JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Data Mining of Supersecondary Structure Homology between Light Chains of Immunoglobulins and MHC Molecules: Absence of the Common Conformational Fragment in the Human IgM Rheumatoid Factor

Hiroshi Izumi,*[,†] Akihiro, Wakisaka,[†] Laurence A. Nafie,[‡,§] and Rina K. Dukor[§]

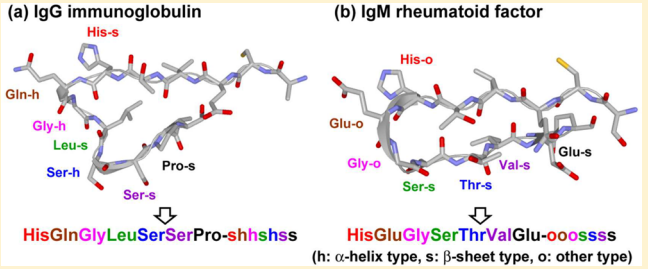
[†]National Institute of Advanced Industrial Science and Technology (AIST), AIST Tsukuba West, 16-1 Onogawa, Tsukuba, Ibaraki 305-8569, Japan

[‡]Department of Chemistry, Syracuse University, Syracuse, New York 13244-4100, United States

[§]BioTools, Inc., 17546 SR 710 (Bee Line Hwy) Jupiter, Florida 33458, United States

Supporting Information

ABSTRACT: It is shown that fuzzy search and data mining techniques of supersecondary structure homology for subunits of proteins using conformational code patterns of α-helix-type ($3\beta5\alpha4\beta$) and β-sheet-type ($6\alpha4\beta4\beta$) fragments can be used to extract correlations between fragments of MHC class I molecules and the light chain of immunoglobulins. The new method of conformational pattern analysis with fuzzy search of structural code homology reflects well the shape of main chain rather than secondary structure in comparison with the DSSP method. Further, the data mining technique using the combination of h- and s-fragment patterns can quantify the supersecondary structure homology between any subunits of proteins with different amino acid sequences. Characteristic fragment patterns (string "shhshss"), which were sandwiched between two identical amino acid sequences His and Pro, were found in light chains of various types of immunoglobulins, α-chain and β-2 microglobulin of MHC class I and α-chain and β-chain of MHC class II, but not in heavy chains of Fab immunoglobulin fragments and T cell receptors (TCR). Leukocyte immunoglobulin-like receptors (LILR) are related by the conformational fragment (string "shhshss") to β-2 microglobulins as a type of pair forms (string "sohsss"). Further, human IgM rheumatoid factor, one of the immunoglobulins, did not strongly exhibit the conformational fragment pattern. Nonclassic MHC class I molecules CD1D, MIC-A, and MIC-B, which have functions to activate NKT, NK, and T cells, did not also clearly show the patterns. These code-driven mining techniques can be utilized as a metadata-generating tool for systems biology to elucidate the biological function of such conformational fragments of MHC I and II molecules, which come in contact with various signal ligands on the surface of T cells and natural killer cells.



INTRODUCTION

Major histocompatibility complex (MHC) classes I[1,2] and II[3] molecules are the key proteins for organism self-recognition and have polymorphisms to defend against a great diversity of microbes. For example, natural killer (NK) cells can recognize and kill tumor cells lacking "self" markers, such as MHC class I, but the basis for this recognition is not completely understood.[2] Several common autoimmune diseases such as rheumatoid arthritis are deeply related to MHC class II and other immune modulators.[3]

The polymorphisms of amino acid sequences and molecular structures for MHC molecules and immunoglobulins are confusing and make the analysis of structural homology and change using the amino acid sequences very difficult. Further, no effective method to compare with supersecondary structure homology of many proteins currently exists. Therefore, we have developed data mining techniques based on backbone conformations to analyze the supersecondary structure homology of proteins with different amino acid sequences. Previously, we have proposed a conformational code for the description of conformations of all kinds of chemical compounds based on structural analysis using vibrational circular dichroism (VCD) of chiral bioactive compounds.[4−7] The conformational code consists of the combination of the codes of regional angle locations and conformational elements (Figure 1), and the conformational elements representing the classification of dihedral angles are substituted for the symbols indicating the bond locations (alphabets of angle locations).[6] For example, the conformational elements 1, 2, 3, 4, 5, and 6 correspond to the conformational terms, *T* (*trans*), G^+ (+*gauche*), G^- (−*gauche*), *sp* (*synperiplanar*), +*ac* (+*anticlinal*),

Received: September 3, 2012
Published: February 10, 2013

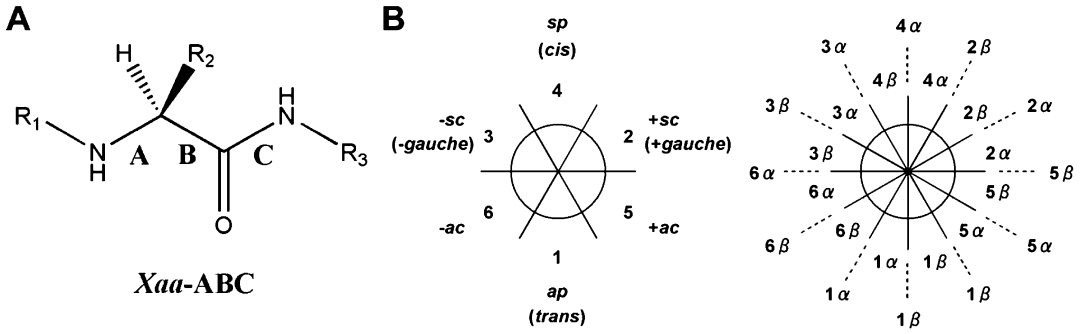


Figure 1. Definition of angle locations and conformational elements for conformational code of main chains of proteins. (A) Angle locations consist of prefix of amino acids (Xaa) and symbols indicating the bond locations (ABC). (B) Conformational elements represent classification of dihedral angles, and the elements 1, 2, 3, 4, 5, and 6 correspond to conformational terms, *T* (*trans*), G^+ (*+gauche*), G^- (*−gauche*), *sp* (*synperiplanar*), *+ac* (*+anticlinal*), and *−ac* (*−anticlinal*), respectively. The terms, α and β, in the conformational elements mean clockwise and counterclockwise, respectively.

and *−ac* (*−anticlinal*), respectively. The terms, α and β, in the conformational elements mean clockwise and counterclockwise, respectively, and conclusively, the local structural differences of optimized conformations using the density functional theory calculations can be discriminated by the classification of dihedral angles with the 12-divided segments similar to the face of a clock (Figure 1B).[6]

Second, evaluation techniques of structural code homology for X-ray crystallographic data of proteins from the Protein Data Bank (PDB) have been examined.[8] The comparison of PDB IDs 2imm and 2mcp with 99.1% homology of amino acid sequences has indicated that the structural code homology of the main chains using the conformational elements with the 12-divided segments is only 11.5%. To correspond to the polymorphism of molecular structures for proteins, fuzzy search techniques of the structural code homology have also been proposed. In this paper, we present the new method of conformational pattern analysis with these techniques for subunits of proteins in comparison with the DSSP (Dictionary of Secondary Structure in Proteins) method. We also indicate the first finding of supersecondary structure homology between light chains of immunogloblins and MHC classes I and II molecules using the fuzzy search and data mining techniques to extract the specific conformational patterns of proteins. Further, absence of this common conformational pattern in the human IgM rheumatoid factor and correlation of a conformational fragment are revealed in view of a related entropically driven protein−protein interaction.

METHODS

The autoconversion conformation information needed for structural code homology between proteins from X-ray crystallographic PDB data[8] was carried out by the following procedures using a spreadsheet software program: (1) extraction process of four atomic coordinates for determination of dihedral angles from PDB data, (2) conversion process from dihedral angles to conformational codes using dot products and cross products of vectors and plane equations, and (3) strict and/or fuzzy search processes of structural code homology.

In process 1, the same rules previously reported were applied.[6] In the fuzzy search of structural code homology (process 3), it was judged that the homology was high if the conformational elements were included in a range of 90°. Specifically, there were 12 sets, {4α, 2β, 2α}, {2β, 2α, 5β}, {2α, 5β, 5α}, {5β, 5α, 1β}, {5α, 1β, 1α}, {1β, 1α, 6β}, {1α, 6β, 6α}, {6β, 6α, 3β}, {6α, 3β, 3α}, {3β, 3α, 4β}, {3α, 4β, 4α}, and {4β, 4α, 2β}, in a range of 90°. It was supposed that a set of these sets was X = {p, q, r}, and a conformational element, which was compared at a specific angle location, was y. If all elements were satisfied with the equation, $y \in X$, it was judged that the structural code homology of the angle location was high. Finally, the structural code homology of the main chain for each amino acid peptide unit was calculated as the logical conjunction of structural code homology at the angle locations A, B, and C in Figure 1A. In the case of cysteine (Cys), the selected bond at angle location A for determination of dihedral angle based on IUPAC priority rule was just different. Therefore, the comparison of structural code homology for fragments of the main backbone chain between different amino acid sequences except Cys was possible.

The specific structural patterns were extracted by the fuzzy search of structural code homology with the template conformational patterns. The α-helix-type (string "h"), β-sheet-type (string "s"), and other fragment patterns (string "o") were judged by using the template patterns 3β5α4β (α-helix-type) and 6α4β4β (β-sheet-type).

The data mining of subunit homology with similar conformational patterns was carried out by the following procedures with the core programming code: (1) removal of all of the string "o" from the string represented by h-, s-, and o-fragment patterns, defined above, and simultaneous creation of a list (Python term) with the elements of strings composed of combination of h- and s-fragment patterns except o-fragment pattern and (2) comparison of the lists (Python term) between two subunits of proteins using the SequenceMatcher module in the Python Standard Library.[9,10]

The data mining of fragment homology between subunits of proteins, in the case that the characteristic fragment patterns composed of the combination of h- and s-fragment patterns were not acquired in advance, was carried out by the following procedures: (1) selection of the elements of strings with 6 characters or more, (2) matching of the elements of strings between two subunits of proteins, and (3) comparison of the lists (Python term) with the matched elements and above between two subunits of proteins using the SequenceMatcher module in the Python Standard Library.[9,10]

RESULTS AND DISCUSSION

Supersecondary Structure Homology of Immunoglobulins. The conformational code is composed of the

dx.doi.org/10.1021/ci300420d | J. Chem. Inf. Model. 2013, 53, 584−591

combination of the codes of regional angle locations and the 12-divided segments (conformational elements) and is available for the description of the optimized local structures of various compounds using the density functional theory (DFT) calculation. The 12-divided segments were introduced for the discrimination of conformers in the case that two conformations could be optimized in the same region as for *ap* (1) and *sp* (4) by theoretical calculations occasionally.[6] As shown in Figure 2, the structures of light chains of immunogloblins 2imm and

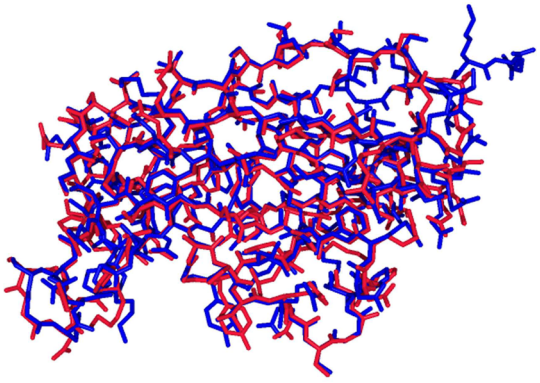


Figure 2. Structural comparison of light chains of immunogloblins 2imm (blue) and 2mcp (red).[8] The structural difference between 2imm and 2mcp with 99.1% homology of amino acid sequences can be described as the symbol by using the conformational code.[10].

2mcp with 99.1% homology of amino acid sequences were similar, but the flexibility of main chain and side chain existed.[8] To describe the difference of the conformations, the conformational code was applied to the X-ray crystal structures of proteins. The comparison of structural code homology of the main chains using the conformational elements with the 12-divided segments extracted the fine conformational difference such as in the case of 2imm and 2mcp where the strict structural code homology was 11.5%.[10] On the other hand, comparison of 2imm and 2mcp, using the fuzzy search of structural code homology in which it was judged that the high homology if the conformational elements were included in a range of 90°, and the structural code homology of the main chain for each amino acid peptide unit that was calculated as the logical conjunction of structural code homology at the angle locations A, B, and C in Figure 1A, indicated that the structural code homology of the main chains was 94.7%.[10] The residual fragments (5.3%) showed the large conformational difference such as the *cis*−*trans* relationship at angle location B of Ser (Figure 3).[10] The advantage of the homology analytical method using conformational code to convert 3D data to 1D data over alternative processing that directly uses the coordinates and implement "coarse" criteria by equivalently relaxed matching ranges is to save the analytical result of conformational fragments to data storage economically.

In the case of cysteine (Cys), the selected bond at angle location A for determination of the dihedral angle based on IUPAC priority rule was just different, and the comparison of structural code homology for fragments of the main backbone chain between different amino acid sequences except Cys was possible.[10] Further, this fuzzy search of structural code homology could also isolate the large structural change of the main chains, but the description was complicated to understand

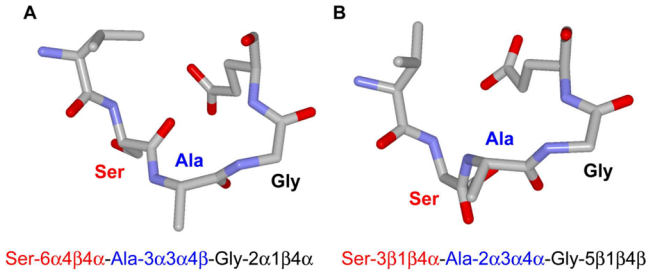


Figure 3. Fragment structures of main chains (14−16 aa.) for light chains of immunogloblins (A) 2imm and (B) 2mcp with same amino acid sequences.[8] The fuzzy search of structural code homology extracted the large conformational difference of Ser and Ala between 2imm and 2mcp.

the difference (Figure 3). Therefore, the simpler description method was explored.

The fragment pattern 3*β*5*α*4*β* in the *α*-helix structure and the pattern 6*α*4*β*4*β* in the *β*-sheet structure for each amino acid peptide unit statistically appeared in the PDB data of immunogloblins (2imm, 2mcp, 1a7o, and 3iy2), insulins (2vk0 and 2c8r), and chaperonin GroEL (1kid, 1jon, and 1srv).[11] The specific structural patterns for each amino acid peptide unit could be extracted by the fuzzy search of structural code homology with the template of conformational patterns (3*β*5*α*4*β* and 6*α*4*β*4*β*) instead of the patterns from the reference subunit of PDB data mentioned above. The fuzzy search of structural code homology for many proteins including immunogloblin, insulin, chaperonin, and MHC molecules using the template patterns 3*β*5*α*4*β* (*α*-helix-type) and 6*α*4*β*4*β* (*β*-sheet-type) for each amino acid peptide unit indicated that most protein structures except *α*-helix and *β*-sheet also consisted of the combination of *α*-helix-type and *β*-sheet-type fragment patterns, and the other patterns were few (Figure 4).[10] This finding became the hint for the simple description method of the conformational difference of main chains. All the conformational patterns of main chain for each amino acid peptide unit could be classified by using *α*-helix-type (pale green: string "h"), *β*-sheet-type (orange: string "s"), and other fragment patterns (white: string "o").

To evaluate the advantage of this new method, the proposed method was compared with the method using string based on secondary structure information predicted by DSSP.[12] The DSSP program is used to standardize secondary structure assignment of proteins widely. Figure 5 indicates the result of conformational pattern analysis with DSSP analysis for light chains of immunogloblins 2imm and 2mcp.[12] The DSSP method assigned the secondary structure including hydrogen-bond well, but some blanks that corresponded to loop or irregular were found (Figure 5). On the other hand, the data of conformational pattern analysis reflected the shape of main chain well. For example, fragment structures of main chains (25−40 aa.) for light chains of 2imm and 2mcp were assigned as the same patterns (string "shsshsshhhossoss") by the conformational pattern analysis, though the assignment of secondary structure predicted by DSSP was different (Figure 6). From this result, it was suggested that the conformational pattern analysis would be useful for the data mining of conformations with the complicated shapes of main chains such as loop structure.

Common Fragment Patterns between MHC Molecules and Light Chain of Immunoglobulins. Sequence-

dx.doi.org/10.1021/ci300420d | J. Chem. Inf. Model. 2013, 53, 584−591

Num.	1wbxA	2w9eL	3hc0L	2agjL	1adqL	2bsrA	2bsrB	1aqdA	1aqdB	1wbxA	2w9eL	3hc0L	2agjL	1adqL	2bsrA	2bsrB	1aqdA	1aqdB
257	TYR	TYR	TYR	TYR	TYR	TYR	TYR	TYR	TYR	6a3a4b	6b3a4a	6b3a4a	6a3a4a	6a3a4a	6a3a4b	6b3a4b	6b3a4b	6a3a4b
258	THR	THR	ALA	ALA	SER	THR	ALA	ASP	THR	6b3a4b	6b3a4b	6b3a4b	6b3a4b	6b3a4b	6b3a4b	1a4b4b	6b4b4b	6b3a4b
259	CYS	CYS	CYS	CYS	CYS	CYS	CYS	CYS	CYS	5a3a4b	5a3a4a	5b3a4b	5a3a4b	5a3b4b	5a3a4a	5b3a4b	5a3a4b	5a3a4b
260	ARG	GLU	GLU	GLU	GLN	HIS	ARG	ARG	GLN	6a3a4a	6b3b4a	6a3a4a	6a3a4a	6a3a4b	6b3a4a	6a3a4a	6a3b4a	6a3a4a
261	VAL	ALA	VAL	VAL	VAL	VAL	VAL	VAL	VAL	6b3a4a	6a3a4b	6a3a4b	6b3a4b	6a3a4b	6a3a4b	6a3a4b	6a3a4b	6a3a4b
262	TYR	THR	THR	THR	THR	GLN	ASN	GLU	GLU	6b3a4a	6a3a4b	6b3a4a	6a3a4b	6a3b4b	6a3a4b	6b3a4a	6a3a4b	6a3a4b
263	HIS	HIS	HIS	HIS	HIS	HIS	HIS	HIS	HIS	1a3a4a	6b4b4b	1a3a4b	6b3a4a	6a6a4b	1a3a4a	1a3a4a	6b3a4a	6b3a4a
264	GLU	LYS	GLN	GLN	GLU	GLU	VAL	TRP	PRO	3a5a4b	3b5a4a	3a5a4a	3a1b4a	2a6a4a	3a1b4a	3a5a4a	3a5a4b	3a1b4b
265	GLY	THR	GLY	GLY	GLY	GLY	THR	GLY	SER	3b1b4a	3a5a4a	3b1b4b	6a1a4b	2a5b4a	3b1b4b	3b1b4a	3b1a4b	3b1a4b
266	LEU	SER	LEU	LEU	SER	LEU	LEU	LEU	VAL	6a3b4b	6a3a4b	6a3a4a	6a3a4b	3b3a4b	6a3a4b	6a3b4b	6a3a4a	6b3a4a
267	PRO	THR	SER	SER	THR	PRO	SER	ASP	THR	3a5a4b	3b1b4a	3b1b4b	3b5a4b	6a3b4b	3b5a4b	3b1b4b	3a5a4b	3b5a4a
268	GLU	SER	SER	SER	VAL	LYS	GLN	GLU	SER	6a3a4b	6b3a4b	6b4b4a	6b4b4b	6a3a4b	6b4b4b	1a4b4b	6b4b4b	6b3a4a
269	PRO	PRO	PRO	PRO	GLU	PRO	PRO	PRO	PRO	3b3a4b	3a3b4a	3b3a4b	3a3a4b	6b3a4b	3b3a4b	3a3a4b	3a3a4a	3a3a4b
270	LEU	ILE	VAL	VAL	LYS	LEU	LYS	LEU	LEU	3b3a4b	3b3a4a	6a3a4b	6a3a4b	6b3a4b	6a4b4b	6a3a4a	6a3a4b	3b4b4b
271	THR	VAL	THR	THR	THR	THR	ILE	LEU	THR	6b3a4a	6b3a4b	6a3a4b	6b3a4b	6a3a4b	6b3a4b	6a3b4a	6a3a4b	6b3a4b
272	LEU	LYS	LYS	LYS	VAL	LEU	VAL	LYS	VAL	6b4b4a	6b3a4a	6a3a4a	6b3a4b	6b3b4b	6b4b4a	6a3b4a	6a3a4a	6b3a4b
273	ARG	SER	SER	SER	ALA	ARG	LYS	HIS	GLU	6b4b4b	6b4b4b	6b4b4b	1a4b4b	6a3a4b	6b4b4b	3b3a4b	6a4b4b	6a4b4b
274	TRP	PHE	PHE	PHE	PRO	TRP	TRP	TRP	TRP	3b3a4b	6b4b4b	6b4b4b	6b4b4b	3a1b4b	3b3b4a	3b3a4b	6b3a4b	6b3a4b
275	GLU	ASN	ASN	ASN	THR	GLU	ASP	GLU	ARG	6a3b4a	6a4b4b	6a3a4b	6a3a4a	1a4b4b	6b3b4b	6b3b4a	6b4b4b	6a3b4b
276	PRO	ARG	ARG	ARG	GLU	PRO	ARG	PHE	ALA	3b1bX	3b6a4b	3a3a4a	3a3a4b	6b5a4a	3b5aX	3a1b4a	3b3a4a	3a3aX
277		ASN	GLY	GLY	CYS		ASP	ASP			6b5a4b	5b1bX	2a5a4a	2a4b4b		6a1a4b	3b2aX	
278		GLU		GLU	SER		MET				6b6aX		3b6b4b	1a4bX		6b1aX		
279				CYS									5b4aX					

Figure 4. Supersecondary structure homology between light chains of immunoglobulins (2w9e: mouse; 3hc0, 2agj, 1adq: human)[14−17] and MHC class I (1wbx: mouse; 2bsr: human)[13,18] and II (1aqd: human)[19] molecules near C-terminal region. Though homology of these amino acid sequences is low (pale purple), the characteristic fragment pattern [263−269 aa. (string "shhshss"), pale green: α-helix-type (string "h"), orange: β-sheet-type (string "s")] is common. As a notable exception, IgM rheumatoid factor (1adq: human) do not have the fragment pattern. Terms "a" and "b" mean "α" and "β" for conformational elements, respectively. Term "X" indicates indeterminable conformational element.

Num.	2imm[a]	2mcp[a]	2imm[b]	2mcp[b]	2imm[c]	2mcp[c]
1	ASP	ASP				
2	ILE	ILE	s	s		
3	VAL	VAL	s	s		
4	MET	MET	s	s	E	E
5	THR	THR	s	s	E	E
6	GLN	GLN	s	s	E	E
7	SER	SER	o	o	E	E
8	PRO	PRO	s	s		
9	SER	SER	h	h	S	S
10	SER	SER	s	s	E	E
11	LEU	LEU	s	s	E	E
12	SER	SER	s	s	E	E
13	VAL	VAL	s	s	E	E
14	SER	SER	s	h		
15	ALA	ALA	s	o	T	S
16	GLY	GLY	o	o	T	S
17	GLU	GLU	s	s		
18	ARG	ARG	s	s		
19	VAL	VAL	s	s	E	E
20	THR	THR	s	s	E	E
21	MET	MET	s	s	E	E
22	SER	SER	s	s	E	E
23	CYS	CYS	o	o	E	E
24	LYS	LYS	s	s	E	E
25	SER	SER	s	s	E	E
26	SER	SER	h	h	S	S
27	GLN	GLN	s	s	S	S
28	SER	SER	s	s		
29	LEU	LEU	h	h		
30	LEU	LEU	s	s	E	B
31	ASN	ASN	s	s	E	
32	SER	SER	h	h	T	S
33	GLY	GLY	h	h	T	S
34	ASN	ASN	h	h	T	S
35	GLN	GLN	o	o	T	
36	LYS	LYS	s	s	E	
37	ASN	ASN	s	s	E	B
38	PHE	PHE	o	o		
39	LEU	LEU	s	s	E	E
40	ALA	ALA	s	s	E	E
41	TRP	TRP	s	s		E
42	TYR	TYR	s	s	E	E
43	GLN	GLN	s	s	E	E
44	GLN	GLN	s	s	E	E
45	LYS	LYS	s	s		
46	PRO	PRO	s	s	T	T
47	GLY	GLY	o	o	T	T
48	GLN	GLN	s	s	S	S
49	PRO	PRO	s	s		
50	PRO	PRO	s	s		
51	LYS	LYS	s	s	E	E
52	LEU	LEU	s	s	E	E
53	LEU	LEU	h	h	E	E
54	ILE	ILE	s	s	E	E
55	TYR	TYR	s	s	E	E
56	GLY	GLY	o	o	T	T
57	ALA	ALA	o	h	T	T
58	SER	SER	h	h	T	T
59	THR	THR	s	s	E	E
60	ARG	ARG	s	s	E	E
61	GLU	GLU	s	s		
62	SER	SER	s	s	T	T
63	GLY	GLY	o	o	T	T
64	VAL	VAL	s	s		S
65	PRO	PRO	s	s		
66	ASP	ASP	h	h	T	T
67	ARG	ARG	h	h	T	T
68	PHE	PHE	s	s	E	E
69	THR	THR	s	s	E	E
70	GLY	GLY	s	s	E	E
71	SER	SER	s	s	E	E
72	GLY	GLY	o	o	E	E
73	SER	SER	s	s	E	E
74	GLY	GLY	o	o	T	T
75	THR	THR	h	h	T	T
76	ASP	ASP	s	s	E	E
77	PHE	PHE	s	s	E	E
78	THR	THR	s	s	E	E
79	LEU	LEU	s	s	E	E
80	THR	THR	s	s	E	E
81	ILE	ILE	s	s	E	E
82	SER	SER	h	h	S	S
83	SER	SER	s	o	S	S
84	VAL	VAL	s	s		
85	GLN	GLN	s	s		
86	ALA	ALA	h	h	G	T
87	GLU	GLU	h	h	G	T
88	ASP	ASP	h	h	G	
89	LEU	LEU	s	s		
90	ALA	ALA	o	s	E	E
91	VAL	VAL	s	s	E	E
92	TYR	TYR	s	s	E	E
93	TYR	TYR	s	s	E	E
94	CYS	CYS	o	o	E	E
95	GLN	GLN	s	s	E	E
96	ASN	ASN	s	s	E	E
97	ASP	ASP	h	h		
98	HIS	HIS	h	h	S	S
99	SER	SER	s	s	S	S
100	TYR	TYR	o	o	S	S
101	PRO	PRO	s	s	S	S
102	LEU	LEU	s	s		
103	THR	THR	s	s	E	E
104	PHE	PHE	s	s	E	E
105	GLY	GLY	s	s		
106	ALA	ALA	h	h		
107	GLY	GLY	o	o		
108	THR	THR	s	s	E	
109	LYS	LYS	s	s	E	E
110	LEU	LEU	s	s	E	E
111	GLU	GLU	s	s	E	E
112	LEU	ILE	s	s	E	E
113	LYS	LYS	s	s		
114	ARG	ARG	o	s		

Figure 5. Comparison of conformational pattern analysis with DSSP analysis for light chains of immunoglobulins 2imm and 2mcp.[12] All amino acid peptide units are finely assigned using the conformational pattern analysis. [a]Amino acid sequence. [b]Conformational pattern analysis (h: α-helix-type, s: β-sheet-type, and o: other-type). [c]DSSP [H: α-helix, E: extended strand, participates in beta ladder, B: residue in isolated β-bridge, G: 3-helix (3/10 helix), I: 5 helix (π-helix), T: hydrogen bonded turn, and S: bend].

Matcher module in the Python Standard Library[9] is available for comparing pairs of sequences of various types and for finding homology between two strings. For example, the ratio method (Python term) returns a measure of the sequences' similarity "0.866" between two strings "private Thread currentThread" and "private volatile Thread currentThread." using the SequenceMatcher module.[9] This method was applied to the strings represented by h-, s-, and o-fragment patterns,

dx.doi.org/10.1021/ci300420d | J. Chem. Inf. Model. 2013, 53, 584−591

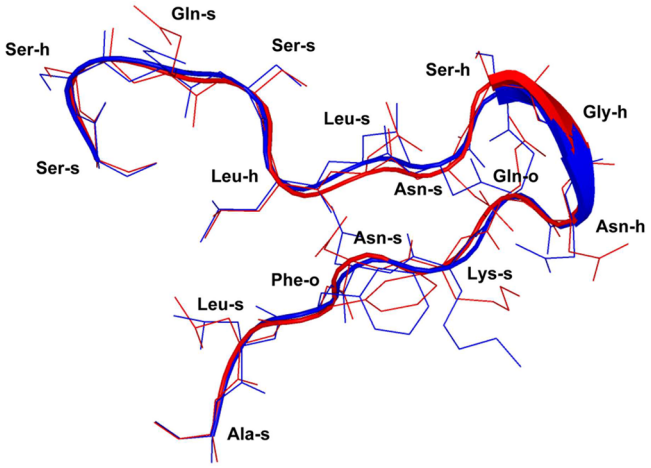


Figure 6. Comparison of fragment structures of main chains (25−40 aa.) for light chains of immunogloblins 2imm (blue) and 2mcp (red).[8] Though the assignment of secondary structure predicted by DSSP is different, the fragments of main chains between 2imm and 2mcp are characterized as the same patterns (string "shsshsshhhossoss") by the conformational pattern analysis.

defined above, such as string "shsshsshhhossoss" to measure the supersecondary structure homology between subunits of proteins. However, the direct comparison of two strings from very similar α-chains of MHC class I molecules 1wbx (mouse) and 1wby (mouse) indicated only the sequences' similarity value of 0.15.[13] One of the reasons for the low value was suggested that not all atomic coordinates for X-ray crystallographic data of proteins are often determinable in general. To correspond to such a situation, all of string "o" was removed from the string represented by h-, s-, and o-fragment patterns, and a list (Python term) with the elements of strings composed of combination of h- and s-fragment patterns except the o-fragment pattern was created simultaneously. The SequenceMatcher module was applied to the lists (Python term) for subunits of proteins. The quantification of supersecondary structure homology of MHC class I molecule (1wbx: mouse, α-chain)[13] between subunits of proteins using this procedure suggested that the correlation of similar α-chains of MHC class I molecules was relatively high (1wby: 0.84, 2cik: 0.73, and 2bsr: 0.64) but not the other subunit type of proteins (Table 1). The quantification of supersecondary structure homology of the light chain of immunoglobulin (2w9e: mouse)[14] between subunits of proteins (516 subunits) also indicated the correlation of similar light chains of immunoglobulins with homology values of 0.69−0.32.[10] The data mining technique could roughly quantify the supersecondary structure homology between any subunits of proteins.

Further, the data mining technique to find the fragment homology between subunits of proteins, in the case that the characteristic fragment patterns composed of the combination of h- and s-fragment patterns were not acquired in advance, was developed. The data mining was carried out by the following procedures: (1) selection of the elements of strings with six characters or more, (2) matching of the elements of strings between two subunits of proteins, and (3) comparison of the lists (Python term) with the matched elements and above between two subunits of proteins using the SequenceMatcher module.

The quantification of fragment homology of MHC class I molecule (1wbx: mouse, α-chain)[13] between subunits of proteins using this data mining technique extracted the correlation between fragments of MHC class I molecule[13] and the light chain of immunoglobulin[14] (Table 1). The common conformational fragments of main chains (red) in Figure 7 were located in the constant domains. In these common fragments, the characteristic fragment patterns (string "shhshss"), which are combined with α-helix-type and β-sheet-type patterns and are sandwiched between two identical amino acid sequences His and Pro, were also extracted in light chains of other types of immunoglobulins,[14−17] α-chain and β-2 microglobulin of MHC class I,[13,18] and α-chain and β-chain of MHC class II[19] (Figure 4). The characteristic fragment patterns (string "shhshss") were represented as several different descriptions "XTTXSSX" (1wbx_A), "XSSXSSX" (2w9e_L), and "XTTSSSX" (2agj_L) using the DSSP method (T: hydrogen bonded turn, S: bend, and X: blank).[12] On the other hand, heavy chains of Fab immunoglobulin fragments[14−17] and T cell receptors (TCR)[20] did not have these conformational fragments.[10] The common protruding structural regions of the characteristic fragment patterns suggested some kind of ligand interaction (Figure 8).

The cluster of differentiation 4 (CD4)[21] and CD8[22] that serves as coreceptors for the TCRs did not interact with the conformational fragments.[10] Although, killer-cell immunoglobulin-like receptors (KIR),[23] which are inhibitory receptors on the surface of natural killer cells that also interact with MHC class I molecules, did not contact with the conformational fragments.[10] On the other hand, leukocyte immunoglobulin-like receptors (LILR, immunoglobulin-like transcripts,

Table 1. Quantification of Supersecondary Structure Homology of MHC Class I Molecule (1wbx: mouse, α-chain)[13] between Subunits of Proteins and between Fragments[8]

PDB ID	protein	subunit homology	fragment homology
1wbxA	MHC class I (α chain)	1.00	1.00
1wbyA	MHC class I (α chain)	0.84	1.00
2cikA	MHC class I (α chain)	0.73	1.00
2bsrA	MHC class I (α chain)	0.64	1.00
2w9eL	immunogloblin (light chain)	0.31	0.80
1kn2L	abzyme (light chain)	0.32	0.38
2mcpH	immunogloblin (heavy chain)	0.33	0.33
2khmB	fibroin	0.29	0.29
1aqdA	HLA class II (α chain)	0.29	0.29
1kidA	GroEL	0.09	0.27
3e20A	SUP35	0.26	0.26
1gzmA	rhodopsin	0.24	0.26
1bk9A	PBP	0.25	0.25
1etsH	thrombin	0.25	0.25
2bimA	P53	0.14	0.25
2vb1A	lysozyme C	0.25	0.25
1o9kP	E2F-1	0.24	0.24
1n5oX	BRCA1	0.23	0.23
2vk0A	insulin (A chain)	0.23	0.23
1alcA	α-lactalbumin	0.22	0.22
1o9kA	Rb	0.22	0.22
1gqeA	RF2	0.21	0.21
2onjA	Sav1866	0.17	0.17
2wrnZ	EF-Tu	0.16	0.16
2w9eA	prion	0.11	0.11

dx.doi.org/10.1021/ci300420d | J. Chem. Inf. Model. 2013, 53, 584−591

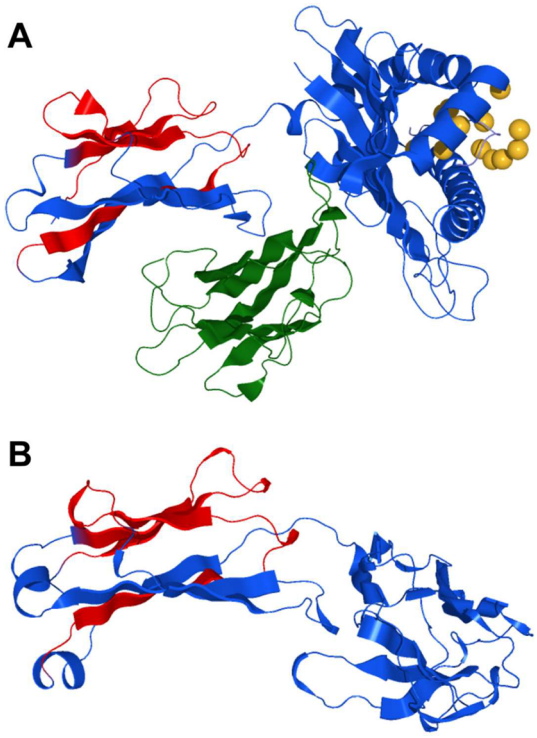


Figure 7. Common conformational fragments of main chains (red) between α-chain of MHC class I (1wbx: mouse)[13] and light chain of immunoglobulin (2w9e: mouse).[14] (A) α-chain (blue) and β-2 microglobulin (green) for MHC class I molecule and (B) light chain of immunoglobulin (blue).[8] Though amino acid sequences of these proteins were completely different, the data mining of fragment patterns extracted common structures of the main chains.

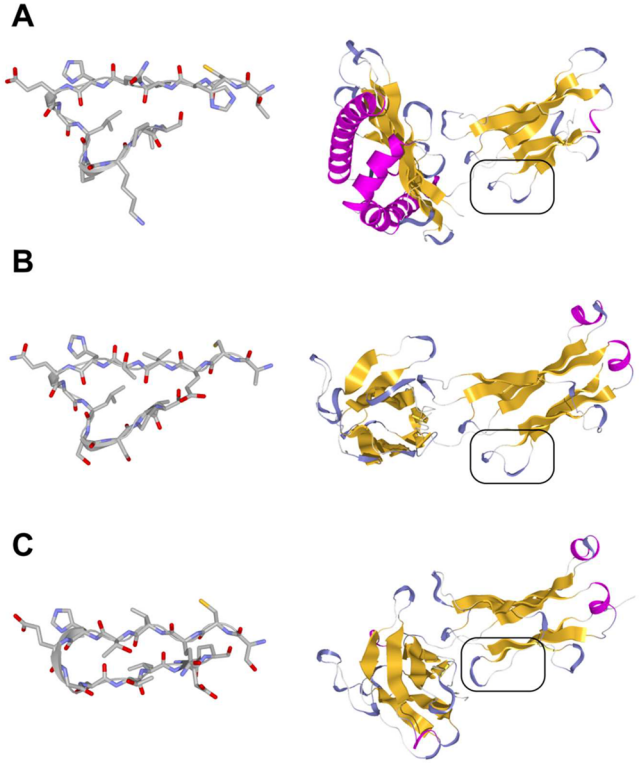


Figure 8. Fragment structures of main chains (258−270 aa.) for α-chain of MHC class I and light chains of immunoglobulins.[8] (A) MHC class I (2bsr: human),[18] (B) IgG immunoglobulin (3hc0: human),[15] and (C) IgM rheumatoid factor (1adq: human).[17] Characteristic fragment patterns (string "shhshss") sandwiched between two identical amino acid sequences His and Pro are protruded on the molecular surfaces. Of particular interest, the IgM rheumatoid factor loses the fragment pattern (string "ooossss").

CD85),[24−26] which also bind to MHC class I molecules and transduce a negative signal that inhibits stimulation of an immune response, connected with the conformational fragments (string "shhshss") in β-2 microglobulins as a type of pair forms (string "sohsss") (Figure 9). Before and after the interaction, the conformational fragment patterns of the main chains were unchanging, and it was suggested that the pair interaction contributed to the entropically driven recognition of LILR.[26] Interestingly, the structure of LILR (2dyp) moderately correlated with that of KIR (1efx) (fragment homology value: 0.67), and LILR and KIR contained the other type of characteristic common fragment patterns (string "sshhohhss") with different amino acid sequences.[10]

Further, human IgM rheumatoid factor,[17] one of the immunoglobulins, did not, in particular, have the conformational fragment pattern (string "ooossss") (Figure 8). Nonclassic MHC class I molecules CD1D,[27] MHC class I-related chain A (MIC-A), and MIC-B,[28] which have functions to activate NKT, NK, and T cells, did not also clearly show the patterns, though human leukocyte antigen E (HLA-E),[29] HLA-G,[24] hemochromatosis protein (HFE),[30] Zn-α2-glycoprotein (ZAG),[31] and neonatal Fc receptor of IgG (FCRN)[32] have the patterns.[10] UL18 protein,[25] which is encoded in the genome of human cytomegalovirus, indicated the similar patterns to MIC-A and MIC-B (string "soossss") with different amino acid sequences.[10] The loop structures are known to be highly flexible in general. The fuzzy search and data mining techniques taken into account thermal factor and other dynamical information available from molecular dynamics simulation would be efficient for the elucidation of biological function of such conformational fragments in future.

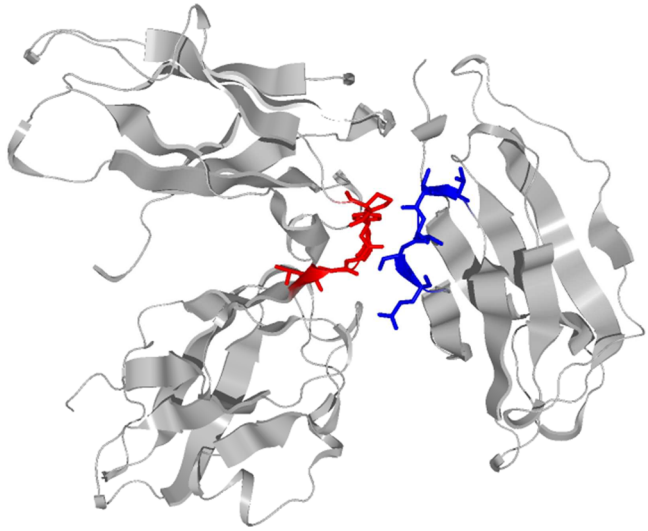


Figure 9. Interaction of LILR with β-2 microglobulin of MHC class I (2dyp: human).[24] The unchanging conformational fragment (red: string "sohsss") of main chain in LILR connected with the conformational fragment (blue: string "shhshss") in β-2 microglobulin as a type of pair forms.

dx.doi.org/10.1021/ci300420d | J. Chem. Inf. Model. 2013, 53, 584−591

CONCLUSION

In the present study, we showed that the PDB protein structures for conformational similarity rather than only for sequence similarity can be mined by the encoding techniques of conformational structures. The new method of conformational pattern analysis would be useful for the data mining of conformations with the complicated shapes of main chains such as loop structure because the method reflected the shape of main chain well. We also indicated the ligand interaction of closely related proteins with the conformation-similar fragments even though the primary structures show only moderate levels of correlation. The fuzzy search and data mining techniques of supersecondary structure homology using the new conformational code are available for the classification of confusing polymorphisms of amino acid sequences and molecular structures for proteins such as MHC I and II molecules and immunogloblins. Further, these code-driven mining techniques can be utilized as a metadata-generating tool for systems biology because MHC I and II molecules contact with various signal ligands on the surface of T cells and natural killer cells.

ASSOCIATED CONTENT

Supporting Information

Core programming code, comparison and quantification of supersecondary structure homology, interactions of characteristic conformational fragments, and comparison of conformational fragments. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

* E-mail: izumi.h@aist.go.jp.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was partly supported by JSPS KAKENHI Grant Number 23615012. We thank reviewers for their helpful comments.

REFERENCES

(1) Fine, J. H.; Chen, P.; Mesci, A.; Allan, D. S. J.; Gasser, S.; Raulet, D. H.; Carlyle, J. R. Chemotherapy-induced genotoxic stress promotes sensitivity to natural killer cell cytotoxicity by enabling missing-self recognition. *Cancer Res.* **2010**, *70*, 7102−7113.

(2) Wang, B.; Primeau, T. M.; Myers, N.; Rohrs, H. W.; Gross, M. L.; Lybarger, L.; Hansen, T. H.; Connolly, J. M. A single peptide−MHC complex positively selects a diverse and specific CD8 T cell repertoire. *Science* **2009**, *326*, 871−874.

(3) Muenz, C.; Luenemann, J. D.; Getts, M. T.; Miller, S. D. Antiviral immune responses: Triggers of or triggered by autoimmunity? *Nat. Rev. Immunol.* **2009**, *9*, 246−258.

(4) Izumi, H.; Yamagami, S.; Futamura, S.; Nafie, L. A.; Dukor, R. K. Direct observation of odd−even effect for chiral alkyl alcohols in solution using vibrational circular dichroism spectroscopy. *J. Am. Chem. Soc.* **2004**, *126*, 194−198.

(5) Izumi, H.; Ogata, A.; Nafie, L. A.; Dukor, R. K. Vibrational circular dichroism analysis reveals a conformational change of the baccatin III ring of paclitaxel: Visualization of conformations using a new code for structure-activity relationships. *J. Org. Chem.* **2008**, *73*, 2367−2372.

(6) Izumi, H.; Ogata, A.; Nafie, L. A.; Dukor, R. K. A revised conformational code for the exhaustive analysis of conformers with one-to-one correspondence between conformation and code: Application to the VCD analysis of (S)-ibuprofen. *J. Org. Chem.* **2009**, *74*, 1231−1236.

(7) Izumi, H.; Ogata, A.; Nafie, L. A.; Dukor, R. K. Structural determination of molecular stereochemistry using VCD spectroscopy and a conformational code: Absolute configuration and solution conformation of a chiral liquid pesticide, (R)-(+)-malathion. *Chirality* **2009**, *21*, E172−E180.

(8) *PDBj*. Protein Data Bank Japan. http://www.pdbj.org/index_j.html (accessed November 9, 2012).

(9) 7.4. difflib — Helpers for computing deltas. http://docs.python.org/library/difflib.html (accessed November 9, 2012).

(10) Core programming code, comparison and quantification of supersecondary structure homology, interactions of characteristic conformational fragments, and comparison of conformational fragments are shown in the Supporting Information.

(11) Izumi, H.; Ogata, A. Conformation Homology Evaluation Method/Apparatus and Structure Pattern Analysis Method/Apparatus. JP 2011193868, October 6, 2011.

(12) Kabsch, W.; Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**, *22*, 2577−2637.

(13) Meijers, R.; Lai, C.; Yang, Y.; Liu, J.; Zhong, W.; Wang, J.; Reinherz, E. L. 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.* **2005**, *345*, 1099−1110.

(14) Antonyuk, S. V.; Trevitt, C. R.; Strange, R. W.; Jackson, G. S.; Sangar, D.; Batchelor, M.; Cooper, S.; Fraser, C.; Jones, S.; Georgiou, T.; Khalili-Shirazi, A.; Clarke, A. R.; Hasnain, S. S.; Collinge, J. Crystal structure of human prion protein bound to a therapeutic antibody. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 2554−2558.

(15) Jordan, J. L.; Arndt, J. W.; Hanf, K.; Li, G.; Hall, J.; Demarest, S.; Huang, F.; Wu, X.; Miller, B.; Glaser, S.; Fernandez, E. J.; Wang, D.; Lugovskoy, A. Structural understanding of stabilization patterns in engineered bispecific Ig-like antibody molecules. *Proteins* **2009**, *77*, 832−841.

(16) Ramsland, P. A.; Terzyan, S. S.; Cloud, G.; Bourne, C. R.; Farrugia, W.; Tribbick, G.; Geysen, H. M.; Moomaw, C. R.; Slaughter, C. A.; Edmundson, A. B. Crystal structure of a glycosylated Fab from an IgM cryoglobulin with properties of a natural proteolytic antibody. *Biochem. J.* **2006**, *395*, 473−481.

(17) Corper, A. L.; Sohi, M. K.; Bonagura, V. R.; Steinitz, M.; Jefferis, R.; Feinstein, A.; Beale, D.; Taussig, M. J.; Sutton, B. J. Structure of human IgM rheumatoid factor Fab bound to its autoantigen IgG Fc reveals a novel topology of antibody-antigen interaction. *Nat. Struct. Biol.* **1997**, *4*, 374−381.

(18) Stewart-Jones, G. B. E.; Di Gleria, K.; Kollnberger, S.; McMichael, A. J.; Jones, E. Y.; Bowness, P. Crystal structures and KIR3DL1 recognition of three immunodominant viral peptides complexed to HLA-B*2705. *Eur. J. Immunol.* **2005**, *35*, 341−351.

(19) Murthy, V. L.; Stern, L. J. The class II MHC protein HLA-DR1 in complex with an endogenous peptide: Implications for the structural basis of the specificity of peptide binding. *Structure* **1997**, *5*, 1385−1396.

(20) Garcia, K. C.; Degano, M.; Pease, L. R.; Huang, M.; Peterson, P. A.; Teyton, L.; Wilson, I. A. Structural basis of plasticity in T cell receptor recognition of a self peptide MHC antigen. *Science* **1998**, *279*, 1166−1172.

(21) Wang, J. H.; Meijers, R.; Xiong, Y.; Liu, J. H.; Sakihama, T.; Zhang, R.; Joachimiak, A.; Reinherz, E. L. Crystal structure of the human CD4 N-terminal two-domain fragment complexed to a class II MHC molecule. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 10799−10804.

(22) Gao, G. F.; Tormo, J.; Gerth, U. C.; Wyer, J. R.; McMichael, A. J.; Stuart, D. I.; Bell, J. I.; Jones, E. Y.; Jakobsen, B. K. Crystal structure of the complex between human CD8 alpha alpha and HLA-A2. *Nature* **1997**, *387*, 630−634.

(23) Boyington, J. C.; Motyka, S. A.; Schuck, P.; Brooks, A. G.; Sun, P. D. Crystal structure of an NK cell immunoglobulin-like receptor in complex with its class I MHC ligand. *Nature* **2000**, *405*, 537−543.
(24) Shiroishi, M.; Kuroki, K.; Rasubala, L.; Tsumoto, K.; Kumagai, I.; Kurimoto, E.; Kato, K.; Kohda, D.; Maenaka, K. Structural basis for recognition of the nonclassical MHC molecule HLA-G by the leukocyte Ig-like receptor B2 (LILRB2/LIR2/ILT4/CD85d). *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 16412−16417.
(25) Yang, Z.; Bjorkman, P. J. Structure of UL18, a peptide-binding viral MHC mimic, bound to a host inhibitory receptor. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 10095−10100.
(26) Shiroishi, M.; Kuroki, K.; Tsumoto, K.; Yokota, A.; Sasaki, T.; Amano, K.; Shimojima, T.; Shirakihara, Y.; Rasubala, L.; van der Merwe, P. A.; Kumagai, I.; Kohda, D.; Maenaka, K. Entropically driven MHC class I recognition by human inhibitory receptor leukocyte Ig-like receptor B1 (LILRB1/ILT2/CD85j). *J. Mol. Biol.* **2006**, *355*, 237−248.
(27) Borg, N. A.; Wun, K. S.; Kjer-Nielsen, L.; Wilce, M. C. J.; Pellicci, D. G.; Koh, R.; Besra, G. S.; Bharadwaj, M.; Godfrey, D. I.; McCluskey, J.; Rossjohn, J. CD1d-lipid-antigen recognition by the semi-invariant NKT T-cell receptor. *Nature* **2007**, *448*, 44−49.
(28) Bauer, S.; Groh, V.; Wu, J.; Steinle, A.; Phillips, J. H.; Lanier, L. L.; Spies, T. Activation of NK cells and T cells by NKG2D, a receptor for stress-inducible MICA. *Science* **1999**, *285*, 727−729.
(29) Sullivan, L. C.; Hoare, H. L.; McCluskey, J.; Rossjohn, J.; Brooks, A. G. A structural perspective on MHC class lb molecules in adaptive immunity. *Trends Immunol.* **2006**, *27*, 413−420.
(30) Lebrón, J. A.; Bennett, M. J.; Vaughn, D. E.; Chirino, A. J.; Snow, P. M.; Mintier, G. A.; Feder, J. N.; Bjorkman, P. J. Crystal structure of the hemochromatosis protein HFE and characterization of its interaction with transferrin receptor. *Cell* **1998**, *93*, 111−123.
(31) Hassan, M. I.; Bilgrami, S.; Kumar, V.; Singh, N.; Yadav, S.; Kaur, P.; Singh, T. P. Crystal structure of the novel complex formed between zinc alpha 2-glycoprotein (ZAG) and prolactin-inducible protein (PIP) from human seminal plasma. *J. Mol. Biol.* **2008**, *384*, 663−672.
(32) Ye, L.; Liu, X.; Rout, S. N.; Li, Z.; Yan, Y.; Lu, L.; Kamala, T.; Nanda, N. K.; Song, W.; Samal, S. K.; Zhu, X. The MHC class II-associated invariant chain interacts with the neonatal Fc gamma receptor and modulates its trafficking to endosomal/lysosomal compartments. *J. Immunol.* **2008**, *181*, 2572−2585.