FUNCTIONAL ANNOTATION OF INTRINSICALLY DISORDERED DOMAINS BY THEIR AMINO ACID CONTENT USING IDD NAVIGATOR

ASHWINI PATIL
Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai Minato-ku, Tokyo 108-8639, Japan
Email: ashwini@hgc.jp

SHUNSUKE TERAGUCHI*
Host Defense Lab, WPI Immunology Frontier Research Center (IFReC), Osaka University,3-1 Yamadaoka, Suita, Osaka 565-0871, Japan
Email: teraguch@ifrec.osaka-u.ac.jp

HUY DINH
Systems Immunology Lab, WPI Immunology Frontier Research Center (IFReC), Osaka University, 3-1 Yamadaoka, Suita, Osaka 565-0871, Japan
Email: dinh@ifrec.osaka-u.ac.jp

KENTA NAKAI
Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai Minato-ku, Tokyo 108-8639, Japan
Email: knakai@ims.u-tokyo.ac.jp

DARON M STANDLEY
Systems Immunology Lab, WPI Immunology Frontier Research Center (IFReC), Osaka University, 3-1 Yamadaoka, Suita, Osaka 565-0871, Japan
Email: standley@ifrec.osaka-u.ac.jp

Function prediction of intrinsically disordered domains (IDDs) using sequence similarity methods is limited by their high mutability and prevalence of low complexity regions. We describe a novel method for identifying similar IDDs by a similarity metric based on amino acid composition and identify significantly overrepresented Gene Ontology (GO) and Pfam domain annotations within highly similar IDDs. Applications and extensions of the proposed method are discussed, in particular with respect to protein functional annotation. We test the predicted annotations in a large-scale survey of IDDs in mouse and find that the proposed method provides significantly greater protein coverage in terms of function prediction than traditional sequence alignment methods like BLAST. As a proof of concept we examined several disorder-containing

* The authors wish it to be known that, in their opinion, the first two authors contributed equally to this work.

proteins: GRA15 and ROP16, both encoded in the parasitic protozoa *T. gondii*; Cyclon, a mostly uncharacterized protein involved in the regulation of immune cell death; STIM1, a protein essential for regulating calcium levels in the endoplasmic reticulum. We show that the overrepresented GO terms are consistent with recently-reported biological functions. We implemented the method in the web server IDD Navigator. IDD Navigator is available at http://sysimm.ifrec.osaka-u.ac.jp/disorder/beta.php.

1. Introduction

Intrinsically disordered domains (IDDs) make up a significant portion of many eukaryotic proteins but present a unique challenge to structural bioinformatics-based functional annotation methods, which generally operate on the structure-function paradigm. For example, several groups, including ours, have used 3D modeling in combination with structural alignment to infer biological or biochemical functions of query sequences[1-3]. While it is in principal possible to follow this approach even in the absence of a well-defined structure by simply omitting the intermediate structural modeling step, and substituting structural alignment with sequence alignment, in practice this is problematic; the high rates of evolution coupled with higher than expected correspondence between low complexity regions (LCRs) and IDDs[4], along with the high false positive rate for LCRs in conventional sequence similarity scores make it difficult to quantify the significance of the resulting hits.

We recently investigated two alternatives to conventional sequence alignment in order to extract functional information from IDD sequences[5]. First, we examined short sequence motifs that were over-represented in a set of predicted IDDs. The over-represented motifs corresponded to amino acid repeats in general, but it was not possible to associate them with known functions using Pfam domains or Gene Ontology (GO) annotations. Second, we investigated the similarity between the overall amino acid composition of IDDs using a simple histogram distance method. Surprisingly, some structured Pfam domains associated with proteins that had been clustered only by the amino acid composition of their IDD regions were highly overrepresented compared with a random grouping of IDDs. This presents the possibility of IDDs with similar amino acid composition sharing a related function. Motivated by these results, we constructed a web server that allows a database of predicted IDDs to be searched by several similarity scores that are functions of the amino acid composition for a given IDD. Highly similar IDDs are returned along with their Pfam and GO annotations. The significance of the occurrence of each annotation can then be quantified by comparison with a sample of background IDDs. While the proposed approach is not expected to be as sensitive or as specific as is sequence alignment of structured domains, it can nevertheless have practical utility, especially when used in conjunction with structure-based functional annotations or in cases where no alternatives for function prediction exist. To demonstrate its utility, we carried out a large-scale survey of the predicted GO annotations as well as small-scale analysis of 4 IDD-containing proteins.

2. Methodology

2.1. *Preparation of IDD dataset*

64,322 amino acid sequences of mouse proteins were downloaded from UniProtKB[6]. A representative set of 18,126 non-redundant protein sequences was prepared by clustering at 40% sequence identity using the cd-hit program[7]. For each sequence, IDDs were predicted using the Disopred2 program[8] and 22,057 predicted IDDs of length greater than 30 were retained for analysis. In IDD Navigator, these IDD sequences and the annotations of the corresponding proteins were used as a database for finding similar IDDs and predicting functions for each query IDD.

2.2. *Similarity scores*

In this work, we predicted functions of each query IDD based on the most similar IDDs in the above IDD dataset. We defined similarities for a pair of IDDs through similarity scores. In our previous work, we introduced a similarity score between IDDs based on the frequency of amino acid residues[5]. In the current study, we have defined two similarity scores, which are described below. For each query, IDD Navigator calculates similarity scores against the above 22,057 stored IDDs and by default shows the top 100 most similar IDDs.

2.2.1. *Similarity score based on Euclidean distance*

The frequency $f_i(a)$ of an amino acid residue a in the sequence of IDD i is defined as follows:

$$f_i(a) = \frac{N(a)}{len_i} \tag{1}$$

Here, $N(a)$ is the number of the amino acid residues of type a in the sequence and len_i is the length of the sequence. The Euclidean distance based similarity score between sequences i and j is defined by

$$sim_{i,j} = -100\sqrt{\sum_{a=1}^{20}\left(f_i(a) - f_j(a)\right)^2} \tag{2}$$

The distance was multiplied by a negative number in order to convert it to a similarity score. This similarity score performs comparably with the Gaussian similarity used in the previous work (data not shown).

2.2.2. *BLAST score*

Additionally, we used the protein BLAST (blastp) score for comparison purposes. Blastp was run on each IDD against the database of all 22,057 IDDs with an e-value cutoff of 0.01. The BLAST score was used in the same manner as the Euclidian distance-based score.

2.3. *Pfam domain and Gene Ontology term prediction*

By assuming that there is a correlation between the similarity scores defined above and similarity in functions of the corresponding proteins, IDD Navigator provides a prediction of the functions of

a query IDD. We counted the numbers of Pfam domains and GO terms of unique proteins associated with a list of the most similar IDDs (top 100) evaluated by one of the similarity scores. The significantly overrepresented Pfam domains and GO terms constitute the primary output of IDD Navigator. The statistical significance of the predicted functions was estimated by using hypergeometric distribution functions. Here, we modeled the probability of having a particular function by randomly selecting the same number of proteins from 18,126 representatives, counting the fraction annotated by the function in question, and then calculating the corresponding p-values. Thus, the annotations of all proteins with IDDs were used as a reference dataset.

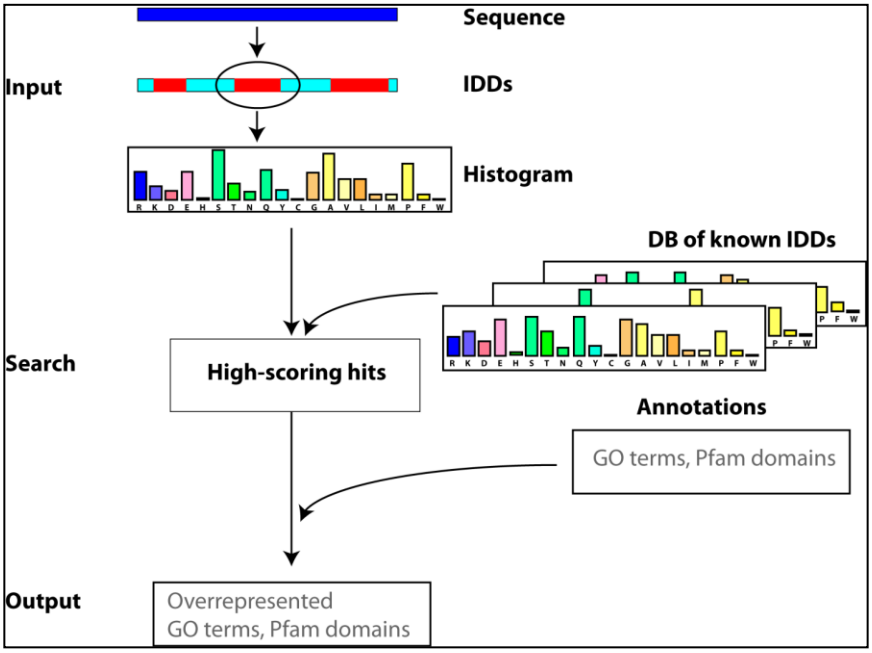

Figure 1. Flowchart of the of the web server. An IDD can be input or IDD prediction can be performed on a full-length sequence. A histogram is then computed for the input IDD, and it is compared with a database of known IDDs. High-scoring hits are then extracted and annotations from GO and Pfam are collected. The overrepresented GO terms and Pfam domains are then identified and output.

2.4. *Evaluation of function prediction*

From the initial dataset of 18,126 mouse proteins, 5,164 had predicted IDDs and Gene Ontology term annotations in UniProt and were used in evaluating the prediction performance of IDD Navigator. For each IDD within the 5,164 proteins, GO terms were predicted as those significantly enriched ($p < 0.01$) among the top 100 most similar IDDs based on a particular (Euclidean or BLAST) score. We tested the similarity of GO terms predicted for each IDD in a protein with the actual GO terms assigned to it. For proteins with multiple IDDs, unique predicted GO terms from all IDDs were combined and compared to the actual GO annotations of the protein. In order to calculate the significance of our results, we compared the predicted GO terms with random terms. We prepared a list of unique GO terms from the actual GO annotations of the 5,614 mouse proteins. We then randomly picked GO terms for each IDD that were equal in number to the predicted GO terms for that IDD. The semantic similarity was calculated between the actual GO terms and the list of randomly selected GO terms.

Clusters of IDDs with similar amino acid content were made based on the Euclidean distance similarity score using hierarchical clustering with Ward's method. The number of clusters was empirically set at 10. Average semantic similarity between predicted and real GO terms was calculated for each cluster and compared to those of random GO terms. Amino acid propensity for amino acid *a* in the i^{th} cluster was calculated as follows:

$$P_{ai} = \frac{N_{ai}}{N_a} \quad (3)$$

where N_{ai} is the count of the amino acid *a* in cluster *i* and N_a is the total count of amino acid *a* in all clusters. The z-score for the propensity of each amino acid was calculated over its average propensity in all clusters and plotted as a heatmap in figure 4.

Semantic similarity between lists of GO terms was calculated using the R package GOSemSim[9]. GO terms are arranged in the form of a directed acyclic graph (DAG), and GOSemSim calculates the similarity between two lists of GO terms using their location in the DAG and their relationship with the ancestor terms. GOSemSim assigns a value between 0 and 1 with higher values indicating greater similarity between groups of GO terms. The predictions of Cellular Component (CC), Molecular Function (MF) and Biological Process (BP) terms were evaluated independently.

2.5. *Web server*

IDD Navigator can be accessed at http://sysimm.ifrec.osaka-u.ac.jp/disorder/beta.php. A flowchart of steps performed by the web server is shown in Figure 1. The histograms used to score the similarity between IDDs are represented as colored bar graphs. Several methods have been provided to evaluate the similarity score between IDDs based on amino acid content.

3. Results and Discussion

3.1 IDD Navigator Function prediction

In order to determine if the significantly enriched GO terms given by IDD Navigator for a query IDD can be used as function predictions, we first evaluated the accuracy of the predictions with p-value < 0.01. We compared the semantic similarity between the real GO terms and those predicted by IDD Navigator for a set of 5,164 mouse proteins using the Euclidean distance score. Since some proteins have several IDDs, we also tested the performance of the combined list of GO terms predicted for all IDDs within a protein. Finally we compared these results with the similarity of an equal number of random GO terms assigned to each protein.

Figure 2 shows the results of the evaluation. GO terms predicted by the Euclidian distance score show greater semantic similarity to the real GO terms of proteins, on average, than randomly assigned GO terms ($p << 0.001$, t-test). This result is consistent across all categories of GO terms. Combined predictions from all IDDs in a protein performed best, though those predicted using single IDDs showed an almost equivalent performance. Among the categories of GO terms, CC terms were predicted with the greatest accuracy, followed by MF and BP terms, respectively. The lower performance of BP terms was possibly due to the large number of GO terms present in this category. Based on these results we concluded that the GO terms given by IDD Navigator could be used as valid function predictions for unannotated proteins. We further, checked the relationship

between the length of the IDD and the accuracy of prediction of the GO terms but were unable to find a significant correlation (GO BP: 0.10; GO MF: 0.08; GO CC: 0.12).

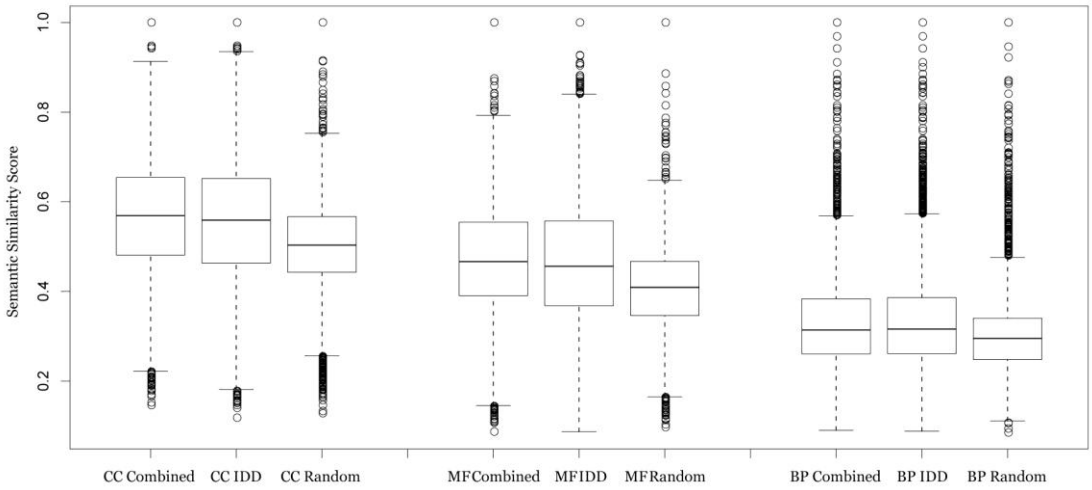

Figure 2. Average semantic similarity between Euclidian distance-predicted and actual GO terms in Cellular Component (CC), Molecular Function (MF) and Biological Process (BP) categories. Combined indicates a combination of all unique GO terms predicted for all IDDs within a protein. IDD denotes GO terms predicted for a single IDD in a protein. Random shows the performance of the same number of random GO terms assigned to a protein.

3.2 Comparing different methods in IDD Navigator

We compared the function prediction performance of the Euclidean distance method and the BLAST score in IDD Navigator. The Euclidean distance method represents the similarity of IDDs by amino acid content, irrespective of the sequence. On the other hand, the BLAST score identifies similar IDDs by direct sequence alignment to the query IDD. The BLAST score option in IDD Navigator is equivalent to performing a BLAST search using an IDD against all the IDDs in the current dataset with low complexity filtering turned off and predicting a function based on the top 100 hits obtained.

In Figure 3, the Euclidian distance score is compared with the BLAST score for each GO term. For a given GO category (CC, MF, BP), Figure 3 shows the combined score (combining all the IDDs from one protein) for Euclidian distance and BLAST, the IDD scores (where IDDs for a given protein are treated separately), and a randomized score. The randomized score corresponds to the semantic similarity between the actual GO terms of the proteins and randomly chosen GO terms that are equal in number to those predicted by the IDD Navigator using the Euclidean and BLAST scoring schemes. The two methods have different randomized scores because of the differences in the number of GO terms predicted by each. IDD Navigator using the BLAST score predicts considerably fewer GO terms as significantly over-represented than using the Euclidean score. Due to the arrangement of the GO terms in the form of a DAG and the method used by GOSemSim to calculate the semantic similarity, the similarity score is dependent on the number of terms, often increasing as the number of terms increases. In order to address this drawback of the

semantic similarity scoring, we calculated separate randomized similarity scores for predictions made using the two methods.

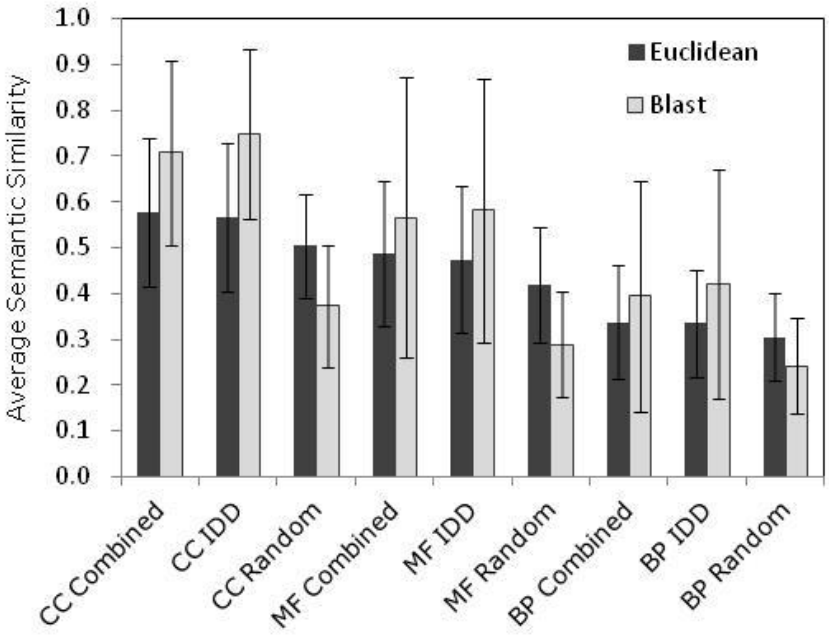

Figure 3. Comparison between Euclidian distance and BLAST scores. Average semantic similarities between predicted and real GO terms for mouse proteins are shown for the two scoring methods along with randomized values. Euclidean: Euclidean distance score for amino acid composition similarity, BLAST: BLAST score for IDD sequence similarity. Term definitions same as in Figure 2.

The BLAST semantic similarity scores are closer to the actual terms than those predicted by the Euclidean distance method. However, there is a large difference in the coverage of proteins for the two types of methods. Significantly enriched GO terms are obtained for 4,383 proteins using the Euclidean method, while over-represented GO terms were identified for IDDs in only 27 proteins using the BLAST score at a significance threshold of 0.01. The very small number of BLAST hits is not surprising. Proteins within the mouse genome that are not obvious homologs (i.e., do not form clusters in the cd-hit step) do not, generally, have sequentially similar IDDs. The high coverage of proteins for which GO terms can be predicted using the Euclidian distance score, on the other hand, highlights the utility of IDD Navigator in novel function prediction.

Several methods have been proposed to assign probable GO terms to proteins without annotations[10-12]. However, these methods rely on BLAST or PSI-BLAST either partially or completely. Predicting the function of proteins with large disordered domains in the absence of conserved or annotated domains and annotated homologs is currently an open problem. Figure 2 indicates that IDD Navigator can provide function associations in at least a subset of such cases. It can also potentially be used to assign new functions to proteins based on the IDDs that they contain. Although functional annotations have been associated with specific IDDs[13, 14] and function prediction using intrinsic disorder in proteins has been studied before[15-17], IDD Navigator uses a novel strategy based on amino acid content to assign probable GO terms to unannotated proteins, and is thus complementary to existing approaches.

3.3 Function prediction for IDD clusters

Our method can also be used to cluster the IDDs into groups that may potentially be associated with specific functions similar to a previous study identifying "flavors of disorder" [18]. To see if this is the case, we clustered the 22,057 IDDs into 10 clusters based on their amino acid similarity. Figure 4 shows the amino acid enrichment for each cluster. Each cluster has a distinct pattern of over-represented amino acids. For each cluster, we calculated the average semantic for each IDD similarity between the predicted GO terms and its observed GO terms. Table 1 shows the numerical results.

The GO term similarity between those predicted by IDD Navigator compared to the observed terms is better than random in all the clusters for all 3 types of GO terms. The GO similarity score in Cluster 4 are particularly high. The IDDs in this cluster are enriched in charged residues, especially Lys, and are primarily parts of ribosomal proteins based on GO term enrichment analysis [19]. However, these results also indicate that the performance of function prediction varies with the amino acid content of the IDD. A good performance for certain clusters may be attributed to the fact that the clusters contain fewer IDDs or proteins containing the IDDs in these clusters are better annotated. On the other hand, a poor similarity score between the observed and predicted GO terms may be the result of the IDDs participating in multiple functions despite similar amino acid content. The performance of the function prediction by IDD Navigator may also change with the number of clusters used to partition the dataset. We did not, for example, find 3 major clusters using our methods, so a direct comparison with the 3 flavors of disorder proposed by Vucetic et al. was not straightforward.

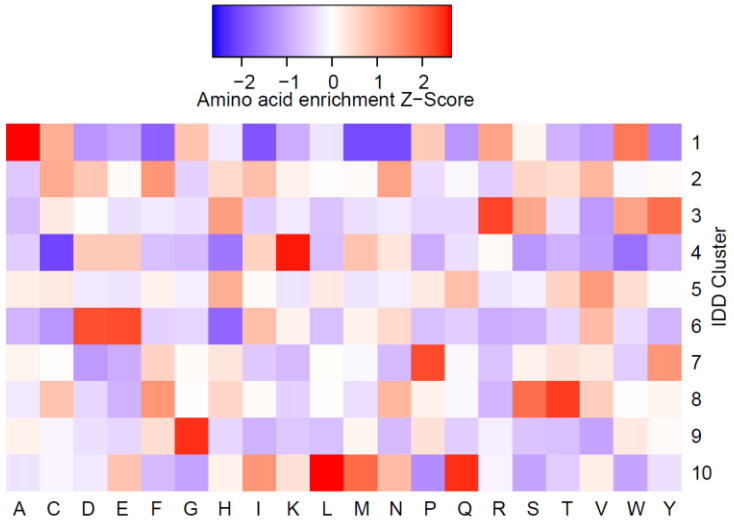

Figure 4. Amino acid enrichment in 10 clusters of IDDs. Each square represents the amino acid enrichment in the form of a z-score.

Cluster	***Average Semantic Similarity***		
	BP	**MF**	**CC**
1	0.339	0.463	0.562
2	0.329	0.465	0.562
3	0.330	0.448	0.528
4	0.354	0.512	0.627
5	0.325	0.468	0.548
6	0.354	0.456	0.588
7	0.346	0.514	0.586
8	0.337	0.473	0.566
9	0.347	0.480	0.548
10	0.345	0.490	0.603
Average	0.335	0.474	0.565
Random	0.305	0.419	0.504

Table 1. Average semantic similarity scores between predicted and real GO terms for IDDs in mouse proteins assigned to 10 clusters based on amino acid content similarity by Euclidean distance score. Average and random values shown in gray.

3.4 Case Studies

The ultimate test of IDD Navigator is to analyze a functionally uncharacterized sequence *de novo* and validate the predicted function experimentally. Since such a test is beyond the scope of the current work, we have selected 4 IDD-containing proteins whose biological function has only recently been reported. Each protein was submitted using default options. In the case of GRA15, some subsequent adjustment of the Disopred false-positive threshold was also used.

3.4.1 GRA15 from T. gondii

GRA15 is a ~500 amino acid polymorphic protein localized in the dense granules of the parasitic protozoa *T. gondii*, and secreted into host cells upon infection. The effect of GRA15 on the host immune system depends on the particular *T. gondii* strain (types I-III). Recently, the effects of type II GRA15 (GRA15-II) on host immune signaling pathways have been described[20, 21]. No hits to GRA15 were found using conventional sequence alignment searches[21]. The sequence of GRA15-II lacking a 51-residue N-terminal signaling peptide was submitted to IDD Navigator using the Disopred option. A single 475-residue IDD was predicted for residues 24-499. The top hits were associated with transcription factors PAX2, PAX7, and A2A410. The next most significant hit was to an AXH-containing protein, which interacts with CIC, a transcriptional repressor. There were also a number of hits involved in signal transduction. The GO molecular function terms "protein binding" (p-value 2.4e-4) and "kinase activity" (p-value 4.9e-4) scored highly, while "cell fate determination" scored highest in the biological process category (p-value 0.8e-3). Taken together, the GRA15-II results are consistent with a role in signal transduction, transcription, or both. Although the biochemical function of GRA15-II has not yet been elucidated, it was recently revealed that its biological role is to activate the transcription factor NF-κB in infected hosts[16]. GRA15-II-mediated NF-κB activation in macrophages inhibited apoptosis, and increased cell migration. Moreover, it was argued that GRA15-II is likely to act in a complex with the kinase IKK and signaling protein TRAF6[21], which are important regulators of the NF-κB signaling pathway. GRA15-II-mediated NF-κB activation in macrophages inhibited apoptosis, and increased cell migration[20, 21]. These known biological functions of GRA15-II are overall consistent with the IDD Navigator results. One potential application of the IDD Navigator query would be to suggest putative protein-protein interaction sites. However, in order to do this, we had to first lower the Disopred2 false-positive threshold from the default value of 5% to 2%, in order to obtain several smaller, but higher-confidence IDDs. This resulted in three distinct IDDs (residues 34-206, 209-389, 402-499). Of these, the last one was of particular interest, as the top GO annotation was "apoptosis" which includes both positive and negative regulators of cell death. The specific hits were comprised of 6 proteins, including "NF-κB-interacting protein 1". Interestingly, the region of greatest divergence between type I/III and II GRA15 consists of an 84–amino acid indel near the C-terminus[21]. This finding is consistent with the observation that type I/II *T. gondii* has been found to *suppress* NF-κB activity, suggesting that the C-terminal IDD may be critical for differentiating effects on host immune cell responses. This result shows that IDD Navigator results are sensitive to the domain boundaries, which is also true for structure-based function prediction. Fortunately, the false-positive threshold in the Disopred program is a convenient way to vary the boundaries.

3.4.2 Cyclon from M. musculus

Cyclon is a coiled coil-containing protein that is known to positively regulate expression of Fas, an immune cell surface protein that mediates activation-induced cell death [22, 23]. However, a direct role in Fas transcription has not been shown, and Cyclone lacks any known DNA binding domains. When we submitted the Cylon sequence to IDD Navigator, we found that GO terms relating to epigenetic regulation dominated the list. For example, in the molecular function
Category, the GO top terms and associate p-values included "histone methyltransferase activity (H3-K4 specific)" (2.4e-4); "transcription corepressor activity" (1.4e-3); "chromatin binding" (2.2e-3); "histone-lysine N-methyltransferase activity" (2.8e-3). The top biological process terms were: "chromatin modification" (1.2e-4); "cellular component organization" (1.6e-3); "peptidyl-lysine methylation" (1.6e-3). The top cellular component terms were "histone deacetylase complex" (0.7e-5) and "histone methyltransferase complex" (0.8e-4). A BLAST search against the nr database resulted in a number of hits to uncharacterized proteins in vertebrates and insects. The one exception was the chromatin assembly factor-I p150 from Culex quinquefasciatus (southern house mosquito). Although this annotation could not be confirmed by further BLAST or literature searches, it agreed qualitatively with the IDD Navigator result. The coiled-coil domains are most likely involved in protein-protein interactions, so Cyclon may act as an adaptor that recruits proteins involved in protein remodeling or histone modification. A potentially useful direction would be to systematically screen all such proteins for cyclon-like IDD regions, as this subset may contain epigenetic regulators of Fas gene expression.

3.4.3 STIM1 from M. musculus

STIM1 is a calcium-sensing protein that spans the ER membrane. The N-terminal luminal portion contains the calcium-binding domain and the cytoplasmic C-terminal portion is predicted to contain coiled-coil domains and two IDDs. The first IDD, but not the second, was found to have a significant (p-value 4e-11) over-representation of the GO cellular component term "Cytoskeleton". This result is potentially interesting, as a recent report showed that microtubules affect STIM1-mediated calcium entry[24] and in another report, STIM1 was shown to affect the organization of microtubules [25]. When the second IDD was submitted to IDD Navigator, the cellular component term "Cytoplasm" was over-represented with a p-value of 4e-7, which is consistent with the known localization of this domain. Both findings are relevant to one of the key functions of STIM1, the formation of ER projections known as 'puncta' that grow toward the plasma membrane when calcium stores are depleted. It is already known that calcium depletion promoted STIM1-STIM1 interactions on the luminal side of the ER membrane [26]. Although speculative at this point, the coiled-coils and IDDs may interact directly with components of the cytoskeleton upon calcium depletion in order to induce ER membrane remodeling.

3.4.4 ROP16 from T. gondii

Like GRA15, Rhoptry protein 16 (ROP16) is encoded in *T. gondii*, and secreted into host cells upon infection as part of the parasite's counter-defense against host immunity.
The N-terminal 357 residues of ROP16 are predicted to be disordered. When we submitted residues 1-357 to IDD Navigator, we found no overrepresented GO terms for biological function. However, the cellular component term "cell junction" was over-represented with a p-value of 2.6e-8. This result is potentially relevant since ROP proteins are excreted through cell-cell junctions

formed between host and parasite. The N-terminal portion of ROP16 has not been characterized, and the IDD Navigator results suggest that the IDD region may be involved in the initial secretion into the host cell.

4. Conclusions

In this report we have described an extension of our earlier work on IDD clustering toward the goal of functionally annotating IDDs. The large-scale survey results indicate clearly that, on average, GO terms that are overrepresented in IDD Navigator hits compared with background values are closer to true GO terms than a random selection of IDDs. This, in turn, implies that the underlying scoring function can pick up some information that is encoded in the IDD sequences without using sequence alignment. Distinguishing useful information from background noise, however, requires careful interpretation, as the 4 examples show. Submitting the whole GRA15-II sequence and running with default parameters produced results that are, in general, consistent with the known biological function of GRA15-II: the protein appears to function as an adaptor and may effect transcription or signal transduction pathways; however, the IDD Navigator hits did not give a clear clue as to which pathways are effected, or which parts of the protein are involved in specific protein-protein interactions. When the Disopred2 false-positive threshold was lowered from 5% to 2%, however, individual domains appeared with sequence-specific associated GO terms and Pfam domains. The appearance of apoptosis as the top GO term, with a specific hit to an NF-κB-interacting protein was encouraging, given recent reports that GRA15 suppresses apoptosis by activating NF-κB. However, it must be acknowledged that, in the absence of some knowledge of basic biological function, it would be difficult to isolate this hit from the others. The cylon result was somewhat more straightforward to interpret due to the dominance of hits related to epigenetic control of gene expression. Interaction with histone modifying proteins would explain the protein's role in regulation of Fas. However, this hypothesis requires further experiment before it can be validated. Similarly, the hits to STIM1 and ROP16 help to generate hypotheses. In the case of STIM1, the first IDD might interact with microtubules, which would be consistent with the coiled-coil domains predicted nearby. In the case of ROP16, the IDD might be is necessary for migrating through the cell-cell junction. Fortunately, each of these hypotheses can be experimentally tested, and we are currently collaborating with experimental groups to validate our predictions.

Future versions of IDD Navigator will include all sequences from Uniref100, allowing better function predictions across multiple species. In order to prepare and maintain such a large database of IDDs and also to process query sequences rapidly, the disorder prediction calculation time must be reduced. For this reason, in the future, we plan to incorporate faster disorder prediction methods[27]. We also aim to improve the scoring functions, which are highly simplistic at this stage.

5. Acknowledgements

We would like to thank M. Yamamoto for helpful discussions concerning GRA15, and H. Fujii for advice regarding Cyclon function. Computation time was partly provided by the Super Computer System, Human Genome Center, Institute of Medical Science, University of Tokyo. This research

was partially supported by the Japan Society for the Promotion of Science (JSPS) through its "Funding Program for World-Leading Innovative R&D on Science and Technology (FIRST Program).

References

1 J. Dundas, L. Adamian, and J. Liang, J Mol Biol **406**, 713 (2010).
2 A. Roy, A. Kucukural, and Y. Zhang, Nat Protoc **5**, 725 (2010).
3 D. M. Standley, R. Yamashita, A. R. Kinjo, et al., Bioinformatics **26**, 1258 (2010).
4 M. A. DePristo, M. M. Zilversmit, and D. L. Hartl, Gene **378**, 19 (2006).
5 S. Teraguchi, A. Patil, and D. M. Standley, BMC Bioinformatics **11 Suppl 7**, S7 (2010).
6 M. Magrane and U. Consortium, Database (Oxford) **2011**, bar009 (2011).
7 W. Li and A. Godzik, Bioinformatics **22**, 1658 (2006).
8 J. J. Ward, J. S. Sodhi, L. J. McGuffin, et al., J Mol Biol **337**, 635 (2004).
9 G. Yu, F. Li, Y. Qin, et al., Bioinformatics **26**, 976 (2010).
10 M. Chitale, T. Hawkins, C. Park, et al., Bioinformatics **25**, 1739 (2009).
11 T. Hawkins, S. Luban, and D. Kihara, Protein Sci **15**, 1550 (2006).
12 W. T. Clark and P. Radivojac, Proteins **79**, 2086 (2011).
13 H. Xie, S. Vucetic, L. M. Iakoucheva, et al., J Proteome Res **6**, 1882 (2007).
14 A. K. Dunker, C. J. Brown, J. D. Lawson, et al., Biochemistry **41**, 6573 (2002).
15 A. Lobley, M. B. Swindells, C. A. Orengo, et al., PLoS Comput Biol **3**, e162 (2007).
16 B. Meszaros, I. Simon, and Z. Dosztanyi, PLoS Comput Biol **5**, e1000376 (2009).
17 A. Mohan, C. J. Oldfield, P. Radivojac, et al., J Mol Biol **362**, 1043 (2006).
18 S. Vucetic, C. J. Brown, A. K. Dunker, et al., Proteins **52**, 573 (2003).
19 W. Huang da, B. T. Sherman, and R. A. Lempicki, Nucleic Acids Res **37**, 1 (2009).
20 K. D. Jensen, Y. Wang, E. D. Wojno, et al., Cell Host Microbe **9**, 472 (2011).
21 E. E. Rosowski, D. Lu, L. Julien, et al., J Exp Med **208**, 195 (2011).
22 A. Hoshino and H. Fujii, FEBS Lett **581**, 975 (2007).
23 S. Saint Fleur, A. Hoshino, K. Kondo, et al., Blood **114**, 1355 (2009).
24 C. Galan, N. Dionisio, T. Smani, et al., Biochem Pharmacol **82**, 400 (2011).
25 Z. Hajkova, V. Bugajev, E. Draberova, et al., J Immunol **186**, 913 (2010).
26 P. B. Stathopulos, L. Zheng, G. Y. Li, et al., Cell **135**, 110 (2008).
27 B. Xue, R. L. Dunbrack, R. W. Williams, et al., Biochim Biophys Acta **1804**, 996 (2010).