation (Figure S3). Regarding the models, IDUA-ITASSER had the most similar behavior in comparison to the IDUA-Crystal, and 1Y24 presented the greatest divergence.

Supplementary Figure S6 related to this article can be found, in the online version, at <http://dx.doi.org/10.1016/j.jmglm.2014.10.004>.

Therefore, molecular dynamics provided a series of additional data, which allowed us to compare our models with the crystal structure, highlighting the limitations of each model. Alternatively, a crystal of other proteins from the same family could be used as a reference structure, providing information on the dynamic behavior of the domains of interest (*e.g.*, TIM Barrel). Even in the absence of a reference crystal structure, these analyses could be used as additional tools for model viability assessment. For instance, secondary structure analysis throughout the simulation was shown to be an efficient way to detect alignment problems and segments that were wrongly folded (Figure S3).

3.4. Applicability to other targets

Differences between the results of alternative methods and alignments indicate hidden obstacles to automated production of reliable models, highlighting the importance of careful modeling and viability assessment. The combined use of Phyre 2 [24] and Modeller [22] for accurate alignment and homology modeling provided an improved model of IDUA 3D structure. This goal was achieved using a template with only 22% sequence identity, and the model quality was confirmed by different methods: Ramachandran, ModFOLD, Verify3D, molecular dynamics and direct comparison with crystal structure. Our results suggest the applicability of this combined approach to other important targets that still lack high similarity templates. For instance, the GAGs degradation pathway is composed of 11 enzymes [45,46], five of which have not yet had their structure determined by experimental methods: iduronate 2-sulfatase, heparan N-sulfatase, α -N-acetylglucosaminidase, acetyl CoA:alpha-glucosaminide N-acetyltransferase and N-acetylglucosamine-6-sulfatase, where deficiencies cause MPS II, MPS IIIA, MPS IIIB, MPS IIIC and MPS IIID, respectively. These are highly important targets for structural analysis because mutations of these enzymes are related to different types of MPS [45,46]. Furthermore, mutations of enzymes responsible for the GAGs synthesis are also implicated in human genetic disorders and could also be a target of molecular modeling for structural analysis [47].

4. Conclusions

Molecular homology modeling of proteins with low identity templates remains a highly difficult challenge, which must be addressed by the combined use of alternative approaches. The correct alignment between target and template is a critical step and has a direct impact on model quality. In addition, the careful use of reliable evaluation tools over the best ranked models is a critical step, which also benefits from the combined use of tools with different characteristics. Finally, the conservancy of known secondary structure patterns through molecular dynamics provides a powerful tool for model quality assessment.

In the present work, four alternative models of the human α -L-iduronidase were evaluated and compared with the recently published crystal structure (PDB: 4JXP). Superposition of catalytic domains from the best models (IDUA-PM and IDUA-ITASSER) with the reference crystal structure presented an RMSD of 5.78 Å and 6.35 Å (TIM barrel backbone), respectively. These values are twice smaller than the previously published model (PDB: 1Y24), which presented an RMSD of 12.54 Å (TIM barrel backbone). In conjunction with other conducted evaluations, this result highlights the performance gain of our approach and justifies the use of a similar protocol for modeling of proteins that lack high identity templates, such as other enzymes of the GAGs degradation pathway.

Acknowledgments

This work was supported through scholarships from CNPq, CAPES and FAPERGS. We would also like to thank the Centro Nacional de Supercomputação (CESUP-RS) for allowing access to its computational resources.

References

- [1] L. Pauling, H.A. Itano, et al., Sickle cell anemia: a molecular disease, *Science* 110 (1949) 543–548.
- [2] V.M. Ingram, A specific chemical difference between the globins of normal human and sickle-cell anaemia haemoglobin, *Nature* 178 (1956) 792–794.
- [3] J.C. Kendrew, G. Bodo, H.M. Dintzis, R.G. Parrish, H. Wyckoff, D.C. Phillips, A three-dimensional model of the myoglobin molecule obtained by X-ray analysis, *Nature* 181 (1958) 662–666.
- [4] M.F. Perutz, M.G. Rossmann, A.F. Cullis, H. Muirhead, G. Will, A.C. North, Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5-Å resolution, obtained by X-ray analysis, *Nature* 185 (1960) 416–422.
- [5] J. Baldwin, C. Chothia, Haemoglobin: the structural changes related to ligand binding and its allosteric mechanism, *J. Mol. Biol.* 129 (1979) 175–220.
- [6] J.T. Lecomte, D.A. Vuletich, A.M. Lesk, Structural divergence and distant relationships in proteins: evolution of the globins, *Curr. Opin. Struct. Biol.* 15 (2005) 290–301.
- [7] J.G. Mullins, Structural modelling pipelines in next generation sequencing projects, *Adv. Protein Chem. Struct. Biol.* 89 (2012) 117–167.
- [8] J. Moult, A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction, *Curr. Opin. Struct. Biol.* 15 (2005) 285–289.
- [9] A. Kryshchuk, K. Fidelis, J. Moult, CASP9 results compared to those of previous CASP experiments, *Proteins* 79 (Suppl. 10) (2011) 196–207.
- [10] A. Runthala, Protein structure prediction: challenging targets for CASP10, *J. Biomol. Struct. Dyn.* 30 (2012) 607–615.
- [11] J. Moult, K. Fidelis, A. Kryshchuk, T. Schwede, A. Tramontano, Critical assessment of methods of protein structure prediction (CASP) – round X, *Proteins* 82 (Suppl2) (2013) 1–6.
- [12] A.J. Harvey, M. Hrmova, R. De Gori, J.N. Varghese, G.B. Fincher, Comparative modeling of the three-dimensional structures of family 3 glycoside hydrolases, *Proteins* 41 (2000) 257–269.
- [13] E.F. Neufeld, J. Muenzer, The mucopolysaccharidoses, in: C.R. Scriver, A.L. Beaudet, W.S. Sly, D. Valle, B. Childs, K.W. Kinzler, et al. (Eds.), *The Metabolic and Molecular Bases of Inherited Disease*, vol. III, McGraw-Hill, Medical Publishing Division, 2001, p. 3421.
- [14] G. Parenti, Treating lysosomal storage diseases with pharmacological chaperones: from concept to clinics, *EMBO Mol. Med.* 1 (2009) 268–279.
- [15] E.H. Schuchman, R.J. Desnick, Mucopolysaccharidosis type I subtypes presence of immunologically cross-reactive material and *in vitro* enhancement of the residual α -L-iduronidase activities, *J. Clin. Invest.* 81 (1988) 98–105.
- [16] G.M. Pastores, P. Arn, M. Beck, J.T.R. Clarke, N. Guffon, P. Kaplan, et al., The MPS I registry: design, methodology, and early findings of a global disease registry

- for monitoring patients with Mucopolysaccharidosis Type I, *Mol. Genet. Metab.* 91 (2007) 37–47.
- [17] B.P. Rempel, L.A. Clarke, S.G. Withers, A homology model for human alpha-L-iduronidase: insights into human disease, *Mol. Genet. Metab.* 85 (2005) 28–37.
- [18] S.S. Chandar, K. Mahalingam, Mucopolysaccharidosis type I: homology modeling and docking analysis of the lysosomal enzyme, human α -L-iduronidase, *Afr. J. Pharm. Pharmacol.* 6 (2012) 2027.
- [19] H. Bie, J. Yin, X. He, A.R. Kermode, E.D. Goddard-Borger, S.G. Withers, et al., Insights into mucopolysaccharidosis I from the structure and action of alpha-L-iduronidase, *Nat. Chem. Biol.* 9 (2013) 739–745.
- [20] U. Consortium, Update on activities at the Universal Protein Resource (UniProt) in 2013, *Nucleic Acids Res.* 41 (2013) D43–D47.
- [21] Y. Zhang, I-TASSER server for protein 3D structure prediction, *BMC Bioinformatics* 9 (2008) 40.
- [22] M.A. Marti-Renom, A.C. Stuart, A. Fiser, R. Sanchez, F. Melo, A. Sali, Comparative protein structure modeling of genes and genomes, *Annu. Rev. Biophys. Biomol. Struct.* 29 (2000) 291–325.
- [23] R.A. Laskowski, J.A. Rullmann, M.W. MacArthur, R. Kaptein, J.M. Thornton, AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR, *J. Biomol. NMR* 8 (1996) 477–486.
- [24] L.A. Kelley, M.J. Sternberg, Protein structure prediction on the Web: a case study using the Phyre server, *Nat. Protoc.* 4 (2009) 363–371.
- [25] J.U. Bowie, R. Luthy, D. Eisenberg, A method to identify protein sequences that fold into a known three-dimensional structure, *Science* 253 (1991) 164–170.
- [26] R. Luthy, J.U. Bowie, D. Eisenberg, Assessment of protein models with three-dimensional profiles, *Nature* 356 (1992) 83–85.
- [27] L.J. McGuffin, The ModFOLD server for the quality assessment of protein structural models, *Bioinformatics (Oxford, England)* 24 (2008) 586–587.
- [28] S. Pronk, S. Pall, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, et al., GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit, *Bioinformatics* 29 (2013) 845–854.
- [29] D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A.E. Mark, H.J.C. Berendsen, GROMACS: fast, flexible, and free, *J. Comput. Chem.* 26 (2005) 1701–1718.
- [30] W. Humphrey, A. Dalke, K. Schulten, VMD: visual molecular dynamics, *J. Mol. Graph.* 14 (1996) 33–38, 27–8.
- [31] W.L. DeLano, S. Bromberg, PyMOL User's Guide, DeLano Scientific LLC, San Francisco, 2004.
- [32] E.F. Pettersen, T.D. Goddard, C.C. Huang, G.S. Couch, D.M. Greenblatt, E.C. Meng, et al., UCSF Chimera – a visualization system for exploratory research and analysis, *J. Comput. Chem.* 25 (2004) 1605–1612.
- [33] N.J. Terlato, G.F. Cox, Can mucopolysaccharidosis type I disease severity be predicted based on a patient's genotype? A comprehensive review of the literature, *Genet. Med.* 5 (2003) 286–294.
- [34] N. Guex, M.C. Peitsch, SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling, *Electrophoresis* 18 (1997) 2714–2723.
- [35] C.N. Cavasotto, S.S. Phatak, Homology modeling in drug discovery: current trends and applications, *Drug Discovery Today* 14 (2009) 676–683.
- [36] A. Sali, T.L. Blundell, Comparative protein modelling by satisfaction of spatial restraints, *J. Mol. Biol.* 234 (1993) 779–815.
- [37] C. Chothia, A.M. Lesk, The relation between the divergence of sequence and structure in proteins, *EMBO J.* 5 (1986) 823–826.
- [38] M.J. Sippl, Recognition of errors in three-dimensional structures of proteins, *Proteins* 17 (1993) 355–362.
- [39] S. Della-Longa, A. Arcovito, Structural and functional insights on folate receptor alpha (FRalpha) by homology modeling, ligand docking and molecular dynamics, *J. Mol. Graphics Modell.* 44 (2013) 197–207.
- [40] K. Kulleperuma, S.M. Smith, D. Morgan, B. Musset, J. Holyoake, N. Chakrabarti, et al., Construction and validation of a homology model of the human voltage-gated proton channel hHv1, *J. Gen. Physiol.* 141 (2013) 445–465.
- [41] T. Kwon, A.L. Harris, A. Rossi, T.A. Bargiello, Molecular dynamics simulations of the Cx26 hemichannel: evaluation of structural models with Brownian dynamics, *J. Gen. Physiol.* 138 (2011) 475–493.
- [42] C.B. Platania, S. Salomone, G.M. Leggio, F. Drago, C. Bucolo, Homology modeling of dopamine D2 and D3 receptors: molecular dynamics refinement and docking evaluation, *PLoS ONE* 7 (2012) e44316.
- [43] A. Raval, S. Piana, M.P. Eastwood, R.O. Dror, D.E. Shaw, Refinement of protein structure homology models via long, all-atom molecular dynamics simulations, *Proteins* 80 (2012) 2071–2079.
- [44] N. Maita, T. Tsukimura, T. Taniguchi, S. Saito, K. Ohno, H. Taniguchi, et al., Human alpha-L-iduronidase uses its own N-glycan as a substrate-binding and catalytic module, *Proc. Natl. Acad. Sci. U. S. A.* 110 (2013) 14628–14633.
- [45] Z. Banecka-Majkutewicz, J. Jakobkiewicz-Banecka, M. Gabig-Ciminska, A. Wegrzyn, G. Wegrzyn, Putative biological mechanisms of efficiency of substrate reduction therapies for mucopolysaccharidoses, *Arch. Immunol. Ther. Exp. (Warsz)* 60 (2012) 461–468.
- [46] J. Jakobkiewicz-Banecka, E. Piotrowska, M. Gabig-Ciminska, E. Borysiewicz, M. Slominska-Wojewodzka, M. Narajczyk, et al., Substrate reduction therapies for mucopolysaccharidoses, *Curr. Pharm. Biotechnol.* 12 (2011) 1860–1865.
- [47] S. Mizumoto, S. Ikegawa, K. Sugahara, Human genetic disorders caused by mutations in genes encoding biosynthetic enzymes for sulfated glycosaminoglycans, *J. Biol. Chem.* 288 (2013) 10953–10961.