

Abundance and functional roles of intrinsic disorder in allergenic proteins and allergen representative peptides

Bin Xue,^{1,2,3} Daniel Soeria-Atmadja,⁴ Mats G. Gustafsson,⁵ Ulf Hammerling,⁴
A. Keith Dunker,^{2,3} and Vladimir N. Uversky^{1,2,3,6*}

¹ Department of Molecular Medicine, University of South Florida, Tampa, Florida 33612

² Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, Indiana 46202

³ Institute for Intrinsically Disordered Protein Research, Indiana University School of Medicine, Indianapolis, Indiana 46202

⁴ Department of Toxicology, National Food Administration, Uppsala SE-751 26, Sweden

⁵ Cancer Pharmacology and Computational Medicine, Department of Medical Sciences, Uppsala University, Uppsala Academic Hospital, Uppsala SE-751 85, Sweden

⁶ Institute for Biological Instrumentation, Russian Academy of Sciences, Pushchino 142290, Moscow Region, Russia

ABSTRACT

The pathological process of allergies generally involves an initial activation of certain immune cells, tied to an ensuing inflammatory reaction on renewed contact with the allergen. In IgE-mediated hypersensitivity, this typically occurs in response to otherwise harmless food- or air-borne proteins. As some members of certain protein families carry special properties that make them allergenic, exploring protein allergens at the molecular level is instrumental to an improved understanding of the disease mechanisms, including the identification of relevant antigen features. For this purpose, we inspected a previously identified set of allergen representative peptides (ARPs) to scrutinize protein intrinsic disorder. The resulting study presented here focused on the association between these ARPs and protein intrinsic disorder. In addition, the connection between the disorder-enriched ARPs and UniProt functional keywords was considered. Our analysis revealed that ~20% of the allergen peptides are highly disordered, and that ~77% of ARPs are either located within disordered regions of corresponding allergenic proteins or show more disorder/flexibility than their neighbor regions. Furthermore, among the subset of allergenic proteins, ~70% of the predicted molecular recognition features (MoRFs that consist of short interactive disordered regions undergoing disorder-to-order transitions at interaction with binding partners) were identified as ARPs. These results suggest that intrinsic disorder and MoRFs may play functional roles in IgE-mediated allergy.

Proteins 2011; 00:000–000.
© 2011 Wiley-Liss, Inc.

Key words: allergen; intrinsic disorder; peptide; MoRF; UniProt keyword.

INTRODUCTION

Allergies may affect up to ~20% of the total population in many countries and some hypersensitivity ailments seem to be increasing, particularly in industrialized regions.^{1,2} The symptoms are caused by an inappropriate immune response to otherwise harmless foreign antigens. This can occur in atopic individuals, who have a genetic predisposition to develop allergic immune reactions. In IgE-mediated allergy (also known as Type I or immediate-type allergy), an initial immunological exposure to certain proteins involves a preferential direction of the antibody production toward synthesis of IgE antibodies, a process referred to as sensitization. On renewed contact with the antigen, IgE molecules—attached to high-affinity receptors on tissue mast cells or basophilic granulocytes—can capture the sensitizing antigen and thereby elicit release of various inflammatory substances (mediators). This (second) phase is known as triggering.^{3,4} These mediators will act in concert and cause any or several among a variety of symptoms, such as asthma, rhinitis, rhinoconjunctivitis, eczema, contact dermatitis, angioedema, and abdominal pain. Allergens can also trigger life-threatening anaphylactic reactions in sensitized individuals.⁵ Importantly, a sensitized individual may also respond similarly to other substances, sharing certain IgE-epitopes with the molecule that prompted the initial immune reaction. This phenomenon is designated as IgE cross-reactivity and typically—but not consistently—involves homologous allergenic

Daniel Soeria-Atmadja's present address is R&D Information, AstraZeneca, Snäckviken B-203, Stockholm SE-171 77, Sweden.

Grant sponsor: National Science Foundation; Grant number: EF 0849803; Grant sponsor: Program of the Russian Academy of Sciences for the "Molecular and Cellular Biology."

*Correspondence to: Vladimir N. Uversky, Department of Molecular Medicine, University of South Florida, 12901 Bruce B. Downs Blvd., MDC 07, Tampa, FL 33612. E-mail: vuversky@health.usf.edu.

Received 14 February 2011; Revised 14 April 2011; Accepted 4 May 2011

Published online 10 May 2011 in Wiley Online Library (wileyonlinelibrary.com).

DOI: 10.1002/prot.23077

proteins from phylogenetically related species.^{6,7} An understanding of many cross-reactive patterns has developed in recent years; for example, it is common that subjects allergic to major pollen allergens also react to a variety of fruits and/or vegetables due to the occurrence of structurally similar components.⁸ Recently, we have reported a global delineation of allergen relationships, as demonstrated by IgE-reactivity, in sensitized individuals.⁹

Typically, an underlying principle in allergy follows the amino acid sequence–structure–function paradigm. For example cross-reactivity is not frequently observed below 45–50% sequence identity.^{10,11} Thus, a variety of bioinformatics implements for the identification of potentially cross-reactive allergens have been developed, mostly founded on the amino acid sequence alignment principle, but comparisons based on other protein features are also reported.^{12,13} Notably, virtual peptides with certain shared properties have been derived from collections of allergen proteins by a range of techniques.^{14–18} One of those is based on a special dataset, referred to as the allergen representative peptides (ARPs) identified by cross-sequence alignment between groups of known allergenic and nonallergenic proteins.¹⁶ This set was used for a study aiming at shedding light on whether allergens and peptides derived thereof may harbor amino acid sequence motifs, which would indicate special structural features.

As supported by gradually accumulating data, the classical protein structure paradigm (where a specific protein function is determined by a unique protein structure which is determined by an amino acid sequence) is no longer the only model describing the relation between sequence and function, due to the discovery of intrinsically disordered proteins (IDPs).^{19–32} IDPs are biologically active proteins that do not have rigid three-dimensional (3D) structures under physiological conditions. Instead, they form highly dynamic ensembles of flexible conformations. Many biological functions are enabled by IDPs through transient conformational changes.^{19,33–35} In experiments, IDPs and intrinsically disordered regions (IDRs) can be identified by various techniques, such as X-ray crystallography (missing electron density), NMR (highly flexible regions), protease digestion, circular dichroism, or several other experimental settings.^{36–38}

IDPs are abundant in nature. As estimated by genome-level computational studies, 7–30% prokaryotic proteins contain long disordered regions of more than 30 consecutive residues, whereas in eukaryotes the fraction of such proteins is even higher, about 45–50%.^{21,39–43} Even in the PDB database, which is highly biased toward structured proteins, ~ 70% of proteins have IDRs manifested as regions without coordinates in the corresponding PDB structures.^{44,45} Among them, more than 10% have long segments, encompassing of missing electron density consisting of at least 30 amino acids. IDPs are not only pop-

ulation-dominant, but also functionally important. In all 710 SwissProt functional keywords, 238 were shown to be strongly positively associated with intrinsic disorder, whereas 302 other functional key words were strongly negatively correlated with IDP.⁴⁶ In a report on the human genome, disordered regions were found to be independent and evolutionarily conserved functional units.⁴⁷

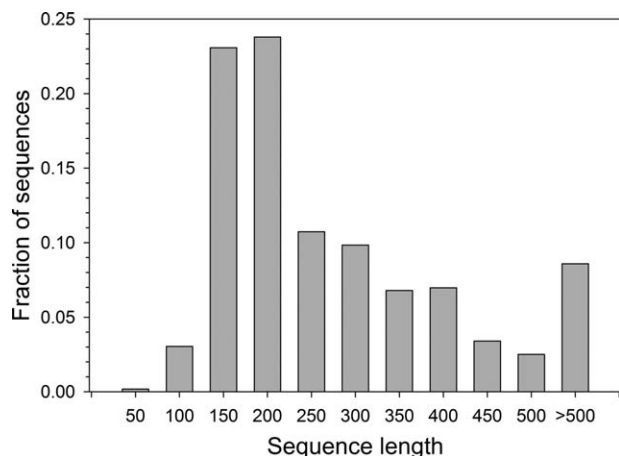
Previous findings reveal a tight correlation between the protein intrinsic disorder and protein binding. A special type of intrinsic disorder-based interaction-prone sequence motifs—known as molecular recognition features (MoRFs)—was shown to play an important role in protein–protein interactions (PPIs).^{42,48} A MoRF is a short structure-prone segment within a long disordered region which usually undergoes a disorder-to-order transition on binding to a partner. Such segments are frequently found in proteins involved in recognition, signaling, and regulation. It was estimated that more than 40% of eukaryotic proteins contain at least one MoRF that forms a helix upon binding to its partner.^{42,49}

The immunological processes of IgE-mediated allergy, eventually precipitating as pathophysiological reactions in response to antigen exposure, depend on a range of specific PPIs. IDRs are known to play crucial roles in PPI networks^{50,51}; thus, we hypothesized that protein intrinsic disorder may be implicated this disease. Especially, IDRs may contribute to hypersensitivity reactions via MoRFs. Because allergens, connected with Type I hypersensitivity, are proteins and because protein intrinsic disorder is both abundant and functionally important in biological systems, we have embarked on analyzing a possible connection between IDRs and allergenicity.

METHODS

ARP dataset

ARPs were generated from pair-wise cross-sequence alignment among sequences from datasets of allergenic and nonallergen proteins, an operation which results in a set of peptides with a very marked predominance in the allergens. The set of protein allergens (the allergen dataset) was created by merging six other publically available datasets.¹⁶ The set of presumably nonallergen (henceforth referred to as nonallergen for simplicity) proteins (the nonallergen dataset) included the nonallergen sequences from the Swall database of commonly consumed commodities, such as rice, apple, carrot, peach, cherry, apricot, spinach, tomato, salmon, cow's milk, and hen's egg.¹⁶ All sequences in these datasets were computationally cut into peptides by sliding windows of the same length. Subsequently, all peptides from the allergen dataset were aligned against that of the nonallergen counterpart. Finally, allergen peptides that differed from those of the nonallergen set were jointly labeled the ARP dataset.¹⁶ In this study, 559 allergenic proteins and 2896

**Figure 1**

Characterization of allergenic proteins and ARPs. Plot represents the analysis of the length distribution of all the ARP-containing proteins in the dataset of 559 allergenic proteins.

ARPs were analyzed. Figure 1 represents the length distribution of all the 559 allergenic proteins and shows that these ARP-containing proteins constitute a very diverse dataset, with length ranging from 50 to >500 residues.

Disorder prediction

Peptides of the ARP dataset encompass only 24 amino acids and thus lie below the (reliable) operative range of most predictors of protein disorder. Thus, ARPs were inspected with regard to this feature by a specially designed disorder predictor for short peptides: short peptide analyzer (SPA).⁵² It embeds the query sequence into a preselected ensemble of sequences, thereby creating a series of elongated combined peptides. These combined peptides were then analyzed by the PONDR[®] VL-XT predictor due to its extreme sensitivity to local composition of amino acids.²¹ The final disordered status of a query peptide is the ensemble-averaged disorder score calculated for the corresponding amino acids in all combined peptides. SPA was shown to remarkably improve the prediction accuracy for short disordered regions,⁵² which is one of the bottlenecks in protein disorder prediction.^{29,43,52} Furthermore, the use of SPA in this study is further justified by the fact that disorder prediction can be context-dependent; that is, the results of the disorder prediction for a given region can be affected by the peculiarities of disorder distribution in the neighboring regions. As a result, the predicted disorder status of terminal residues of a short peptide is heavily influenced by the order/disorder status of the flanking regions. The SPA predictor minimizes these boundary effects by inserting the query sequence in every protein segment of the large set of preselected sequences and by providing

the ensemble-averaged data. The disorder propensities are predicted for all these combined peptides. The disorder predictions corresponding to the short query peptide from all the combined peptides are averaged and are taken as the final disorder scores.⁵² Analysis of the per-residue disorder distribution in 559 allergenic proteins was performed by PONDR[®] VL-XT,⁴⁰ and by a meta-predictor PONDR-FIT.⁴³

MoRF prediction

Being defined as short order-prone motifs that are located within long disordered regions and that undergo a disorder-to-order transition during the binding process, MoRFs are usually more enriched in aliphatic and aromatic amino acids relative to typical IDRs. Because of these sequence peculiarities, MoRFs are frequently observed as sharp dips in the corresponding PONDR[®] VL-XT disorder score plots. Hence, MoRF regions can be reliably identified on the basis of special characteristics of the PONDR[®] VL-XT prediction profiles and on several other specific attributes.^{42,48} Furthermore, the length of MoRFs identified by MoRF-II predictor (18 amino acids) is within the range of ARP database members.

Charge-hydrophathy-cumulative distribution function plot

A charge-hydrophathy (CH)-cumulative distribution function (CDF) plot is the readout of a disorder analysis method that combines outputs of CH-plot^{20,41} and CDF-plot into a single graph.^{41,53} A CH-plot shows results from a binary disorder predictor and represents an input protein as a 2D graph, in which the mean Kate-Doolittle hydrophobicity and the mean absolute net charge are projected onto the X- and Y-coordinates, respectively. In the corresponding CH-plot, fully structured proteins and fully disordered proteins can be separated by a boundary line. The proteins above this boundary line are likely to be coil-like and often extended, whereas the proteins below this line are likely to be structured or at least compact. The vertical distance on the CH-plot from the location of the protein to the boundary line is then used as a measure of disorder (or structure) tendency of a given protein. This distance is further referred to as CH-distance.

CDF-plot is another binary disorder predictor, which describes the disorder status of an entire protein.^{41,53,54} In brief, it is a cumulative percentage of disordered residues at various disordered scores. By definition, ordered proteins will have many structure-promoting residues and many fewer disorder-promoting residues. Therefore, the CDF curve of an ordered protein increases very sharply at low disordered scores, and then goes flat for high disordered scores. For disordered proteins, the CDF curve will describe a shallow upward slope at regions of

low disordered scores and change to a steep ascend at high disordered scores. Hence, on the CDF plot, structured proteins intend to stay on the upper left half, whereas disordered proteins locate at the lower right half of the plot. By comparing the locations of CDFs for a group of fully disordered and fully structured proteins, a boundary line between these two groups of proteins can be identified. Proteins with CDF above the boundary line are structured, whereas those below this demarcation are disordered. For a particular protein, the distance from CDF boundary is also a kind of measure of disordered (structured) status and is referred to as CDF-distance. The accuracy of CDF functions varies across different predictors. In our previous study, VSL2-CDF achieved the highest accuracy which was 5–10% higher than the second best of other five CDF functions.⁵⁴ However, as the disorder status of peptides was evaluated by SPA (this study), being a new PONDR[®] VL-XT-based tool for short peptides, the PONDR[®] VL-XT-based CDF boundary was applied to evaluate the CDF-distance for SPA prediction.

By combining CH-distance and CDF-distance, a new measure named CH-CDF plot was created.⁵⁵ This plot provided very useful information on whether an entire protein sequence is structure-prone or disorder-prone.

RESULTS AND DISCUSSION

Intrinsic disorder in ARPs and ARP-containing proteins

Protein intrinsic disorder plays various important functional roles, being frequently found in proteins/regions involved in various PPIs, including those crucial for signal transduction. Furthermore, the complex cascade of events leading to allergic reactions is also based on a set of specific PPIs, hitherto only fragmentarily unveiled. A possible connection between protein intrinsic disorder and allergenicity is, however, still unexplored. To provide some insight in this assumed connection and to see whether IDRs might be implicated in allergy, our initial approach involved studies on occurrence of predicted disorder in ARPs and ARP-containing proteins. As shown in Figure 2(a), only ~ 20% of ARP-containing proteins were predicted to have low disorder content (i.e., 10% or less of intrinsically disordered residues), whereas ~ 50% of ARP-containing proteins had more than 20% of intrinsically disordered residues in their sequences. On the other hand, the intrinsic disorder in allergenic proteins was mostly confined to short disordered regions, as ~ 80% allergenic proteins were predicted by PONDR[®] VL-XT to have such regions no longer than 20 residues [see Fig. 2(b)]. These values are rather close to those calculated for structured proteins from PDB, but are much lower than those predicted for signaling proteins and proteins related to various human diseases such as cancer, neurodegeneration, cardiovascular disorders, diabetes, and amyloidoses.^{56–60}

Next, the fraction of disordered residues in ARPs was evaluated. Figure 2(a) clearly shows that although about 55% of ARPs have 5% or less of disordered residues (all ARPs have 24 residues, therefore, 5% is just one residue), ~ 20% of ARPs have more than 50% of disordered residues. These data indicate that protein allergens and ARPs are mostly ordered.

This conclusion was further supported by results of the CH-CDF analysis,⁵⁵ which is a useful tool for identifying disorder status of entire protein/peptide [see Fig. 2(c)]. CH-CDF plot can be split into four quadrants: Q1 (upper right), Q2 (lower right), Q3 (lower left), and Q4 (upper left). Proteins/peptides in Q1 are structured according to CDF, but disordered according to CH; proteins/peptides in Q2 are predicted to be structured by both CDF and CH; proteins/peptides in Q3 are disordered according to CDF but structured according to CH; and Q4 represents proteins predicted to be disordered by both methods.⁵⁵ Figure 2(c) reveals that the disorder status of allergenic proteins in the CH-CDF plot roughly matches results of the per-residue disorder predictions and clearly shows that a large fraction of the protein allergens is predicted to assume an ordered structure. Relative to the respective population sizes of ARP-containing proteins, ARPs are similar in Q4, much smaller in Q3, and much larger in Q1 and Q2. This is in line with the above observations that ARPs are frequently structure-prone. By looking at the CH-CDF plot, it is clear that ARPs have more diversified distribution than ARP-containing proteins toward the edges of the plot. This suggests that ARPs have either more charged residues or more hydrophobic residues.

The relative abundance of canonical amino acids in proteins of various datasets can be identified from amino acid composition profiles.^{21,61} As shown in Figure 2(d), compared with ARP-containing proteins, ARPs have more Cys, Glu, and Lys, but less Val, Leu, Met, and Arg. ARP-containing proteins and ARPs also have unique preferences for certain structural classes of amino acids, as compared with the fully structured dataset (FSD) and fully disordered dataset (FDD).^{41,43,54} Clearly, compared with FSD, ARPs are markedly enriched in C and A (>20%) and moderately enriched in G, E, and K (10–20%), being significantly depleted in W and R (-20%) and moderately depleted in Y, L, M, and N (-10 to -20%). Compared with the FDD, sequences in ARP are significantly enriched in N and aromatic and aliphatic residues, moderately enriched in A and G, significantly depleted in R, Q, S, P, and E, and moderately depleted in M.

ARPs are frequently located in flexible regions

After revealing the relative abundance of intrinsic disorder in ARPs and ARP-containing proteins, an emerging question is as follows: are those ARPs located in flexible or structured regions of ARP-containing proteins? This

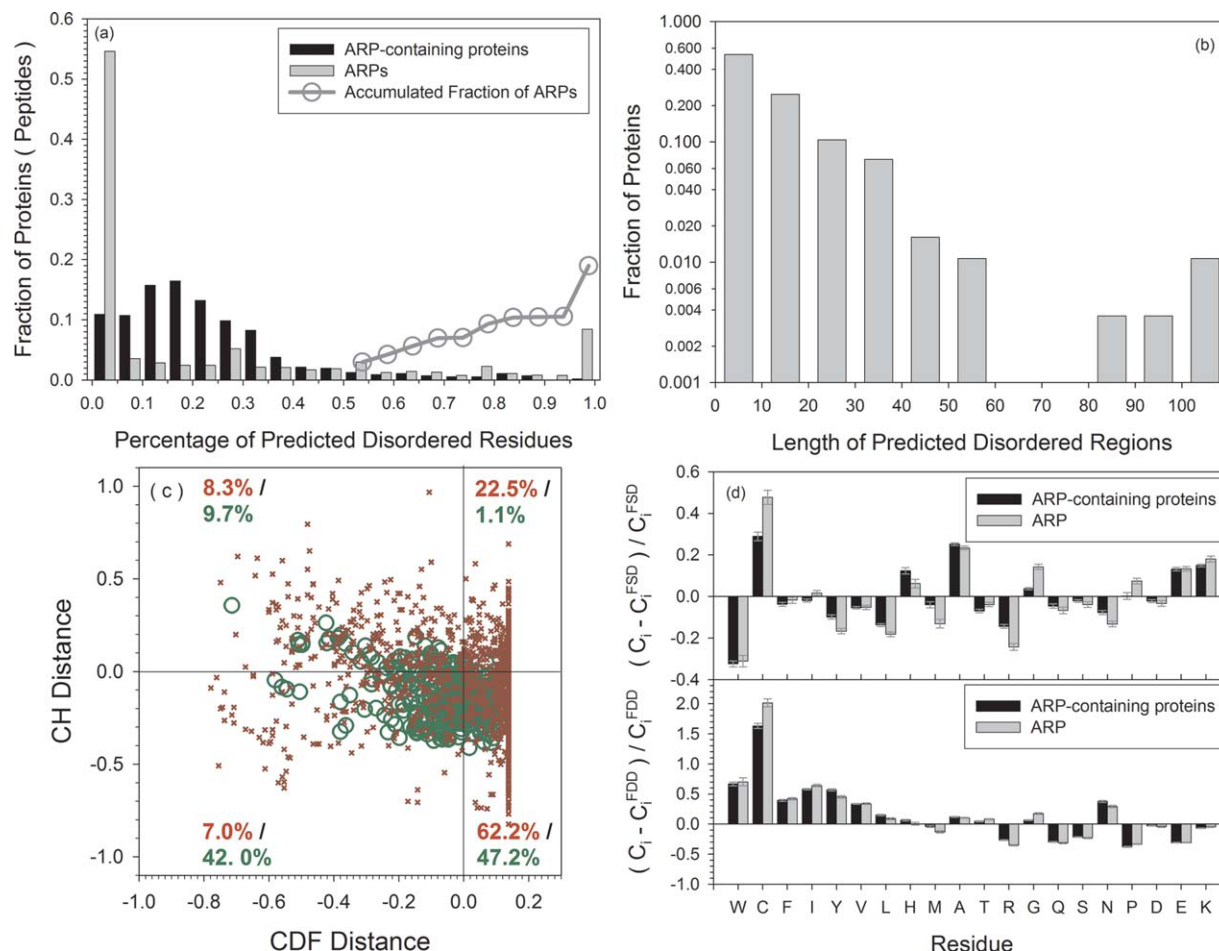


Figure 2

Analyzing intrinsic disorder in ARPs and ARP-containing proteins. **a:** Analysis of the per-residue disorder probabilities of ARPs (gray bars) and ARP-containing proteins (black bars) was performed by PONDR[®] VL-XT.²¹ The circles connected by gray curve are the accumulated fraction of ARPs having 50% or more disordered residues. **b:** Distribution of IDRs of various lengths in ARP-containing proteins. **c:** CH-CDF analysis⁵⁵ for ARPs (brown crosses) and ARP-containing proteins (green circles). The CDF distances are calculated from the PONDR[®] VL-XT-based CDF curves.^{20,41} CH distances are calculated from CH-plot.^{41,53} The CH-CDF plot can be divided into four quadrants: Q1 (upper right), Q2 (lower right), Q3 (lower left), and Q4 (upper left). Proteins in Q1 are structured according to CDF, but disordered according to CH; proteins in Q2 are predicted to be structured according to both CDF and CH; proteins in Q3 are disordered according to CDF but structured according to CH; and Q4 is for disordered proteins predicted according to both methods. Therefore, quadrants Q3 and Q4 include proteins of a general disordered character, with Q3 containing native molten globules, and with Q4 containing native premolten globules and native coils.⁵⁵ The numbers are corresponding percentages of ARPs (brown) and ARP-containing proteins (green) in each quadrant. **d:** Compositional profiling^{21,61} of ARPs and ARP-containing proteins. “C” indicates the content of specific amino acid in the dataset, whereas “C^{ref}” is the content of the same amino acid in the reference dataset. Reference datasets are FSD and FDD.^{41,43,54} The error bars are calculated by bootstrapping. The amino acids on x-axis are arranged in the order of increased flexibility.

question was inspired on the basis of Figure 2(b) showing that $\sim 30\%$ of ARP-containing proteins had long disordered regions of more than 20 consecutive residues. As the length of ARPs is 24 amino acids and 20% of ARPs are highly disordered, as shown in Figure 2(a), the position of ARPs within ARP-containing proteins may disclose features of protein allergens with potential relevance to the pathogenic mechanisms.

To systematically investigate the local environment of ARPs within allergenic proteins (i.e., to see whether ARPs are located inside the disordered or ordered regions), two computational approaches were used. There

are four main potential scenarios for ARPs and their local environment that could be grouped in four classes, where in Class I, a disordered ARP is located inside a disordered region; in Class II, an ordered ARP is located inside a disordered region; in Class III, a disordered ARP is located inside an ordered region; and in Class IV, an ordered ARP is located inside an ordered region. Obviously, these four scenarios are all clearly segregated in the PONDR[®] VL-XT plots, with peptides from Classes II and III appearing as characteristic “dips” and “spikes,” respectively, in the corresponding disorder curves [Fig. 3(a)]. Computationally, differences across local sequence

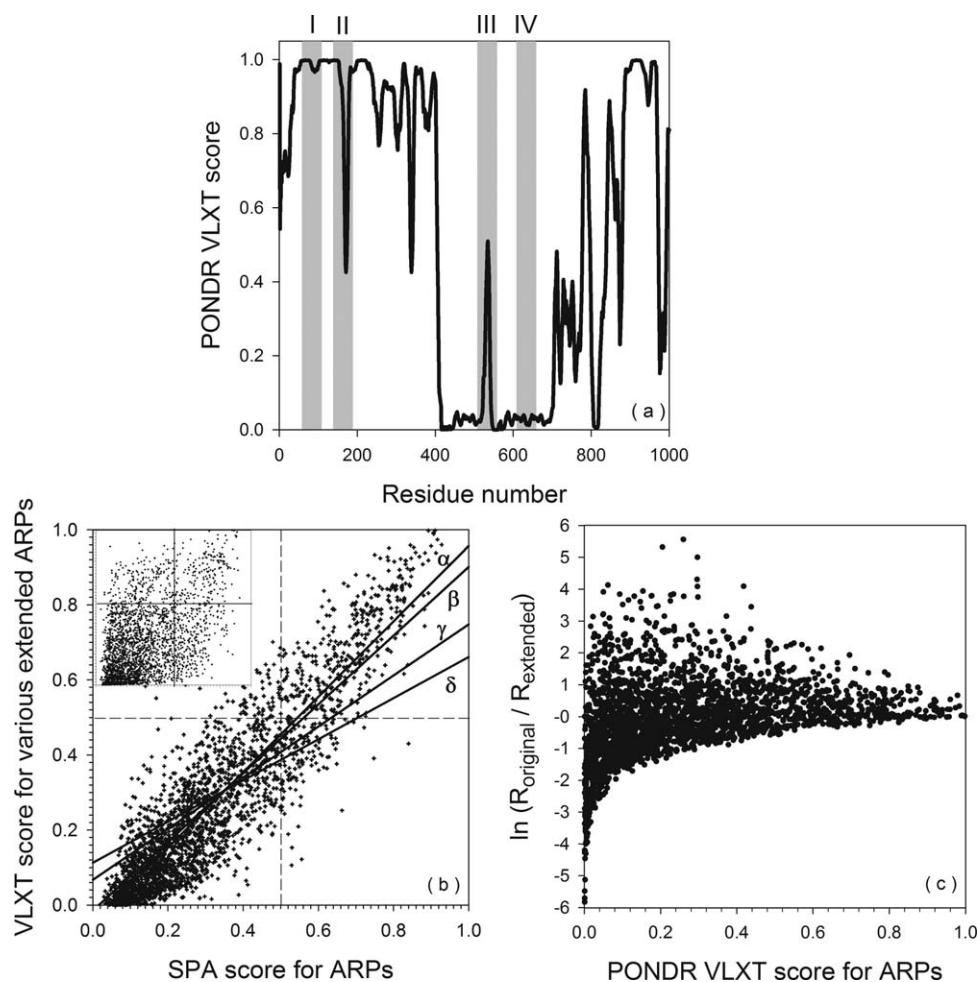


Figure 3

Analysis of the local environment of ARPs within the amino acid sequences of allergenic proteins. **a:** Introducing classification of ARPs based on their local environment within the sequences of parent allergenic proteins: Class I, a disordered ARP is located inside a disordered region; Class II, an ordered ARP is located inside a disordered region; Class III, a disordered ARP is located inside an ordered region; and Class IV, an ordered ARP is located inside an ordered region. PONDRL[®] VL-XT curve of a model protein is shown and positions of mentioned classes are shown as gray shaded areas. **b:** Distribution of IDRs in ARP-containing proteins. The prediction was carried out by SPA⁵² and PONDRL[®] VL-XT.⁴⁰ The coordinates for each spot are disorder scores averaged over the peptide predicted by SPA (*x*-axis) and PONDRL[®] VL-XT (*y*-axis). Influence of neighboring residues was measured by the averaging disorder score over peptides extended on both sides by a specific number of residues. However, the SPA scores on *x*-axes are always from original ARPs. In this plot, data for the original ARP peptides are shown as spots. Line α is the linear regression of these spots. Lines β , γ , and δ represent regressions of peptides extended at both termini by 6, 9, and 15 residues. The inset shows the change of disordered scores after extending the ARPs by 15 residues. **c:** $[\ln(R_{\text{original}}/R_{\text{extended}})]$ versus PONDRL[®] VL-XT score averaged over the original ARP plot. Here, R_{original} is the PONDRL[®] VL-XT score averaged over the original ARP, whereas R_{extended} is the PONDRL[®] VL-XT score averaged over the peptide including the original ARP extended by 15 residues at both termini. For each ARP, R_{original} and R_{extended} are calculated from the corresponding PONDRL[®] VL-XT data for the parent allergenic protein.

environments of ARPs can be evaluated via analysis of extensions of ARPs at both termini, within their respective proteins, on the averaged disorder scores. In greater detail, the original-averaged disorder scores of ARPs were set on the *x*-axis, whereas the averaged disorder scores for various ARP extensions were projected on the *y*-axis. Thus, ascending scores on the *y*-axis indicate a preferential ARP location within long disordered regions, and vice versa. For Classes I- and IV-derived peptides, such extensions are not supposed to reveal noticeable changes in the averaged disorder scores. Conversely, ARPs of the

Class II sort would clearly show decreased disorder scores by such an operation, whereas an increase in averaged disorder score would indicate ARP assignment to Class III.

Results of the above-mentioned analysis are presented in Figure 3(b). The coordinates for each spot in Figure 3(b) are averaged disorder scores over the peptide, as predicted by SPA (*x*-axis) and PONDRL[®] VL-XT (*y*-axis). Influence of neighboring residues was measured by the averaged disorder score over peptides including their respective (fixed number residue) bi-directional extensions.

Scores for ARPs from SPA were on x -axis. Scores from PONDR[®] VL-XT for original ARPs and extended ARPs by 6, 9, and 15 residues on both termini are presented on the y -axis. Figure 3(b) represents the general outcome of this analysis by showing results of the original (nonextended) ARP analysis as dots, and a series of lines describing the linear least square regressions of data for nonextended (line α) and variously extended ARPs: lines β , γ , and δ represent regressions of peptides extended at both termini by 6, 9, and 15 residues, respectively. The linear least square regression functions are as follows:

$$\text{VLXT} = -0.048 + 1.005 \times \text{SPA (line } \alpha),$$

$$\text{VLXT} = -0.015 + 0.916 \times \text{SPA (line } \beta),$$

$$\text{VLXT} = 0.067 + 0.681 \times \text{SPA (line } \gamma), \text{ and}$$

$$\text{VLXT} = 0.113 + 0.548 \times \text{SPA (line } \delta).$$

The decreased slope from line α to line δ indicates that ARPs are more flexible than their respective extended peptides. Thus, when taking the local environment into consideration, an appreciable fraction of ARPs falls within the Class III; that is, disordered ARPs located in reasonably ordered protein regions.

It is also evident, however, that another notable part of the ARP set belongs to Class II (less disordered peptides residing inside a more disordered region). ARPs, extended at both termini by 15 residues, were plotted as PONDR[®] VL-XT scores averaged over peptides extended by 15 residues versus SPA scores averaged over the original ARPs as in the inset of Figure 3(b). Both Figure 3(b) and the inset can be each divided into four quadrants: Q1 (up right); Q2 (low right); Q3 (low left); and Q4 (up left). Apparently, the disorder predictions in Q1 and Q3 are consistent between SPA and PONDR[®] VL-XT. The fractions of original (nonextended) ARPs in the respective quadrants are 14.1%, 5.4%, 77.9%, and 2.6%, whereas those of extended ARPs by 15 residues are 8.6%, 10.8%, 71.5%, and 9.1%. Jointly, these two plots clearly show that ARP extension results in a dramatic redistribution of scores. Actually, an analogous plot of original ARPs shows that most peptides form a cloud around the diagonal, indicating a reasonably high correlation between SPA and VLXT scores. An extension of 15 residues at each ARP terminus, however, translates to data spreading over almost the entire plot. Notably, peptide length extensions lead to higher number of points in quadrants Q2 and Q4 and to a correspondingly lower number of spots in quadrants Q1 and Q3. The observed shift of spots from Q1 to Q2 suggests that the corresponding disordered peptides are located inside the more ordered regions. Analogously, shifts from Q3 to Q4 indicate a more ordered peptide structure, relative to their surroundings. Concisely, results derived from these exercises, involving an in-depth amino acid sequence context-based analytical approach, clearly indicate that many ARPs belong to Classes II and III [Fig. 3(a)].

We also attempted to determine the proportions of ARPs in each of these four IDR classes. To this end, for each ARP in the dataset, its averaged original PONDR[®]

VL-XT score was divided by the averaged disorder score calculated from the PONDR[®] VL-XT data for the corresponding parental allergenic protein over the region that included an ARP and its 15 residues-long N- and C-terminal extensions. In this analysis, peptides from Classes I and IV would be characterized by the ratios close to unity, peptides from Class II would show ratio significantly above 1, whereas Class III peptides would show ratios clearly below 1. Figure 3(c) represents results of this analysis by plotting $[\ln(R_{\text{original}}/R_{\text{extended}})]$ versus PONDR[®] VL-XT score averaged over the original ARP (R_{original}). From this plot, populations of different classes were estimated to 9.9%, 29.7%, 37.2%, and 23.2% for Classes I, II, III, and IV, respectively. This means that (i) a predominant fraction of ARPs ($\sim 67\%$) fall in either of the “dips” or of the “spikes” of the PONDR VL-XT disorder plots; (ii) $\sim 77\%$ of ARPs are located either within disordered regions of corresponding allergenic proteins or show more disorder/flexibility than their neighbor regions within the parent sequence.

Binding motifs in the ARP dataset

To further elucidate the role of intrinsic disorder in PPIs of 559 allergenic proteins used to design the ARP dataset, their potential binding sites were evaluated by the MoRF-II predictor.^{42,48} 228 MoRFs were found in 104 allergenic proteins, of which 169 MoRFs were found to overlap with ARP peptides. Therefore, although only 20% (104/559) of allergenic proteins carry predicted MoRFs, more than 70% (169/228) of these predicted MoRFs were identified as allergenic peptides. By taking into consideration the fact that MoRFs are functional binding motifs, the high abundance of MoRFs in allergenic peptides suggests that protein intrinsic disorder may be functionally implicated in certain aspects of allergy. As it was already pointed out, the length of MoRFs identified by MoRF-II predictor (18 amino acids) is within the range of ARP database members (24 residues). The overlap of ARP peptides with MoRFs is further illustrated in Figure 4. Figure 4(a) represents the fractions of allergenic proteins in which an ARP is overlapped with a MoRF by various numbers of residues and shows that the overlapped regions range in length from 1 to 18 residues. Figure 4(a) also shows that in the majority of cases [64% (108/169)], ARPs contained at least a half of the corresponding MoRFs, and in the 21% (36/169) cases, MoRFs were completely included in ARPs (corresponding overlapping regions were of 18 residues). In Figure 4(b), the 0–1 range, R , of disorder content was split into four smaller ranges, r : $0 \leq r \leq 0.1$; $0.1 < r \leq 0.5$; $0.5 < r \leq 0.9$; and >0.9 . This division reflects the basic statistics found in Figures 1–3 and kept the number of MoRFs in each range roughly the same. Clearly, because of the decrease in the number of ARP peptides

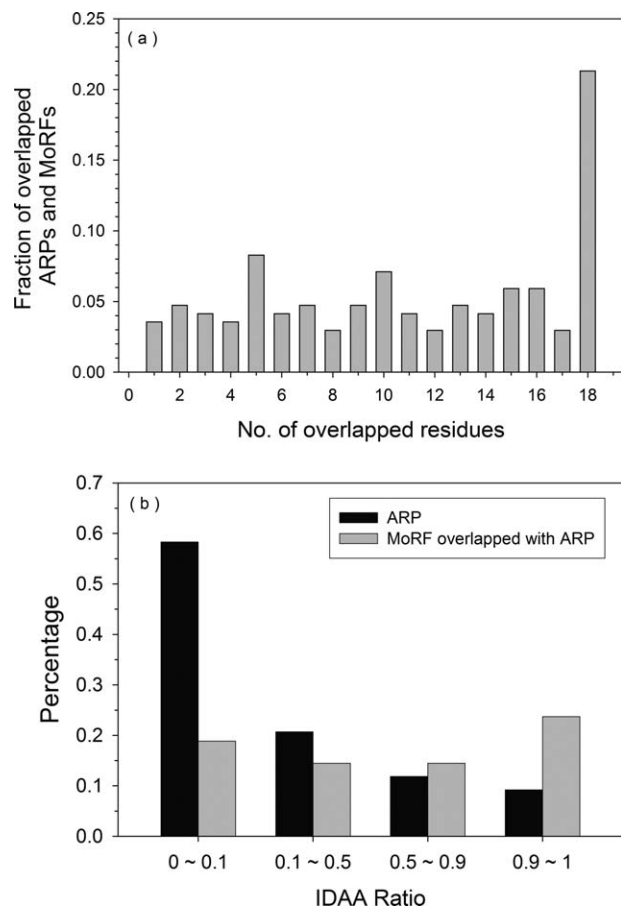


Figure 4

Analysis of the overlap between the ARPs and α -MoRFs. **a:** Plot represents the fraction of proteins in which ARP is overlapped with MoRF by various numbers of residues. **b:** Percentage of ARPs in various ranges of disorder content as evaluated by ratio of predicted disordered residues, and the ratio of ARPs which are overlapped with the predicted MoRFs^{42,48} over the total number of predicted MoRFs. The predicted MoRF is considered to overlap with ARP if they have sequence overlap of at least one residue. Black bars are the percentage of ARPs with corresponding ratio of disordered residues. Gray bars are the percentage of overlapped MoRF and ARP segments over the total number of predicted MoRFs for the sequences containing all ARPs.

with the increase in disorder content, the relative number of MoRFs increased with the disorder content increase.

Out of total 104 allergens carrying potential MoRFs, about 50% are of inhalation type (pollen, mites, and fungi), about 40% occur in foods (nuts, seeds, and crustaceans), whereas the exposure route for the remaining allergens is through contact or sting (data not shown). A major part of these allergens fall into three protein families/superfamilies: seed storage proteins (prolamins and cupins, 30%), group V grass allergens (15%), and tropomyosins (12%; data not shown). Interestingly, no MoRFs were identified in allergens of either profilins or Bet v 1-homolog protein families, known to be connected with extensive cross-reactivity.

To investigate whether MoRFs found in the above-mentioned set of allergens, preferentially appear in antigen-antibody binding protein segments, a search for overlap patches between these entities and experimentally verified IgE-epitopes was conducted. Among 104 allergens, wherein MoRFs had been predicted, only six proteins were associated with IgE-epitopes, according to results derived from the SDAP⁶² web server (visited on 04/05/2011). In three of these allergens, a partial overlap between MoRFs and epitopes was observed, whereas the remaining three cases showed disjoint locations (data not shown). Relatively few allergens are, however, tied to

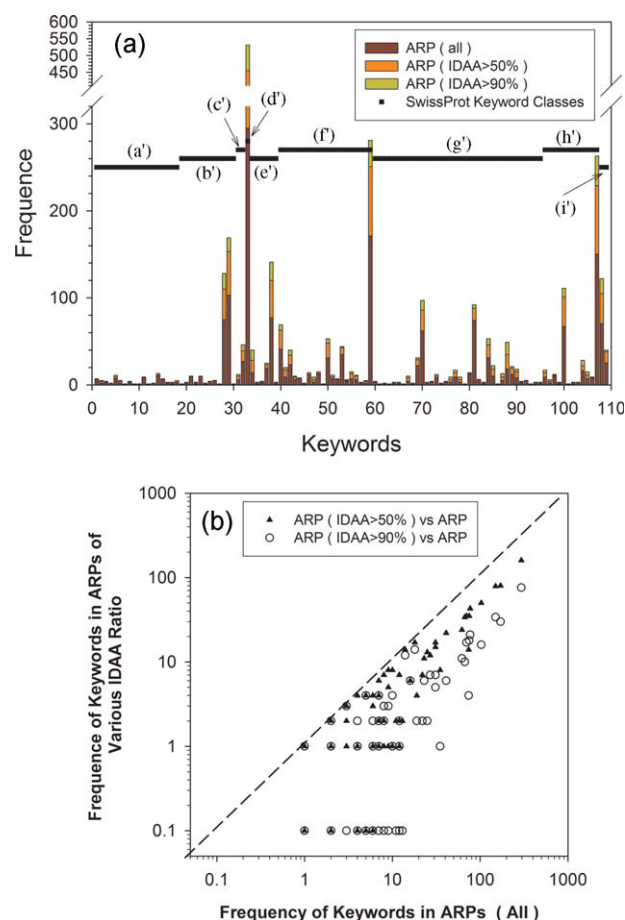


Figure 5

SwissProt keywords in nine classes and their occurrence frequency in various ARP subsets. The subsets include original ARP dataset, ARP subset with ratio of disordered amino acids (IDDA) more than 50% ($IDAA > 50\%$), ARP subset with 90% and more disordered residues ($IDAA > 90\%$). **a:** The x-axis is each keyword, whereas y-axis is the frequency of each keyword grouped into nine classes in one of the three subsets. The nine classes are: (a') biological process; (b') cellular component; (c') coding sequence diversity; (d') disease; (e') domain; (f') ligand; (g') molecular function; (h') PTM; and (i') technical term. **b:** X-axis is the frequency of each keyword in ARP dataset, and y-axis is the frequency of the same keyword in ARP ($IDAA > 50\%$) and ARP ($ARP > 90\%$). The symbols at $y = 0.01$ are artificial points indicating keywords disappeared in the disordered subsets. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Table I

Top 10 UniProt Keywords in Mainly Structured and Mainly Disordered ARP Subsets

Abundance of keywords in subset less than 10% IDAA				Abundance of keywords in subset more than 90% IDAA			
Keyword	Name	Class	Frequency	Keyword	Name	Class	Frequency
KW-0020	Allergen	Disease	253	KW-0020	Allergen	Disease	76
KW-0903	Direct protein sequencing	Technical term	153	KW-1015	Disulfide bond	PTM	34
KW-1015	Disulfide bond	PTM	141	KW-0903	Direct protein sequencing	Technical term	30
KW-0964	Secreted	Cellular component	90	KW-0732	Signal	Domain	21
KW-0732	Signal	Domain	69	KW-0963	Cytoplasm	Cellular component	18
KW-0963	Cytoplasm	Cellular component	64	KW-0002	3D-structure	Technical term	17
KW-0325	Glycoprotein	PTM	63	KW-0964	Secreted	Cellular component	16
KW-0378	Hydrolase	Molecular function	59	KW-0708	Seed storage protein	Molecular function	14
KW-0002	3D-structure	Technical term	56	KW-0175	Coiled-coil	Domain	12
KW-0568	Pathogenesis-related protein	Molecular function	43	KW-0378	Hydrolase	Molecular function	11

detailed IgE-epitope maps. Furthermore, most IgE-epitopes are assumed to be conformational, whereas the majority of those defined by experimental mapping are linear. Thus, our results are not conclusive as regards a direct implication of MoRFs in IgE-epitopes, but nonetheless suggestive of the converse. Notably, however, parts of the protein not directly involved in IgE binding—even those being remotely situated—typically have crucial impact on epitope conformation.

Association of ARPs with UniProt keywords

The distribution of various UniProt keywords in three different ARP subsets was investigated (Fig. 5). The three subsets were original ARP dataset; a subset of peptides with at least 50% disordered residues; and a compilation of peptides with at least 90% disordered residues. Arguably, this data segmentation should enhance prospects to identify an association between intrinsic disorder and keywords in ARPs. There are in total 109 keywords in nine distinct categories associated with ARPs. Clearly, all keyword categories—except for biological process—have at least one member being closely associated with ARPs [Fig. 5(a)]. Furthermore [Fig. 5(a), brown bars], the top five UniProt keywords identified as being associated with the standard ARP dataset are as follows: allergen (KW-0020), direct protein sequencing (KW-0903), disulfide bond (KW-1015), secreted (KW-0964), and signal (KW-0732). Higher content of disordered residues in the ARPs (at least 50% to more than 90%) reduced the absolute number of hits for each UniProt keyword, whereas their relative distribution did not change markedly.

To specifically map the association of UniProt keywords to ARP intrinsic disorder, we compared keywords associated with either structure-prone subsets or disorder-prone counterparts. Thus, two extreme subsets of the original ARP dataset (10% or less and more than 90% disordered residues, respectively) were accordingly interrogated (Table I, top 10 keywords). Structure-prone ARPs thus disclosed connection to glycoprotein (KW-

0325) and pathogenesis-related protein (KW-0568), neither of which appeared associated with the disorder-prone subset. Moreover, the latter subset had seed storage protein (KW-0708) and coiled-coil (KW-0175), which were both absent in the top list of structure-prone ARPs (Table I).

To reveal transitions of keyword distributions among ARP subsets of various disorder contents, a comparison of absolute hits on each keyword in various subsets was conducted [Fig. 5(b)]. Obviously, the general trends of distribution of symbols with more than 20 hits in the original ARP dataset are very similar across these subsets. This suggests that keywords indicated by these symbols may be “neutral” across various subsets. In other words, these keywords are not dictated by disorder/structure status of proteins, but rather by the number of samples in the dataset. However, there are also other trends on the distribution of hits for keywords having 20 or fewer hits in the original ARP dataset. Some symbols occur in a position at or close to the diagonal, whereas many other align with the abscissa. Symbols appearing on or near the diagonal represent keywords not changed by increased disorder content of ARPs; hence, they are mainly connected with disordered peptides. On the other hand, the parallel symbols should be keywords highly associated with structured peptides, because they disappear in ARPs of highly disordered character. Here, if the relative variation of keyword hit is less than 10%, the keyword is treated as unchanged between two subsets.

Clearly, keywords in structure-prone ARPs are totally different from those of disorder-prone counterparts (Table II). Compared with the original ARP subset, three UniProt keyword categories—coding sequence diversity, disease, and technical terms—are completely missing in this table. In the structure-prone subset, there are 12, 8, 0, 7, 16, and 3 keywords associated with biological process (Biol. Proc.), cellular component (Cellular), domain, ligand, molecular function (Mol. Funct.), and post-translational modification (PTM), respectively. In the disorder-prone subset, there are 1, 1, 3, 1, 6, and 3 keywords.

Table II

UniProt Keywords Associated with Structure-Prone and Disorder-Prone Subsets

Keywords mainly associated with structure-prone subset				Keywords associated with disorder-prone subset			
Keyword	Name	Class	Frequency	Keyword	Name	Class	Frequency
KW-0119	Carbohydrate metabolism	Biol. Proc.	6	KW-0653	Protein transport	Biol. Proc.	1
KW-0292	Fruit ripening	Biol. Proc.	2	KW-0272	Extracellular matrix	Cellular	1
KW-0376	Hydrogen peroxide	Biol. Proc.	1	KW-0175	Coiled-coil	Domain	14
KW-0410	Iron transport	Biol. Proc.	7	KW-0176	Collagen	Domain	1
KW-0422	Lactose biosynthesis	Biol. Proc.	1	KW-0812	Transmembrane	Domain	1
KW-0442	Lipid degradation	Biol. Proc.	8	KW-0389	IgE-binding protein	Ligand	3
KW-0443	Lipid metabolism	Biol. Proc.	1	KW-0062	Aspartic protease inhibitor	Mol. Funct.	1
KW-0485	Methanol utilization	Biol. Proc.	2	KW-0064	Aspartyl protease	Mol. Funct.	1
KW-0561	Oxygen transport	Biol. Proc.	11	KW-0505	Motor protein	Mol. Funct.	3
KW-0611	Plant defense	Biol. Proc.	6	KW-0518	Myosin	Mol. Funct.	3
KW-0624	Polysaccharide degradation	Biol. Proc.	2	KW-0626	Porin	Mol. Funct.	1
KW-0843	Virulence	Biol. Proc.	3	KW-0800	Toxin	Mol. Funct.	1
KW-0035	Amyloplast	Cellular	1	KW-0027	Amidation	PTM	2
KW-0052	Apoplast	Cellular	2	KW-0291	Formylation	PTM	1
KW-0134	Cell wall	Cellular	9	KW-0379	Hydroxylation	PTM	1
KW-0493	Microtubule	Cellular	1				
KW-0496	Mitochondrion	Cellular	4				
KW-0576	Peroxisome	Cellular	5				
KW-0934	Plastid	Cellular	1				
KW-1003	Cell membrane	Cellular	1				
KW-0186	Copper	Ligand	7				
KW-0342	GTP-binding	Ligand	1				
KW-0349	Heme	Ligand	12				
KW-0408	Iron	Ligand	13				
KW-0430	Lectin	Ligand	1				
KW-0464	Manganese	Ligand	6				
KW-0683	Retinol binding	Ligand	2				
KW-0022	Alpha-amylase inhibitor	Mol. Funct.	3				
KW-0031	Aminopeptidase	Mol. Funct.	1				
KW-0044	Antibiotic	Mol. Funct.	1				
KW-0049	Antioxidant	Mol. Funct.	1				
KW-0081	Bacteriolytic enzyme	Mol. Funct.	1				
KW-0295	Fungicide	Mol. Funct.	1				
KW-0413	Isomerase	Mol. Funct.	2				
KW-0418	Kinase	Mol. Funct.	3				
KW-0482	Metalloprotease	Mol. Funct.	1				
KW-0494	Milk protein	Mol. Funct.	3				
KW-0540	Nuclease	Mol. Funct.	1				
KW-0575	Peroxidase	Mol. Funct.	6				
KW-0652	Protein synthesis inhibitor	Mol. Funct.	1				
KW-0788	Thiol protease	Mol. Funct.	5				
KW-0789	Thiol protease inhibitor	Mol. Funct.	1				
KW-0929	Antimicrobial	Mol. Funct.	2				
KW-0336	GPI-anchor	PTM	1				
KW-0449	Lipoprotein	PTM	2				
KW-0961	Cell wall biogenesis/degradation	PTM	8				

Thus, the structure-prone subset is tied to many more keywords in biological process, cellular component, ligand, and molecular function, but none in domain, whereas that of the disorder-prone ARPs has the same number or more keywords in domain and PTM. The actual keywords in domain for the disorder-prone subset are coiled-coil (KW-0175), collagen (KW-0176), and transmembrane (KW-0812), the last being part of the cellular component category, according to UniProt taxonomy. Keywords in PTM for the disorder-prone dataset are amidation (KW-0027), formylation (KW-0291), and hydroxylation (KW-0379). Analysis of keywords with

higher occurrence also revealed useful information. In the structure-prone ARPs, carbohydrate metabolism (KW-0119), iron transport (KW-0410), lipid degradation (KW-0442), oxygen transport (KW-0561), plant defense (KW-0611), cell wall (KW-0134), peroxisome (KW-0576), copper (KW-0186), heme (KW-0349), iron (KW-0408), manganese (KW-0464), peroxidase (KW-0575), thiol protease (KW-0788), and cell wall biogenesis/degradation (KW-0961) occurred more than five times; whereas in disorder-prone subset, coiled-coil (KW-0175), IgE-binding protein (KW-0389), motor protein (KW-0505), and myosin (KW-0518) appeared more than three times.

CONCLUSIONS

The identification of disordered and structured segments, which were found to differ between allergenic and nonallergenic proteins, holds potential for opening a new inroad to deepened understanding on allergy mechanisms and for the identification of potential allergens. Structured segments largely adhere to the sequence–structure–function paradigm by assuming a unique and stable structure as determined by the primary sequence, whereas disordered segments can attain any structure among an ensemble of flexible conformations. PPIs involving disordered segments tend to have lower affinity and broad adaptability of specificity, whereas the converse applies to structured segments. Hence, disorder status of allergen protein segments is seemingly important for the understanding of mechanisms involved in the highly diverse set of reactions pertaining to this ailment.

In our current study, 559 sequences containing 2896 peptides (ARPs) were analyzed. By applying a MoRF-II predictor, 228 MoRFs were found in 104 allergen protein sequences. Among 228 MoRFs, 163 overlapped with their corresponding peptides in the ARP dataset. Hence, more than 70% of the predicted MoRFs in allergenic proteins were classified as a peptide in an ARP. As is well documented, MoRFs are consecutive short structure-prone segments within disordered regions and undergo disorder-structure transition on binding onto partners. Thus, our results suggest several important conclusions and hypotheses:

- Typically, a predominant part ($\sim 80\%$) of the respective amino acid sequences of protein allergens is structured.
- Many allergen proteins have distinct disordered regions.
- The MoRFs, present in an appreciable fraction of protein allergens, may play specific roles in IgE-mediated allergy.
- Because of the highly flexible conformations adopted by MoRFs, other flexible peptides with amino acid sequences not addressed in this study may acquire similar conformations.

Because of the highly complex mechanisms involved in IgE-mediated allergy and the diversity of allergens themselves, a consistent presence of IDRs in allergens was unexpected. Our analysis revealed that allergens are mainly ordered but an appreciable fraction thereof carries MoRF-type IDRs. Furthermore, a predominant part of MoRFs found in the protein allergen dataset was seen in ARPs, the latter set being highly enriched with regard to segments appearing in allergens. Our data also demonstrate contextual aspects of IDRs in the allergy in support of the conjectured implication in certain aspects of allergy. Last, a significant part of all ARPs was shown to

carry IDRs encompassing more than half of their respective amino acid residues. Especially these ARPs, as appearing in their respective protein allergens as well as those of the Class II type, may possess features of importance to part(s) of the triggering process including cross-reactivity which does not strictly adhere to expected patterns, that is as deduced by structural resemblance to a sensitizing antigen.

ACKNOWLEDGMENTS

This work was supported in part by the grant from the National Science Foundation (A.K.D. and V.N.U.), and the Program of the Russian Academy of Sciences for the “Molecular and Cellular Biology” (V.N.U.). The authors acknowledge the support of the IUPUI Signature Centers Initiative.

REFERENCES

1. Platts-Mills TA, Erwin EA, Woodfolk JA, Heymann PW. Environmental factors influencing allergy and asthma. *Chem Immunol Allergy* 2006;91:3–15.
2. Genuis SJ. Sensitivity-related illness: the escalating pandemic of allergy, food intolerance and chemical sensitivity. *Sci Total Environ* 2010;408:6047–6061.
3. Nauta AJ, Engels F, Knippels LM, Garssen J, Nijkamp FP, Redegeld FA. Mechanisms of allergy and asthma. *Eur J Pharmacol* 2008;585:354–360.
4. Ozdemir C, Akdis M, Akdis CA. T-cell response to allergens. *Chem Immunol Allergy* 2010;95:22–44.
5. Kay AB. Allergy and allergic diseases. Second of two parts. *N Engl J Med* 2001;344:109–113.
6. Aalberse RC, Akkerdaas J, van Ree R. Cross-reactivity of IgE antibodies to allergens. *Allergy* 2001;56:478–490.
7. Bonds RS, Midoro-Horiuti T, Goldblum R. A structural basis for food allergy: the role of cross-reactivity. *Curr Opin Allergy Clin Immunol* 2008;8:82–86.
8. Egger M, Mutschlechner S, Wopfner N, Gadermaier G, Briza P, Ferreira F. Pollen-food syndromes associated with weed pollinosis: an update from the molecular point of view. *Allergy* 2006;61:461–476.
9. Soeria-Atmadja D, Onell A, Kober A, Matsson P, Gustafsson MG, Hammerling U. Multivariate statistical analysis of large-scale IgE antibody measurements reveals allergen extract relationships in sensitized individuals. *J Allergy Clin Immunol* 2007;120:1433–1440.
10. Aalberse RC. Assessment of allergen cross-reactivity. *Clin Mol Allergy* 2007;5:2. doi:10.1186/1476-7961-5-2.
11. Schein CH, Ivanciuc O, Midoro-Horiuti T, Goldblum RM, Braun W. An allergen portrait gallery: representative structures and an overview of IgE binding surfaces. *Bioinform Biol Insights* 2010;4:113–125.
12. Ladics GS, Selgrade MK. Identifying food proteins with allergenic potential: evolution of approaches to safety assessment and research to provide additional tools. *Regul Toxicol Pharmacol* 2009;54(3 Suppl):S2–S6.
13. Tong JC, Tammi MT. Methods and protocols for the assessment of protein allergenicity and cross-reactivity. *Front Biosci* 2008;13:4882–4888.
14. Stadler MB, Stadler BM. Allergenicity prediction by protein sequence. *FASEB J* 2003;17:1141–1143.
15. Li KB, Issac P, Krishnan A. Predicting allergenic proteins using wavelet transform. *Bioinformatics* 2004;20:2572–2578.
16. Bjorklund AK, Soeria-Atmadja D, Zorzet A, Hammerling U, Gustafsson MG. Supervised identification of allergen-representative peptides for in silico detection of potentially allergenic proteins. *Bioinformatics* 2005;21:39–50.

17. Soeria-Atmadja D, Lundell T, Gustafsson MG, Hammerling U. Computational detection of allergenic proteins attains a new level of accuracy with in silico variable-length peptide extraction and machine learning. *Nucleic Acids Res* 2006;34:3779–3793.
18. Ivanciuc O, Garcia T, Torres M, Schein CH, Braun W. Characteristic motifs for families of allergenic proteins. *Mol Immunol* 2009;46:559–568.
19. Wright PE, Dyson HJ. Intrinsically unstructured proteins: re-assessing the protein structure–function paradigm. *J Mol Biol* 1999;293:321–331.
20. Uversky VN, Gillespie JR, Fink AL. Why are “natively unfolded” proteins unstructured under physiologic conditions? *Proteins* 2000;41:415–427.
21. Dunker AK, Lawson JD, Brown CJ, Williams RM, Romero P, Oh JS, Oldfield CJ, Campen AM, Ratliff CM, Hipps KW, Ausio J, Nissen MS, Reeves R, Kang C, Kissinger CR, Bailey RW, Griswold MD, Chiu W, Garner EC, Obradovic Z. Intrinsically disordered protein. *J Mol Graph Model* 2001;19:26–59.
22. Tompa P. Intrinsically unstructured proteins. *Trends Biochem Sci* 2002;27:527–533.
23. Uversky VN. Natively unfolded proteins: a point where biology waits for physics. *Protein Sci* 2002;11:739–756.
24. Tompa P. Structure and function of intrinsically disordered proteins. Boca Raton: Chapman & Hall/CRC Press; 2009.xxvii,331 pp.
25. Miller M. The importance of being flexible: the case of basic region leucine zipper transcriptional regulators. *Curr Protein Pept Sci* 2009;10:244–269.
26. Ma B, Nussinov R. Regulating highly dynamic unstructured proteins and their coding mRNAs. *Genome Biol* 2009;10:204. (doi:10.1186/gb-2009-10-1-204).
27. Longhi S. Nucleocapsid structure and function. *Curr Top Microbiol Immunol* 2009;329:103–128.
28. Jeffery CJ. Moonlighting proteins—an update. *Mol Biosyst* 2009;5:345–350.
29. He B, Wang K, Liu Y, Xue B, Uversky VN, Dunker AK. Predicting intrinsic disorder in proteins: an overview. *Cell Res* 2009;19:929–949.
30. Uversky VN, Dunker AK. Understanding protein non-folding. *Biochim Biophys Acta* 2010;1804:1231–1264.
31. Dunker AK, Oldfield CJ, Meng J, Romero P, Yang JY, Chen JW, Vacic V, Obradovic Z, Uversky VN. The unfoldomics decade: an update on intrinsically disordered proteins. *BMC Genomics* 2008;9(Suppl 2):S1. doi:10.1186/1471-2164-9-S2-S1.
32. Uversky VN. The mysterious unfoldome: structureless, underappreciated, yet vital part of any given proteome. *J Biomed Biotechnol* 2010;2010:568068. doi:10.1155/2010/568068.
33. Dunker AK, Brown CJ, Obradovic Z. Identification and functions of usefully disordered proteins. *Adv Protein Chem* 2002;62:25–49.
34. Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradovic Z. Intrinsic disorder and protein function. *Biochemistry* 2002;41:6573–6582.
35. Minezaki Y, Homma K, Kinjo AR, Nishikawa K. Human transcription factors contain a high fraction of intrinsically disordered regions essential for transcriptional regulation. *J Mol Biol* 2006;359:1137–1149.
36. Radivojac P, Iakoucheva LM, Oldfield CJ, Obradovic Z, Uversky VN, Dunker AK. Intrinsic disorder and functional proteomics. *Biophys J* 2007;92:1439–1456.
37. Eliezer D. Biophysical characterization of intrinsically disordered proteins. *Curr Opin Struct Biol* 2009;19:23–30.
38. Receveur-Brechot V, Bourhis JM, Uversky VN, Canard B, Longhi S. Assessing protein disorder and induced folding. *Proteins* 2006;62:24–45.
39. Romero P, Obradovic Z, Kissinger CR, Villafranca JE, Dunker AK. Identifying disordered regions in proteins from amino acid sequences. *IEEE Int Conf Neural Networks* 1997;1:90–95.
40. Romero P, Obradovic Z, Li X, Garner EC, Brown CJ, Dunker AK. Sequence complexity of disordered protein. *Proteins* 2001;42:38–48.
41. Oldfield CJ, Cheng Y, Cortese MS, Brown CJ, Uversky VN, Dunker AK. Comparing and combining predictors of mostly disordered proteins. *Biochemistry* 2005;44:1989–2000.
42. Oldfield CJ, Cheng Y, Cortese MS, Romero P, Uversky VN, Dunker AK. Coupled folding and binding with alpha-helix-forming molecular recognition elements. *Biochemistry* 2005;44:12454–12470.
43. Xue B, Dunbrack RL, Williams RW, Dunker AK, Uversky VN. PONDR-FIT: a meta-predictor of intrinsically disordered amino acids. *Biochim Biophys Acta* 2010;1804:996–1010.
44. Obradovic Z, Peng K, Vucetic S, Radivojac P, Brown CJ, Dunker AK. Predicting intrinsic disorder from amino acid sequence. *Proteins* 2003;53 (Suppl 6):566–572.
45. Le Gall T, Romero PR, Cortese MS, Uversky VN, Dunker AK. Intrinsic disorder in the Protein Data Bank. *J Biomol Struct Dyn* 2007;24:325–342.
46. Xie H, Vucetic S, Iakoucheva LM, Oldfield CJ, Dunker AK, Uversky VN, Obradovic Z. Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions. *J Proteome Res* 2007;6:1882–1898.
47. Pentony MM, Jones DT. Modularity of intrinsic disorder in the human proteome. *Proteins* 2010;78:212–221.
48. Cheng Y, Oldfield CJ, Meng J, Romero P, Uversky VN, Dunker AK. Mining alpha-helix-forming molecular recognition features with cross species sequence alignments. *Biochemistry* 2007;46: 13468–13477.
49. Mohan A, Oldfield CJ, Radivojac P, Vacic V, Cortese MS, Dunker AK, Uversky VN. Analysis of molecular recognition features (MoRFs). *J Mol Biol* 2006;362:1043–1059.
50. Dunker AK, Cortese MS, Romero P, Iakoucheva LM, Uversky VN. Flexible nets. The roles of intrinsic disorder in protein interaction networks. *FEBS J* 2005;272:5129–5148.
51. Uversky VN, Oldfield CJ, Dunker AK. Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling. *J Mol Recognit* 2005;18:343–384.
52. Xue B, Hsu W, Lee J, Lu H, Dunker AK, Uversky VN. SPA: short peptide analyzer of intrinsic disorder status of short peptides. *Genes to Cells* 2010;15:635–646.
53. Dunker AK, Obradovic Z, Romero P, Garner EC, Brown CJ. Intrinsic protein disorder in complete genomes. *Genome Inform Ser Workshop Genome Inform* 2000;11:161–171.
54. Xue B, Oldfield CJ, Dunker AK, Uversky VN. CDF it all: consensus prediction of intrinsically disordered proteins based on various cumulative distribution functions. *FEBS Lett* 2009;583:1469–1474.
55. Mohan A, Sullivan WJ, Jr, Radivojac P, Dunker AK, Uversky VN. Intrinsic disorder in pathogenic and non-pathogenic microbes: discovering and analyzing the unfoldomes of early-branching eukaryotes. *Mol Biosyst* 2008;4:328–340.
56. Iakoucheva LM, Brown CJ, Lawson JD, Obradovic Z, Dunker AK. Intrinsic disorder in cell-signaling and cancer-associated proteins. *J Mol Biol* 2002;323:573–584.
57. Cheng Y, LeGall T, Oldfield CJ, Dunker AK, Uversky VN. Abundance of intrinsic disorder in protein associated with cardiovascular disease. *Biochemistry* 2006;45:10448–10460.
58. Uversky VN. Amyloidogenesis of natively unfolded proteins. *Curr Alzheimer Res* 2008;5:260–287.
59. Uversky VN. Intrinsic disorder in proteins associated with neurodegenerative diseases. *Front Biosci* 2009;14:5188–5238.
60. Uversky VN, Oldfield CJ, Dunker AK. Intrinsically disordered proteins in human diseases: introducing the D2 concept. *Annu Rev Biophys* 2008;37:215–246.
61. Vacic V, Uversky VN, Dunker AK, Lonardi S. Composition profiler: a tool for discovery and visualization of amino acid composition differences. *BMC Bioinformatics* 2007;8:211. doi:10.1186/1471-2105-8-211.
62. Ivanciuc O, Mathura V, Midoro-Horiuti T, Braun W, Goldblum RM, Schein CH. Detecting potential IgE-reactive sites on food proteins using a sequence and structure database, SDAP-food. *J Agric Food Chem* 2003;51:4830–4837.