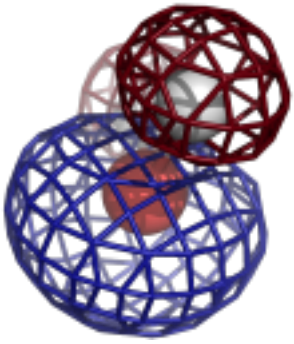
VOLUME FOUR



COMPUTATIONAL CRYSTALLOGRAPHY NEWSLETTER

JULY MMXIII

ENSEMBLE REFINEMENT, CABLAM

Table of Contents

- PHENIX News 28
- Crystallographic meetings 29
- Expert Advice
  - Fitting tips #6 - Potential misfitting by a switch of sidechain vs mainchain 30
- FAQ 32
- Short Communications
  - *CaBLAM*: Identification and scoring of disguised secondary structure at low resolution 33
  - Structural Classification of Allergen IgE Epitopes by Hierarchical Clustering 36
  - New Tool: *phenix.real_space_refine* 43
- Articles
  - *cctbx* tools for transparent parallel job execution in Python. III. Remote access 45
  - *phenix.ensemble_refinement*: a test study of apo and holo BACE1 51

Editor

Nigel W. Moriarty, NWMoriarty@LBL.Gov

PHENIX News

New programs

FEM: Feature Enhanced Maps (Pavel V. Afonine)

Interpretation of a crystallographic map is a means of obtaining an atomic representation of a crystal structure or the map itself may serve as the crystal model. There are number of factors that affect quality of crystallographic maps that in turn affect difficulty (or even feasibility) of their interpretation and quality of resulting model of crystal structure, and include:

- finite resolution of measured reflections;
- incompleteness of data (missing reflections within the resolution range of the measured data);
- experimental errors in measured reflections;
- errors in atomic model parameters.

These factors a) result in artificial peaks in the map that may be confused with the signal and therefore erroneously interpreted in terms of atomic model, b) introduce noise that may obscure the signal and c) may distort the signal in various ways.

Another fundamentally different contributor to the difficulty of map interpretation is that not all the signal has the same strength. For example, a strong signal arising from a heavy atom derivative may easily obscure a very weak signal (that may be at or below the noise level) arising from a partially occupied very mobile ligand or residue side chain alternative conformation or even hydrogen atoms.

Structural Classification of Allergen IgE Epitopes by Hierarchical Clustering

Naveen Chakicherla

Dougherty Valley High School, San Ramon, CA 94582; Intern, Lawrence Berkeley National Laboratory, Berkeley, CA 94720.

Correspondence email: naveenchaki@gmail.com

Abstract

Allergen epitopes derived from the Structural Database of Allergenic Proteins (SDAP) were modeled and classified structurally using single linkage hierarchical clustering. The largest single cluster in this study contained 231 epitopes representing 12 of the 16 species studied. Cluster analysis of 3D models generated for these epitopes revealed that, while sequence identity between the allergen epitopes was generally very low, epitopes in this cluster shared a common tertiary structure in the form of the W-shaped motif with a consensus sequence `CR+AKA--SK+SG`. The putative shared tertiary structural motif in the epitopes in the largest cluster could also contribute to the significant amount of cross-reactivity clinically observed in several allergens.

Introduction

Over the past decade, there has been a dramatic increase in both incidence and knowledge gain in the area of allergy medicine [1, 2]. The sequences of over 100 food allergens have been published and have been compiled into various public databases [2, 3]. Yet effective treatment of allergic response is still not established, prevention (avoidance) being the most touted cure. Surprisingly, analyses of plant food allergens show that most belong to only a small number of protein superfamilies. Most plant allergens belong to one of only four protein superfamilies, namely prolamins, cupins, profilins, and the Bet V1 family [4, 5]. The World Health Organization/Food and Agriculture Organization (WHO/FAO) guidelines define a protein as potentially allergenic if it either has an identity of at least six contiguous amino acids or a minimum of 35% sequence similarity over a window of 80 amino acids when compared with known allergens [6]. The epitopes, or antigenic determinant of protein antigens, are categorized as linear or conformational epitopes [7], which interact with the paratope based on linear sequence or 3-dimensional structure respectively. Conformational epitopes are stretches of amino acids that interact with an antigen and they are recognized on the basis of their 3D structure. Linear epitopes are recognized on the basis of their amino acid sequence. Although most epitopes are of the conformational kind, the allergen classification has been largely based on sequence similarity, and cross-reactivity predictions based on the WHO/FAO rules result in large numbers of false positives. Furthermore, sequence-based detection could miss several true positives. This study proposes the application of a clustering method to the 3-dimensional structures of allergenic proteins in order to better define the characteristics of the IgE epitopes of allergens in general.

Methods

Sequences of allergenic proteins with known and defined epitopes were obtained from the Structural Database of Allergenic Proteins (SDAP) [3]. In total, 41 full-length sequences of 29 different proteins were submitted to the homology modeling server AS2TS [8], resulting in 35 successful model building runs yielding five putative models for each sequence. Structures of epitopes, as listed in the SDAP, were extracted from these models using JMol [9].

In all, 1655 epitope models were generated, of which 170 models were of t-cell epitopes, 80 models of IgG epitopes and the remaining 1405 were of IgE epitopes. Of the total 1655 models, 446 had a file size of zero, since the segment of the allergen containing that epitope had not been modeled in AS2TS. A structural comparison between all remaining 1209 3D models was performed with SUBCOMB [10] with normalized spatial discrepancy (NSD). The 728424 structure superpositioning jobs were carried out using 30 2.4 GHz processors simultaneously, resulting in 1.3 CPU weeks in total, with a wall clock time of 7.5 hours. The resulting distance matrix was used to perform a

Table 1: Allergen types with known epitopes that were downloaded from SDAP and their representation in the largest cluster in the study.

Allergen	Species - Scientific Name	Species - Common Name	In Cluster 282
Ara h 1; Ara h 2.0101; Ara h 3	*Arachis hypogaea*	Peanut	Y
Asp f 1; Asp f 13; Asp f 2; Asp f 3	*Aspergillus fumigatus*	Fungi	Y
Bet v 1	*Betula verrucosa (Betula pendula)*	White birch	Y
Cha o 1	*Chamaecyparis obtusa*	Japanese Cypress	N
Cry j 1; Cry j 2;	*Cryptomeria japonica*	Japanese Cedar, sugi	Y
Jun a 1.010101; Jun a 3	*Juniperus ashei*	Mountain Cedar	N
Fag e 1;	*Fagopyrum esculentum*	Common Buckwheat	Y
Gly m glycinin G1; Gly m glycinin G2	*Glycine max*	Soybean	Y
Hev b 1; Hev b 3; Hev b 5	*Hevea brasiliensis*	Rubber (latex)	Y
Jug r 1	*Juglans regia*	English walnut	Y
Len c 1.0101	*Lens culinaris*	Lentil	N
Par j 1; Par j 2	*Parietaria judaica*	Pellitory-of-the-Wall	Y
Pen a 1	Penaeus aztecus	Shrimp	Y
Ves v 5	*Vespula vulgaris*	Yellow Jacket	N
Gal d 1	*Gallus domesticus*	Chicken	Y
Pen ch 18	*Penicillium chrysogenum (formerly P. notatum)*	Fungi	Y

cluster analysis with single linkage hierarchical clustering as available in the CCTBX [11].

The 19 clusters with ten or more members were further analyzed for composition, sequence identity, and other characteristics. The epitopes included in the largest cluster, cluster 282 were further analyzed. In order to compare their primary structures the PDB structures of the epitopes needed to be converted back to a primary sequence format. A custom application written by Sebastian Raschka, Michigan State University was used in this study to convert pdb files to fasta sequence format [12]. This custom tool takes a folder of multiple PDB files and writes the fasta sequences for all PDBs into one file that is placed in the same folder. Multiple alignment of the representative members of the largest cluster 282 was conducted using Clustal Omega [13]. In order to compare them structurally, the tool MAPSCI [17] was used.

Results

In this study, we analyzed the 29 allergens with epitopes that are to be found in the SDAP allergen database. These 29 allergens with epitopes originate from 16 distinct plant and animal species (table 1). Three-dimensional structural models were generated for the epitopes of these 29 allergens, resulting in 1209 3D models. Single linkage hierarchical clustering of the epitope structures resulted in a distribution of clusters as depicted in table 2. Clusters varied widely in size,

Table 2: Cluster distribution of the 270 clusters resulting from single linkage hierarchical clustering of structural models of all epitopes of the modeled allergens from SDAP.

Cluster Size	Occurrence
231	1
38	1
20	1
15	3
12	1
11	1
10	11
9	1
8	4
7	1
6	2
5	45
4	37
3	25
2	79
1	77

with the largest cluster having 231 members, while the smallest had one. There were 77 clusters with only one member, and these were eliminated. Clusters with ten or more epitope members were further analyzed. Of the 11 clusters with ten members each, nine clusters contained epitopes of the allergen Aspf1 from the fungi *Aspergillus fumigatus*, while the other two contain epitopes of the allergen, lipid transfer protein; P2 from the weed *Parietaria judaica*. The single cluster with 11 members had epitopes from the two allergens Pench18 and Hev b5 from the fungus *Penicillium chrysogenum* and the kiwi and potato homolog from rubber, *Hevea brasiliensis* respectively. The single cluster with 12 members included epitopes Ara h3 and Gly m glycinin G2 from peanut *Arachis hypogea*, and soybean *Glycine max* respectively. The three clusters with 15 members each all contained epitopes from the fungus *Aspergillus fumigatus.* The single cluster with 20 members had epitopes from the allergen Ara h 2 from peanut *Arachis hypogea* and from the allergen Fag e 1, from the common buckwheat *Fagopyrum esculentum*. The largest cluster containing 231 members is described further below.

Discussion

The majority of the clusters in this structural classification study were composed of epitopes from within the same species, some from different parts of the same protein model. This result suggests that epitopes within allergens are repeated at separate locations in the tertiary structure. However, there were some interesting clusters that showed membership from more than one species. The largest cluster, 282, with 231 members, was analyzed further. Of the 16 allergen species in the entire data set studied, epitopes of allergens from 12 species were included in this largest cluster. They included epitopes from *Arachis hypogaea (*peanut*), Aspergillus fumigatus (*fungus*), Betula verrucosa or Betula pendula (*white birch*), Cryptomeria japonica* (Japanese cedar, sugi), *Fagopyrum esculentum* (common buckwheat), *Glycine max* (soybean), *Hevea brasiliensis* (rubber (latex)), *Juglans regia* (English walnut),*Parietaria judaica* (Pellitory-of-the-Wall), *Penaeus aztecus* (shrimp), *Gallus domesticus* (chicken) and *Penicillium chrysogenum* formerly *P. notatum* (fungus). The four species not represented in this cluster included *Chamaecyparis obtuse* (Japanese Cypress), *Juniperus ashei* (mountain cedar), *Lens culinaris* (lentil) and *Vespula vulgaris* (yellow jacket).

The absence of the mountain cedar epitopes and the yellow jacket epitopes in cluster 282 is easily explained, because AS2TS failed to identify structural models in RCSB for the allergens from these two species. Their epitopes were therefore not represented in our final cluster analysis. The Japanese cypress allergen epitopes had 50 models generated in our analysis. However, it should be noted that the epitope in all cases of this allergen were classified by SDAP as t-cell epitopes that presumably has a different structure than the IgE type epitope that is prevalent in all other epitopes in the cluster 282 (the largest cluster). Hong *et. al.* have shown that pepsin digestion eliminates IgE reactivity but maintains T-cell reactivity suggesting that the two epitopes have different structure and or location in the allergen [16]. The lentil allergen did in fact have 12 models generated by AS2TS, one set modeled using the A chain of the adzuki bean 7S globulin 1 protein structure 2ea7-a and the second set modeled using the PDB structure of a hypothetical Coenzyme A-binding protein from *Thermus thermophiles* bacterium, 1iuk-a.

Sequence analysis of all epitopes in cluster 282 using Clustal Omega gave an alignment length of 30 and an average identity of 7.27% (for default settings) and an alignment length of 25 and an alignment length of 24 and average identity of 8.53% for two HMM iteration and one guide-tree iteration. When all FASTA sequences with exact sequence matches to other FASTA from the same allergen and epitope were removed, as were 1-mer epitopes, this resulted in a smaller cluster of 53 unique sequences. This set gave similar results as the entire set with average length of 6.9 residues, an alignment length of 23 and an average identity of 6.17%.

However, multiple structural analysis of essentially the same set of epitopes from cluster 282 using MAPSCI [17] (epitope PDBs with fewer than four C-alphas were removed, resulting in 45 PDBs) showed a consensus motif in several members of this cluster (See figure 1 and table 3). This motif consisted of two parts with a consensus sequence of `CR+AKA- -SK+SG`. The 3D structural visualization of these epitopes reveals a

Table 3: Summary of cluster 282 epitopes analyzed using MAPSCI.

MAPSCI code	Allergen	SDAP Source Number	AS2TS model number	PDB template	PDB template chain	Epitope type	Epitope start residue	Epitope stop residue
uMS01	arah1	p43237	1	1iuk	a	IgE	311	320
uMS02	arah1	p43237	1	1iuk	a	IgE	559	568
uMS04	arah1	p43237	3	3s7i	a	IgE	578	587
uMS05	arah1	p43237	5	1dgw	y	IgE	578	587
uMS06	arah2.0101	9186485	1	1w2q	a	IgE	39	48
uMS07	arah2.0101	9186485	1	1w2q	a	IgE	127	132
uMS09	arah2.0101	9186485	2	3ob4	a	IgE	49	56
uMS10	arah2.0101	9186485	5	1pnb	b	IgE	143	150
uMS11	Gald1	P01005	1	2p6a	d	IgE	64	74
uMS12	Gald1	P01005	1	2p6a	d	IgE	95	99
uMS13	Gald1	P01005	1	2p6a	d	IgG	55	59
uMS14	Gald1	P01005	1	2p6a	d	IgG	70	74
uMS15	Gald1	P01005	1	2p6a	d	IgG	189	198
uMS19	Parj1	P43217	3	1t12	a	IgE	72	79
uMS20	Parj1	O04404	1	1mid	a	IgE	41	47
uMS22	Parj1	Q40905	1	1t12	a	IgE	41	47
uMS23	Parj1	Q40905	1	1t12	a	IgE	72	79
uMS24	Parj2	P55958	1	2alq	a	IgE	45	50
uMS25	Parj2	P55958	1	2alq	a	IgE	61	70
uMS26	Parj2	P55958	1	2alq	a	IgE	77	86
uMS27	Parj2	P55958	1	2alq	a	IgE	83	91
uMS28	Parj2	P55958	1	2alq	a	IgE	122	129
uMS29	Parj2	O04403	1	fk5	a	IgE	45	50
uMS30	Parj2	O04403	1	fk5	a	IgE	61	70
uMS31	Parj2	O04403	1	fk5	a	IgE	77	86
uMS32	Parj2	O04403	1	fk5	a	IgE	83	91
uMS33	Parj2	O04403	1	fk5	a	IgE	122	129
uMS34	Pena1	11893851	3	2b9c	b	IgE	85	99
uMS35	Pench18	7963902	5	1sh7	a	IgE	401	410
uMS37	arah2.0101	15418705	1	1w2q	a	IgE	127	132
uMS38	arah2.0101	15418705	2	3obq4	a	IgE	49	56
uMS39	arah2.0101	15418705	2	3obq4	a	IgE	117	122
uMS40	Aspf1	166486	1	1jbs	a	IgE	51	60
uMS42	aspf1	p04389	1	1jbs	a	IgE	51	60
uMS43	Aspf2	P79017	1	1eb6	a	IgE	58	64
uMS45	Aspf2	P79017	1	1eb6	a	IgE	181	185
uMS46	Aspf2	P79017	1	1eb6	a	IgE	189	192
uMS47	Aspf2	P79017	1	1eb6	a	IgE	200	204
uMS49	Aspf3	664852	1	1eb6	a	IgE	115	125
uMS52	Cryj2	P43212	5	2x3h	a	T-cell	400	414
uMS53	Fage1	2317670	1	3fz3	b	IgE	360	367
uMS54	Fage1	2317670	1	3fz3	b	IgE	411	416
uMS55	Fage1	2317674	1	2e9q	a	IgE	460	465
uMS56	Fage1	2983941	1	3qac	a	IgE	460	465
uMS57	Fage1	2983941	1	3qac	a	IgE	506	511

W-shaped "basket" motif (figure 2). The first part of the motif showed better quality in the multiple alignment, but with lower consensus than the second part. The second half of the motif showed greater conservation and was more consistently present in all the epitopes in the cluster 282, although the quality was overall lower.

Conclusion

The decision tree approach to evaluating the potential allergenicity of genetically engineered food products [14, 15] proposed in 1996 has since

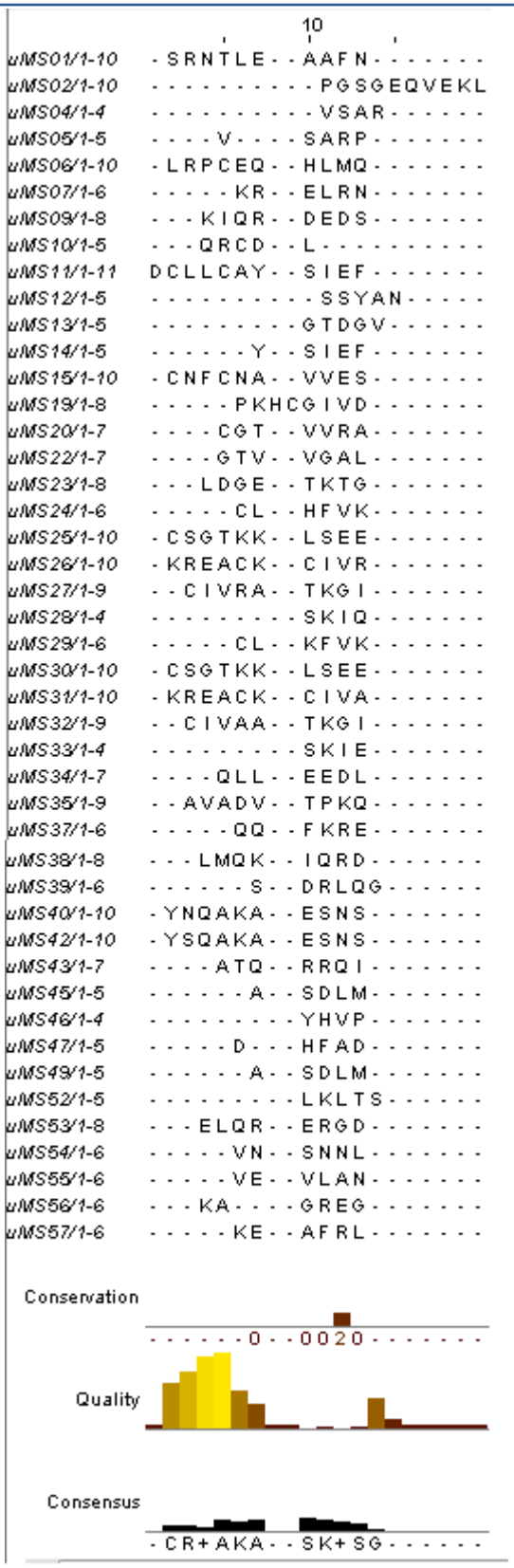


Figure 1. Structure based sequence alignment of members of Cluster 282 using MAPSCI shows a consensus sequence of CR+AKA..SK+SG.

been practiced widely to determine allergenic nature of novel protein products. This strategy uses primary sequence analysis and immunochemistry along with gene source to predict the allergenicity of the novel protein. We have presented here a structural classification of the epitopes from all allergens having epitopes in the Structural Database of Allergenic Proteins, SDAP. The largest single cluster in this study contained 231 epitopes representing 12 of the 16 species studied. Cluster analysis of 3D models generated for these allergens revealed that, while sequence identity between the allergen epitopes was generally very low, epitopes in this cluster shared common tertiary structure in the form of the W-shaped motif with consensus sequence CR+AKA--SK+SG. This study lends support to the recommendation that the decision tree approach to evaluating the potential allergenicity of genetically engineered food products, from the FAO/WHO, should include a stronger component of structural identity, in addition to the current focus on sequence identity. The putative shared tertiary structure motif in the epitopes in the largest cluster in the analysis could also contribute to the significant amount of cross-reactivity clinically observed in several allergens.

Acknowledgements

This study was funded in part by a Student Assistantship in the Physical Biosciences Division at Lawrence Berkeley National Laboratory (Berkeley Lab) in the lab of Dr. Peter H. Zwart. I would like to especially thank my advisor and mentor, Peter H. Zwart, for his guidance and encouragement. I would also like to thank the Berkeley Center of Structural Biology for allowing me the use of their computational facilities for this research. I would like to acknowledge the support of Sebastian Raschka, Ph.D. student at Michigan State University for a tool that converts pdb files to fasta files.

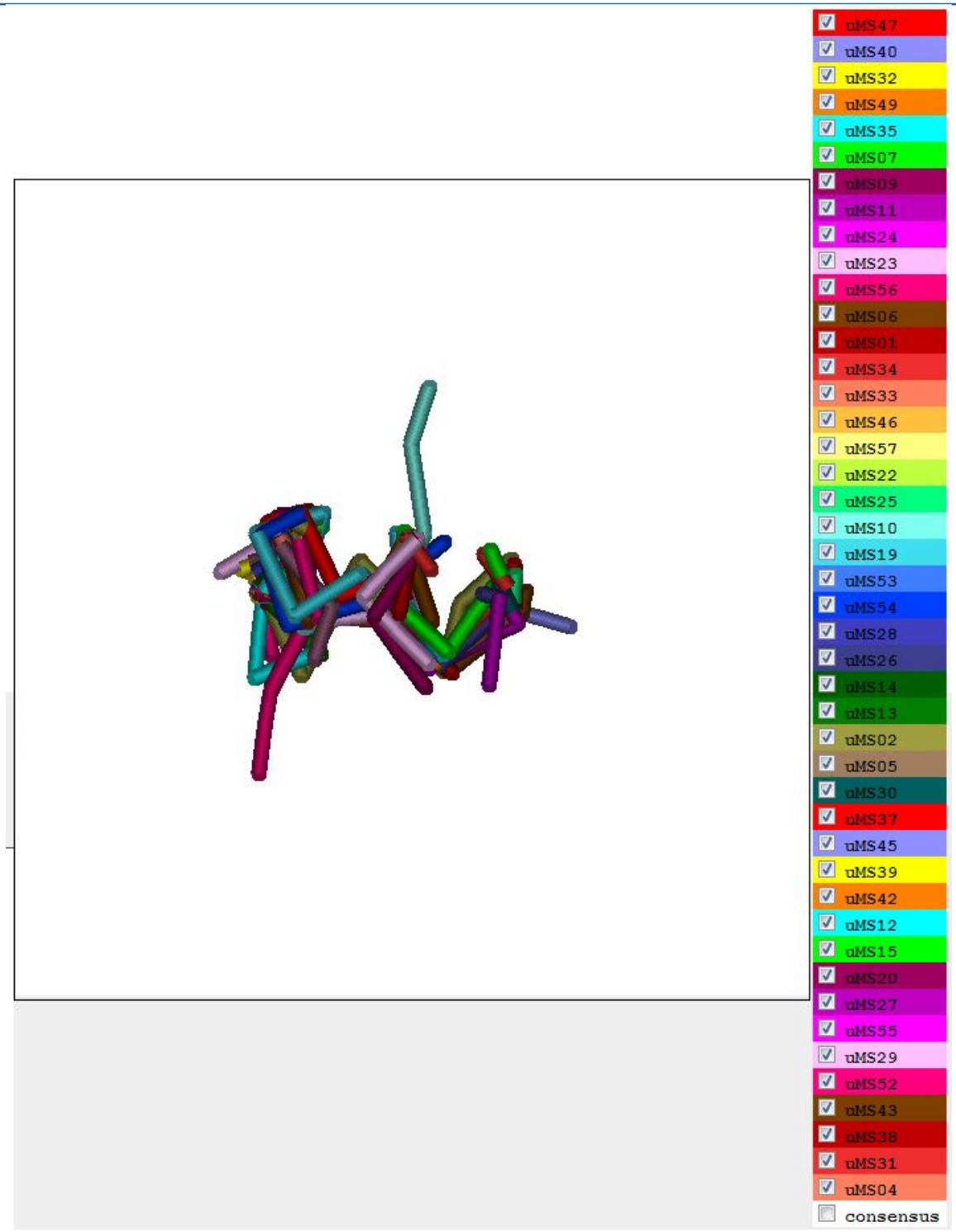


Figure 2. 3D visualization of the members of cluster 282 showing the W-shaped "basket" motif of the cluster. The members of this alignment are available in table 3.

References

1. Ovidiu Ivanciuc, Tzintzuni Garcia, Miguel Torres, Catherine H. Schein, Werner Braun (2009). *Molecular Immunology* 46, 559–568.
2. Christian Radauer, Merima Bublin, Stefan Wagner, Adriano Mari, Heimo Breiteneder (2008) *J Allergy Clin Immunol* 2008;121:847-52.
3. Ivanciuc, O., Schein. C. H., and Braun, W. (2003). *Nucleic Acids Res.* 31(1). 359-362.
4. Heimo Breiteneder, E.N. Clare Mills (2005). *Biotechnology Advances* 23, 395–399.
5. Christian Radauer, and Heimo Breiteneder (2007). *J Allergy Clin Immunol*. 120:518-25.
6. FAO/WHO, 2001, http://www.fao.org/es/ESN/food/pdf/allergygm.pdf; 2003, http://www.codexalimentarius.net/download/report/46/Al0334ae.pdf
7. Ovidiu Ivanciuc, Catherine H. Schein, Tzintzuni Garcia, Numan Oezguen, Surendra S. Negi, Werner Braun. (2009). *Regulatory Toxicology and Pharmacology* 54, S11–S19.
8. Zemla A. (2003). *Nucleic Acids Research*, Vol. 31, No. 13, pp. 3370-3374.
9. Jmol: an open-source Java viewer for chemical structures in 3D. http://www.jmol.org/
10. Kozin, M. B., and Svergun, D. I. (2001) *J. Appl. Crystallogr.* 34, 33.
11. http://cctbx.sourceforge.net/
12. http://sebastianraschka.com/webapps.html
13. Fabian Sievers, Andreas Wilm, David Dineen, Toby J Gibson, Kevin Karplus, Weizhong Li, Rodrigo Lopez, Hamish McWilliam, Michael Remmert, Johannes Söding, Julie D Thompson, and Desmond G Higgins (2011). *Mol Syst Biol.* 7: 539.
14. Dean D. Metcalfe, James D. Astwood, Rod Townsend, Hugh A. Sampson, Steve L. Taylor & Roy L. Fuchs (1996). *Critical Reviews in Food Science and Nutrition* 36, Suppl. 001.
15. http://www.fao.org/docrep/007/y0820e/y0820e05.htm
16. Hong SJ, Michael JG, Fehringer A, Leung DY. (1999). *J Allergy Clin Immunol*. Aug;104(2 Pt 1):473-8.
17. Ivaylo Ilinkin, Jieping Ye, Ravi Janardan (2010). *BMC Bioinformatics*, 11:71.