Disordered RNA binding region prediction with DisoRDPbind

Christopher J Oldfield[1], Zhenling Peng[2] and Lukasz Kurgan[1]*

[1]Department of Computer Science, Virginia Commonwealth University, Richmond, VA 23284, USA

[2]Center for Applied Mathematics, Tianjin University, Tianjin, 300072, P. R. China

*Corresponding author: lkurgan@vcu.edu

Summary

RNA chaperone activity is one of the many functions of intrinsically disordered regions (IDRs). IDRs function without the prerequisite of a stable structure. Instead, their functions arise from structural ensembles. A common theme in IDR function is molecular recognition; IDRs mediate interactions with other proteins, RNA, or DNA. Many computational methods are available to predicted IDRs from protein sequence, but relatively few are available for predicting IDR functions. Available methods focus on protein-protein interactions. DisoRDPbind was developed to predict several protein functions including interactions with RNA. This method is available as a user-friendly web interface, located at http://biomine.cs.vcu.edu/servers/DisoRDPbind/. The development and architecture of DisoRDPbind is briefly presented and its accuracy relative to other RNA binding residue predictors is discussed. We explain usage of the web interface in detail and provide an example of prediction results and interpretation. While DisoRDPbind does not identify RNA chaperones directly, we provide a case study of an RNA chaperone, HCV core protein, as an example of the method's utility in the study of RNA chaperones.

Keywords

Intrinsic disorder; protein-RNA interactions; intrinsically disordered regions; molecular recognition.

1. Introduction

RNA chaperone activity is one of the many functions of intrinsically disordered proteins (IDPs) and intrinsically disordered regions (IDRs) [1]. The sequences of IDPs and IDRs are self-insufficient to form a stably folded structure in isolation and instead exist as structural ensembles, where structures vary over time and over population [2-4]. Available sequence analysis methods accurately predict the locations of IDRs from protein sequence [5-15]. These methods estimate that IDPs and IDRs are prevalent in proteomes [16-23], particularly in Eukaryotes where 25-40% of their proteins contain significant IDRs [16,21]. Despite lack of stable structure, IDPs and IDRs perform many and varied biological functions [24-26,21]. Among their functions, IDPs play an important role in mediating molecular interactions by binding to proteins and nucleic acids [21,27-37]. In particular, many IDPs are known to function as RNA chaperones [38,39]. While the potential for IDRs to recognize RNA is well known, predicting novel RNA binding IDPs from sequence using many of the existing tools [40-48] is problematic; the majority of these tools have been developed for structured proteins interacting with RNA.

While no specific IDP chaperone prediction method is available to date, many studies have demonstrated that IDP function can be predicted from protein sequence [49-51,15,8,52-55]. Until recently, most studies focused on the interaction between IDPs and other proteins, and several methods

are available for prediction of protein recognition regions within IDRs [8]. Less attention has been paid to other types of IDP functions, including interaction with DNA and RNA. To fill this gap, we developed DisoRDPbind, a method that simultaneously predicts protein, DNA, and RNA interacting regions within IDRs [56,57]. This method predicts each type of function separately, allowing identification of the type of interaction partner for each predicted region. Additionally, DisoRDPbind predicts interactions at residue-level resolution, allowing identification of the protein regions responsible for each prediction type. Performance of the method is significantly better than other available methods for RNA interaction region predicted when applied to IDPs [56].

While not specific to RNA chaperones, DisoRDPbind is a useful tool for identification of novel intrinsically disordered RNA chaperones when combined with additional function information. As an example, we examine the Hepatitis C virus (HCV) core protein. This protein is multifunctional, playing a structural role in capsid formation and RNA organization [58], as well as serving as an RNA chaperone [59,60]. The RNA chaperone activity is located in the intrinsically disordered N-terminal region of HCV core protein [59] (Figure 1). The main idea behind methods such as DisoRDPbind is to predict these disordered RNA binding regions directly from the amino acid sequence. We demonstrate that the residue-level DisoRDPbind predictions correctly identify residues in this region of the input protein chain as intrinsically disordered and RNA binding (see Case Study).

The DisoRDPbind method is publicly available as a user-friendly webserver. Further, it was designed as a high-throughput method that can predicted entire proteomes in a matter of hours. Here we briefly review the DisoRDPbind method and provide detailed instructions on usage of its web interface. Finally, we discuss the case study of DisoRDPbind applied to an HCV core protein, an intrinsically disordered RNA chaperone.

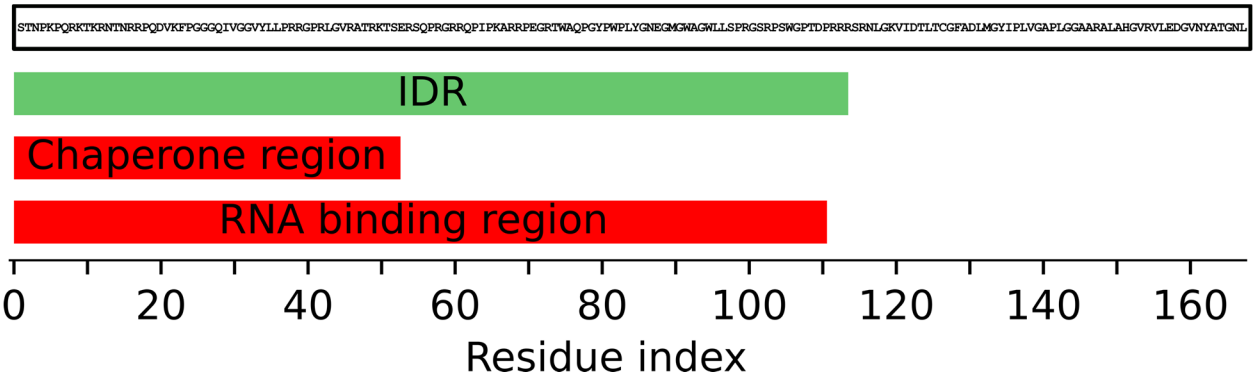

Figure 1. Sequence of HCV core protein, and annotated IDR, RNA chaperone region, and RNA binding region.

2. Materials

DisoRDPbind infers the function of novel sequences through use of historical protein functional data. Available historical data were broken into two sequence dis-similar subsets: the training dataset and the testing dataset. The training dataset is used to build the models DisoRDPbind uses to infer function in novel sequences. The testing data is used to assess the performance of the final model on novel data. Estimation of model accuracy aids in interpreting prediction results for novel proteins; it informs how many correct and incorrect predictions are typically expected for an average protein sequence.

2.1. Datasets

Training and testing datasets for development of DisoRDPbind [56] were extracted from the DisProt database [61], a database of IDPs including function annotations. IDPs from DisProt were clustered at 30% sequence identity and clusters were assigned to either the training or testing set. This procedure is intended to avoid over estimation of method performance by ensuring that orthologous proteins do not appear in both the training and testing sets. It also demonstrates whether DisoRDPbind can make correct predictions in the absence of sequence similarity, i.e., when sequence alignment typically would not produce accurate results. The training dataset included 315 proteins. Two test sets were used, one with 114 proteins, and another of 36 proteins that consisted of only recent additions to the DisProt database. These datasets are available at http://biomine.ece.ualberta.ca/DisoRDPbind/.

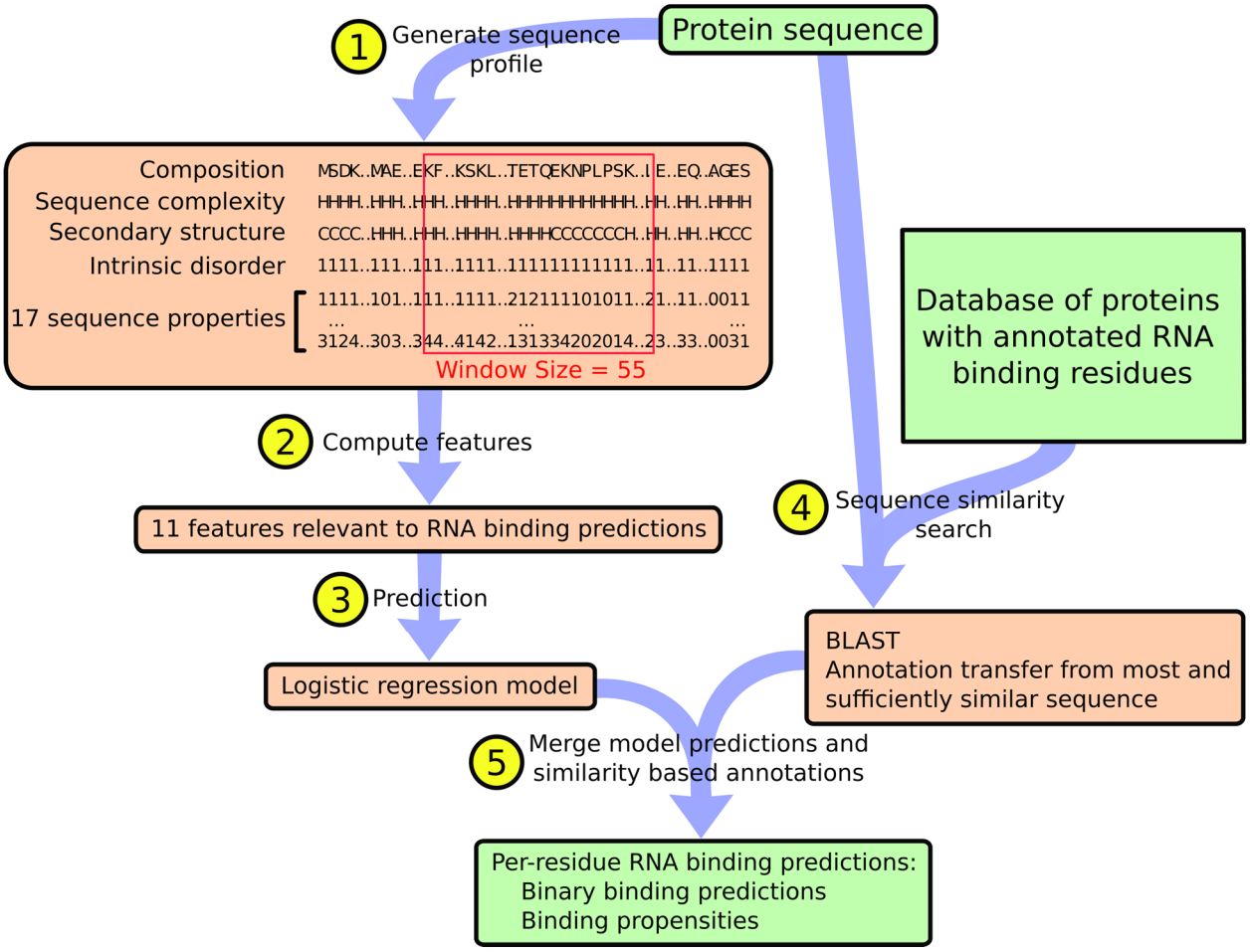

Figure 2. Partial architecture of DisoRDPbind, focused on the RNA binding portion of method. Each stage referenced in the text is indicated with a number in a yellow circle.

2.2. Architecture

The architecture of DisoRDPbind was designed to allow for high-throughput predictions, enabling practical whole proteome IDP function prediction. The architecture has five primary stages (Figure 2), where here we focus on the RNA-binding prediction portion of DisoRDPbind, see Peng *et al* for full

architecture details [56]. Stage 1 develops a profile representation of the protein sequence that covers several relevant sequence properties. Stage 2 extracts a processed set of features that are computed from the profile. These features numerically quantify information that is relevant for prediction, where independent feature sets are selected for each type of functional region. Stage 3 uses these features as input to a trained logistic regression model to produce model-based predictions. Stage 4 is done in parallel to stages 1 through 3, where the input sequence is compared to a database of proteins with annotated RNA binding function. If a sufficiently similar sequence is found in the database, annotations are transferred to the input protein at aligned positions in the sequence. Finally, stage 5 merges model-based predictions with similarity-based predictions to give the final DisoRDPbind prediction.

The feature representation of each sequence (Figure 2, stage 1) involves several calculated features. Intrinsic disorder predictions are made using the IUPred algorithm [5], and low complexity regions – a sequence property correlated with IDRs – are identified using the SEG algorithm [62]. Additional features include residue compositions, secondary structure predictions made with the PSIPRED algorithm [63], and 17 selected amino acid scales from the AAIndex database [64]. Amino acid scales quantify physiochemical properties such as hydrophobicity, net charge, and folding free energy. Sliding windows were applied to average each input features, which transforms residue predictions into local sequence averages, where a window size of 55 residues was used for RNA-binding prediction. Windowed averages are used to make predictions for the center residue of the window. Empirical feature selection (Figure 2, stage 2) was used to remove uninformative and redundant features prior to model training or prediction. A separate set of features was selected empirically for each function prediction method. A small set of 11 features was found to give good results for RNA-binding prediction. Each selected feature set is used in a separate logistic regression model for each function type (Figure 2, stage 3). Logistic regression models are robust to overfitting, are extremely fast, and provide a propensity in the range of 0 to 1 for each residue in a protein. The overall method provides three separate propensity scores, one of which indicates propensity of a residue to be intrinsically disordered and interact with RNA. The final stage of DisoRDPbind merges these predicted propensities with functional annotations found through sequence similarity with the training dataset (Figure 2, stage 4). Input sequences are compared with the training dataset using BLAST [65]. The alignments produced by BLAST are used to transfer functional annotations from training set protein sequences to input protein sequences.

DisoRDPbind output consists of the RNA-, DNA-, and protein-binding propensity scores, as well as binary classification of each residue as RNA-, DNA-, and protein-binding based on a model specific threshold. The thresholds were selected to produce predictions with a low (10%) false positive rate on the training dataset [56]. Residues with propensity scores that exceed the model specific threshold are then classified as either RNA-, DNA-, or protein-binding. Greater propensity scores are indicative of a higher likelihood of binding to a particular molecule type (see Note 1).

2.3. Predictive Quality and Runtime

Prediction quality of DisoRDPbind for RNA-binding residues of IDPs is significantly better than other computational predictors of RNA binding residues that were not specifically designed for IDPs, including BindN+ [41] and RNABindR [46]. The area-under-the-curve (AUC) metric is a threshold agnostic measure of predictive performance. DisoRDPbind RNA-binding predictions produced AUC values around 0.67, depending on the specific test set used, which was significantly better than other methods tested, with AUC values between 0.54 and 0.64 [56]. The other methods tested were developed from structured RNA-binding proteins, whereas DisoRDPbind was developed form intrinsically disordered proteins, which suggests that these predictions may be complementary. A comparison of these methods indicates that this is in fact the case; DisoRDPbind is poorly correlated with other methods of RNA-binding

prediction with a correlation coefficient less than 0.3. This demonstrates that DisoRDPbind is not only accurate, but also complementary to existing RNA-binding prediction methods. This was also recently confirmed in a study of putative RNA-binding protein in the human proteome [37].

The runtime of DisoRDPbind increases quadratically with protein length (see Note 2), ranging between a fraction of a second to several seconds per protein on a modern computer system [56]. This runtime includes predictions of DNA, RNA, and protein interactions. This efficient runtime performance makes proteome scale predictions practical. For example, predictions for the entire human proteome can be obtained in around 40 hours on a modern computer system.

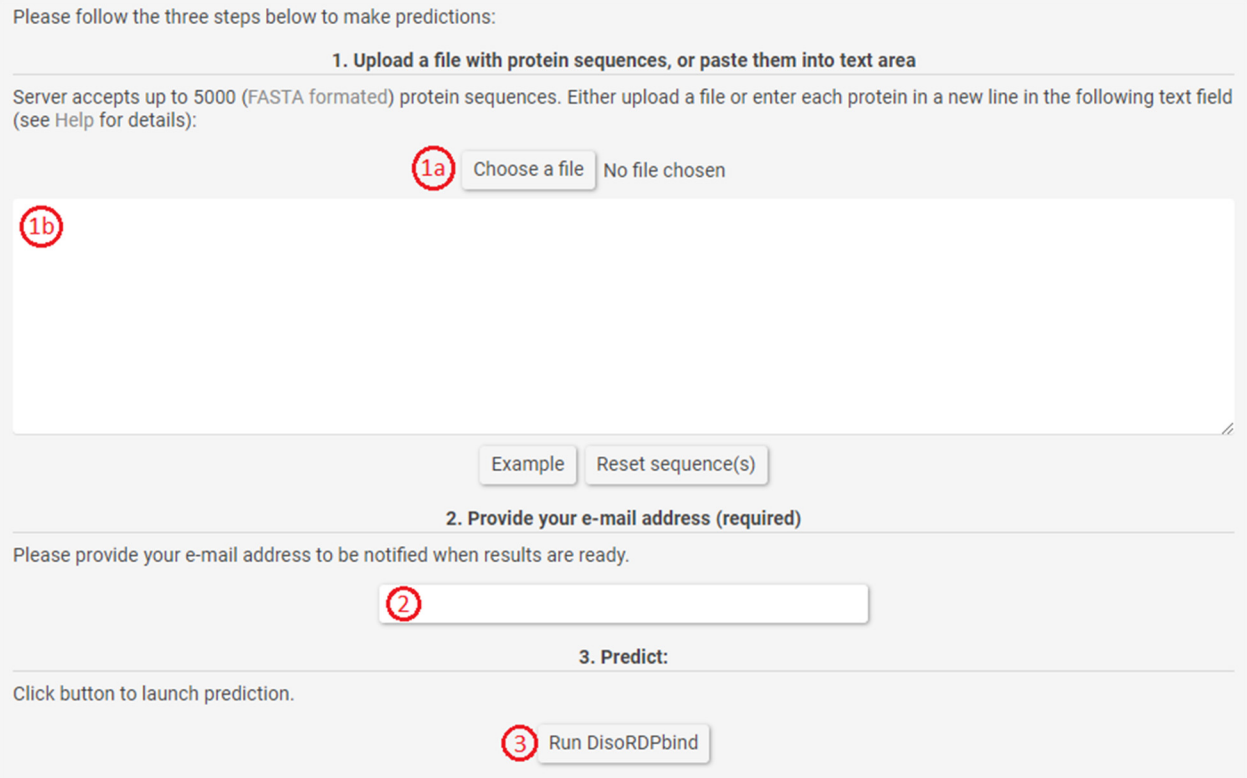

Figure 3. The DisoRDPbind prediction submission form. Red numbers indicate the three necessary steps to submit sequences for predictions, discussed in the text.

2.4. Webserver

The user-friendly web interface for DisoRDPbind can be accessed at http://biomine.ece.ualberta.ca/DisoRDPbind/. The only system requirements for submitting sequences for prediction are: an internet connection and a modern web browser. The interface has been tested with Firefox, Internet Explorer, and Chrome.

Prediction submissions are made at the main page for DisoRDPbind. This page will accept up to 5000 protein sequences in FASTA format, submitted as either a file upload or with a text entry field. Notification of completed predictions are provided by email, so an email address is required for submission of sequences for prediction. Notifications provide a link to prediction results and an explanation of result file format.

Once sequences are submitted, the webserver runs all programs necessary to make DisoRDPbind predictions. Disorder predictions are made with IUPred [5], secondary structure predictions are made with PSIPRED (in single sequence mode) [63], low complexity regions are identified with the SEG method [62], and annotation transfer is made with BLAST [65]. There are no required options for running DisoRDPbind, simply supply sequences for prediction, enter an email address, and click the "Run DisRDPbind" button. The webserver will then run all required programs and send a notification to the supplied email address when predictions are completed.

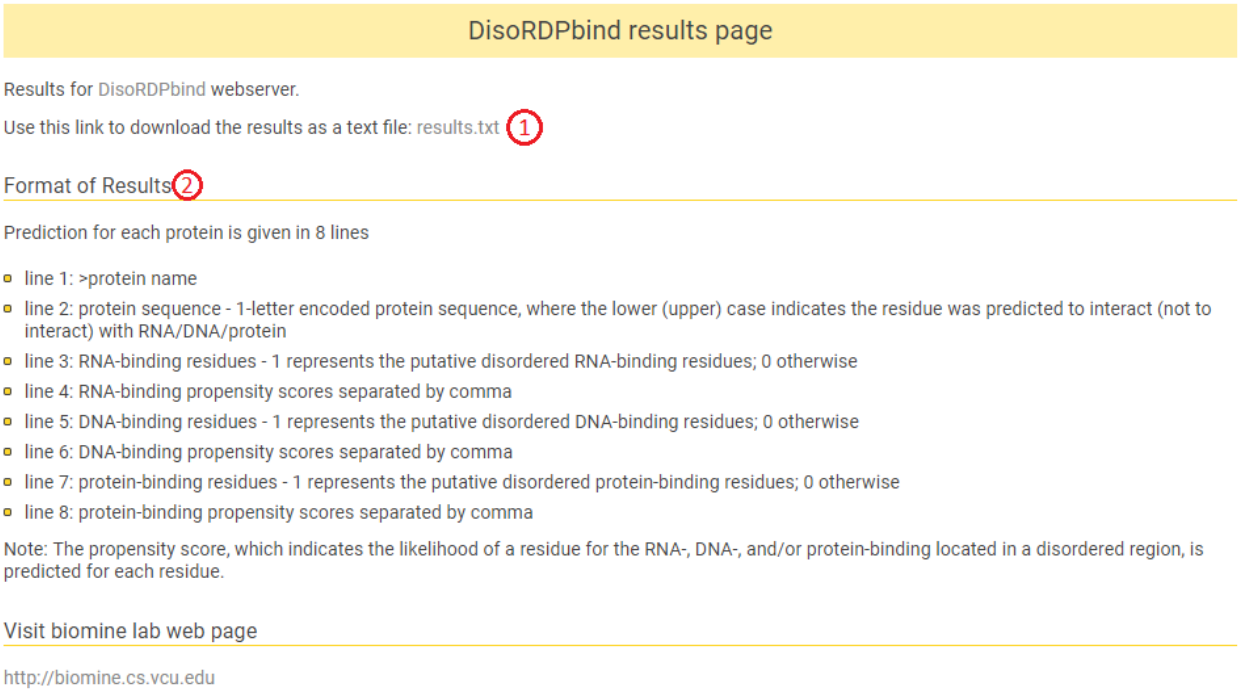

Figure 4. The DisoRDPbind prediction results page. Red numbers indicate important features of this page, discussed in the text.

3. Methods

3.1. Running DisoRDPbind

There are three steps to submit sequences for prediction to the DisoRDPbind server (Figure 3, labels 1a/b, 2, and 3):

1. Provide FASTA formatted sequences (see Note 3) for prediction using 1 a or b, depending on the desired submission method. Clicking the "Reset sequence(s)" button below 1b will clear both submission options. There are limits to both the number of sequences (see Note 4) and maximum length of sequences (see Note 5) submitted for prediction.
    a. Upload a file of FASTA formatted sequences.
    b. Provide FASTA formatted sequences as text. This can be done using the copy and paste function of your operating system; copy from a local file and paste to the text field. For an example of properly formatted sequences, click the "Example" button located below the text field.

2. Provide an email address (see Note 6). This email address is only used to send notification of completed prediction results; you will receive only one notification email per submission.
3. Click "Run DisoRDPbind" to submit sequences and run predictions.

Clicking "Run DisoRDPbind" submit will take the user to a status page, reporting on the current state of the submitted prediction. Submissions are entered into the webservers queue system and the status page will report the current position in the queue and when predictions on the submission have begun. After predictions have completed, the status page will redirect to the prediction results page, and an email will be sent to the notification email address provided. There is no need to keep the status page open while predictions are pending, a notification email is always sent on prediction completion.

3.2. Results Generated by DisoRDPbind

The results page can be reached by leaving a browser open to the status page, or following the link provided in the results email. The results page includes a link to a text file 'results.txt' with prediction results (Figure 4, label 1) and a description of the result file format (Figure 4, label 2). The result file contains RNA-, DNA-, and protein-interaction prediction results for each of the submitted protein sequences. Prediction results include both interaction propensities, ranging from 0 for low propensity and 1 for high propensity, and binary interaction predictions (see Note 7), 0 for non-interacting and 1 for interacting, for each of the three interaction types. Each sequence is represented by eight lines in the results file, where the first 4 are relevant for RNA-interaction prediction:

1. The protein name taken from the FASTA header of each sequence.
2. The protein sequence with interacting residues encoded with character case; lower case residues are predicted to be in intrinsically disordered regions that interact with DNA, RNA, and/or protein and upper case residues are predicted not to be in intrinsically disordered regions or to not interact with DNA, RNA, or protein.
3. RNA-interaction binary predictions, either 1 for interaction or 0 for no interaction.
4. RNA-interaction propensity, ranging between 1 for high interaction propensity to 0 for low interaction propensity.

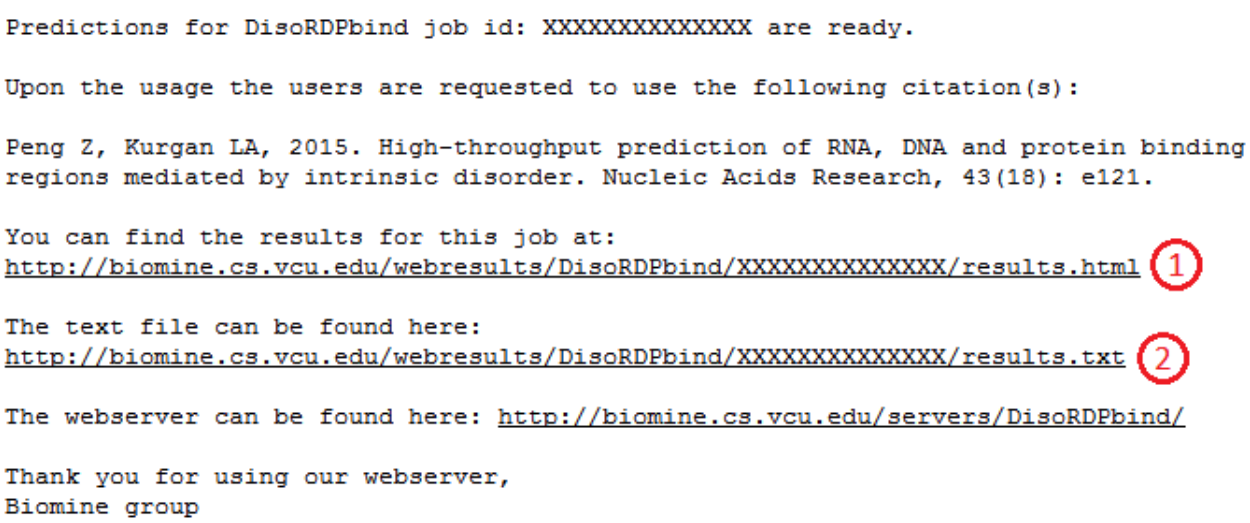

Figure 5. The DisoRDPbind notification email. The email provides links, indicated with red numbers, to prediction results, discussed in the text.

When prediction results have been completed by the server, a notification email (Figure 5) is sent to the email address provided during sequence submission. The email notification contains a link to the results page (Figure 5, label 1) and a direct link to the results text file (Figure 5, label 2). Each job has a unique numerical identifier (Figure 5, "XXXXXXXXXXXXXX") that is given at the top of the notification email and is used in links to identify each submission (see Note 8). In the case of issues with a prediction submission, the prediction identification number is used to trace the corresponding submission.

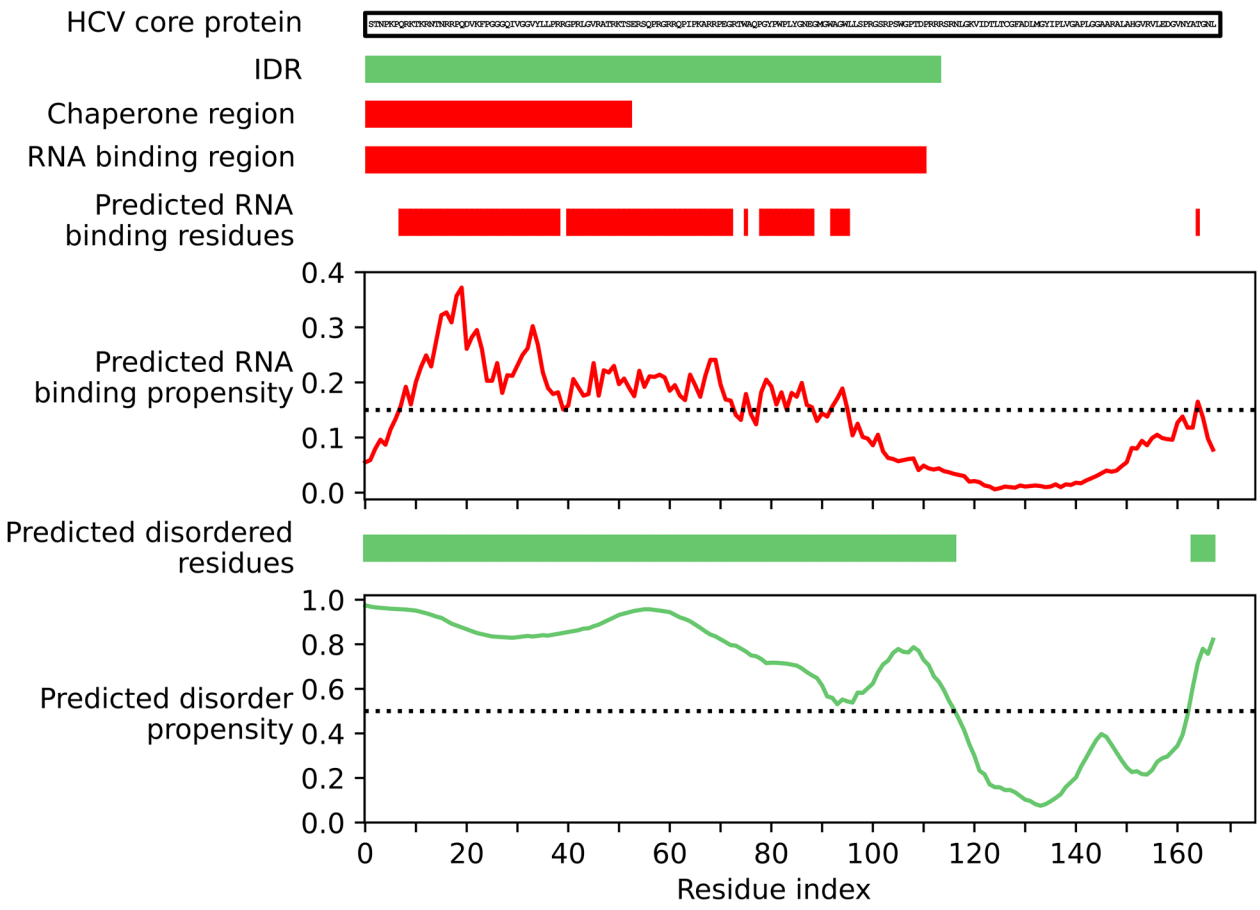

Figure 6. Annotated regions and prediction results for HCV core protein. Annotated regions are an IDR, know RNA binding region, and sufficient RNA chaperone region. DisoRDPbind RNA binding prediction results are shown both as binary (red regions indicate predicted RNA binding residues) and propensity per-residue. The VSL2B prediction of intrinsic disorder is also shown as both a binary prediction (green regions indicate predicted IDRs) and a propensity score, where residues with values greater than 0.5 are predicted to be disordered.

3.3. Case Study

Flaviviridae are a group of single strand RNA, enveloped viruses that infect mammals, including humans. This group includes HCV, which chronically infects up to 130-170 million people world-wide resulting in over 350,000 deaths annually [66]. The core protein of *Flaviviridae* serves as both a capsid protein and as an RNA chaperone [59]. The proteomes of *Flaviviridae* are encoded as a polyprotein, where the core protein is located at or near the extreme N-terminus and is released by proteolysis [58]. The core

protein of HCV is characterized by a basic N-terminus and hydrophobic C-terminus. While the domain organization of *Flaviviridae* may differ, they are generally characterized by the presence of a basic region. These basic regions have been shown to be intrinsically disordered by circular dichroism and to carry RNA binding activity. A shorter N-terminal region has been shown to be sufficient for RNA chaperone activity by base-pairing assays [59]. The intrinsically disordered, RNA binding, and chaperone regions are shown at the top of Figure 6. We note that this figure was created using a specialized graphical software package. DisoRDPbind's users have access to the corresponding text-based output that is summarized in Figure 4.

Application of DisoRDPbind to HCV core protein demonstrates good agreement between the HCV core RNA binding region and predicted RNA binding residues (middle of the Figure 6). RNA binding propensity values for this protein range 0.006 to 0.372, with nearly all of the highest scores located in the known RNA binding region. These propensity values are used to obtain a binary prediction – either RNA binding or non-RNA binding – for each residue in the protein by application of a threshold of 0.151. This threshold was select to balance identification of novel binding residues against spurious predictions (Note 7). DisoRDPbind predicts residues throughout the known RNA binding region to interact with RNA. The predicted residues include nearly all of the region known to have RNA chaperone function. Residues not predicted to be RNA binding by DisoRDPbind are primarily located at the extremes of the defined RNA binding region. This suggests the hypothesis that the necessary and sufficient RNA binding region could be the shorter region suggested by DisoRDPbind, which could be tested experimentally.

Disorder predictions for HCV core protein performed with the VSL2b method [7] also agree well with the characterized disordered region (bottom of Figure 6). Disordered propensity scores are converted to binary disorder predictions – either disorder or structured – for each residue of a protein in the same manner as DisoRDPbind, but with a threshold value of 0.5. While disorder predictions give an accurate estimation of the location of IDRs and structured regions of a protein, they do not carry a direct indication of protein function. For function prediction, specialized prediction methods, like DisoRDPbind, can be used to decompose IDR into functional regions.

4. Notes

1. In the analysis of individual proteins, it may be useful to examine propensity scores in addition to binary predictions. Elevated propensity scores that do not exceed the prediction threshold (and consequently which do not result in the binary prediction of binding) may be indicative of function when combined with other data. The threshold were originally selected to ensure low (10%) false positive rate on the training dataset, resulting in a conservative set of binary predictions of binding. Thus, high propensity scores suggest that the corresponding residues have elevated likelihood for binding, however, the user should expect higher levels of false positives among these predictions.
2. A formula for estimating the run time in milliseconds of DisoRDPbind for a given sequence was determined to be [56]:

$$0.007n^2 + 0.9028n + 301.06$$

   where n is the number of amino acids in the protein. For $n = 200$, the estimate is 0.79 sec, and for $n = 1000$, the estimate is 8.9 secs. Predictions for proteins for each webserver submission are run serially, so applying the above formula to each sequence in the submission and taking the sum will provide a run time estimate.
3. The FASTA format is described at https://en.wikipedia.org/wiki/FASTA_format. Briefly, the format consists of a series of sequence label lines, beginning with ">", followed by the sequence beginning on the next line.

4. Up to 5000 FASTA formatted sequences can be submitted at one time to the web interface. Submission sizes exceeding this limit will result in an error notification from the server and no predictions will be run by the server. For submission of more than 5000 sequences, it will be necessary to break the sequences into multiple submissions each with 5000 or fewer sequences.
5. The programs used to generate predictor inputs limit the maximum length of protein sequences submitted to the webserver. Submitted sequences should be limited to fewer than 10,000 residues.
6. Although single sequence predictions can be made in as little as a fraction of a second, prediction of 5000 sequences will typically require several hours. Rather than requiring an active browser connection, notification of completed predictions are provided via email. The email message will contain instruction on how to access prediction results.
7. Binary predictions are directly related to propensity scores; propensities greater than the predictor-specific threshold are classified as interacting (binary value of 1) and propensities less than the same threshold are classified as non-interacting (binary value of 0). For RNA binding prediction, the threshold is set at a propensity of 0.151. This threshold was selected to give a 10% false positive rate on the training dataset.
8. Please save this email or included links. Predictions will be accessible via these links for at least 3 months after prediction. It is also recommend to save the status page URL, which can be used in the case of a typo in notification email address resulting in no notification email receipt.

Acknowledgements

This research was supported in part by the Robert J. Mattauch Endowment funds and the National Science Foundation grant 1617369 to Lukasz Kurgan.

References

1. van der Lee R, Buljan M, Lang B, Weatheritt RJ, Daughdrill GW, Dunker AK, Fuxreiter M, Gough J, Gsponer J, Jones DT, Kim PM, Kriwacki RW, Oldfield CJ, Pappu RV, Tompa P, Uversky VN, Wright PE, Babu MM (2014) Classification of Intrinsically Disordered Regions and Proteins. Chemical Reviews 114 (13):6589-6631.
2. Dunker AK, Obradovic Z (2001) The protein trinity--linking function and disorder. Nature biotechnology 19 (9):805-806.
3. Wright PE, Dyson HJ (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. J Mol Biol 293 (2):321-331.
4. Uversky VN, Gillespie JR, Fink AL (2000) Why are "natively unfolded" proteins unstructured under physiologic conditions? Proteins 41 (3):415-427.
5. Dosztanyi Z, Csizmok V, Tompa P, Simon I (2005) The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. J Mol Biol 347 (4):827-839.
6. Walsh I, Martin AJ, Di Domenico T, Tosatto SC (2012) ESpritz: accurate and fast prediction of protein disorder. Bioinformatics 28 (4):503-509.
7. Peng K, Radivojac P, Vucetic S, Dunker AK, Obradovic Z (2006) Length-dependent prediction of protein intrinsic disorder. BMC Bioinformatics 7:208.
8. Meng F, Uversky VN, Kurgan L (2017) Comprehensive review of methods for prediction of intrinsic disorder and its molecular functions. Cell Mol Life Sci 74 (17):3069-3090.
9. Lieutaud P, Ferron F, Uversky AV, Kurgan L, Uversky VN, Longhi S (2016) How Disordered is My Protein and What is Its Disorder For? A Guide Through the “Dark Side” of the Protein Universe. Intrinsically Disordered Proteins 4 (1):e1259708.
10. Monastyrskyy B, Kryshtafovych A, Moult J, Tramontano A, Fidelis K (2014) Assessment of protein disorder region predictions in CASP10. Proteins 82 Suppl 2:127-137.

11. Necci M, Piovesan D, Dosztanyi Z, Tompa P, Tosatto SCE (2017) A comprehensive assessment of long intrinsic protein disorder from the DisProt database. Bioinformatics.
12. Fan X, Kurgan L (2014) Accurate prediction of disorder in protein chains with a comprehensive and empirically designed consensus. J Biomol Struct Dyn 32 (3):448-464.
13. Meng F, Uversky V, Kurgan L (2017) Computational Prediction of Intrinsic Disorder in Proteins. Curr Protoc Protein Sci 88:2 16 11-12 16 14.
14. Mizianty MJ, Stach W, Chen K, Kedarisetti KD, Disfani FM, Kurgan L (2010) Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources. Bioinformatics 26 (18):i489-496.
15. Jones DT, Cozzetto D (2015) DISOPRED3: precise disordered region predictions with annotated protein-binding activity. Bioinformatics 31 (6):857-863.
16. Peng Z, Mizianty MJ, Kurgan L (2014) Genome-scale prediction of proteins with long intrinsically disordered regions. Proteins 82 (1):145-158.
17. Xue B, Dunker AK, Uversky VN (2012) Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life. J Biomol Struct Dyn 30 (2):137-149.
18. Pancsa R, Tompa P (2012) Structural Disorder in Eukaryotes. PLoS ONE 7 (4):e34687.
19. Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J Mol Biol 337 (3):635-645.
20. Tompa P (2012) Intrinsically disordered proteins: a 10-year recap. Trends in biochemical sciences 37 (12):509-516.
21. Peng Z, Yan J, Fan X, Mizianty MJ, Xue B, Wang K, Hu G, Uversky VN, Kurgan L (2015) Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life. Cell Mol Life Sci 72 (1):137-151.
22. Hu G, Wang K, Song J, Uversky VN, Kurgan L (2018) Taxonomic Landscape of the Dark Proteomes: Whole-Proteome Scale Interplay Between Structural Darkness, Intrinsic Disorder, and Crystallization Propensity. Proteomics:e1800243.
23. Yan J, Mizianty MJ, Filipow PL, Uversky VN, Kurgan L (2013) RAPID: fast and accurate sequence-based prediction of intrinsic disorder content on proteomic scale. Biochim Biophys Acta 1834 (8):1671-1680.
24. Dyson HJ, Wright PE (2005) Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol 6 (3):197-208.
25. Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradovic Z (2002) Intrinsic disorder and protein function. Biochemistry 41 (21):6573-6582.
26. Xie H, Vucetic S, Iakoucheva LM, Oldfield CJ, Dunker AK, Uversky VN, Obradovic Z (2007) Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions. Journal of proteome research 6 (5):1882-1898.
27. Chen JW, Romero P, Uversky VN, Dunker AK (2006) Conservation of intrinsic disorder in protein domains and families: II. Functions of conserved disorder. Journal of Proteome Research 5 (4):888-898.
28. Cumberworth A, Lamour G, Babu MM, Gsponer J (2013) Promiscuity as a functional trait: intrinsically disordered regions as central players of interactomes. Biochemical Journal 454:361-369.
29. Dyson HJ (2012) Roles of intrinsic disorder in protein-nucleic acid interactions. Molecular Biosystems 8 (1):97-104.
30. Fuxreiter M, Toth-Petroczy A, Kraut DA, Matouschek AT, Lim RYH, Xue B, Kurgan L, Uversky VN (2014) Disordered Proteinaceous Machines. Chem Rev 114 (13):6806-6843.
31. Haynes C, Oldfield CJ, Ji F, Klitgord N, Cusick ME, Radivojac P, Uversky VN, Vidal M, Iakoucheva LM (2006) Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. PLoS Comput Biol 2 (8):890-901.

32. Peng Z, Oldfield CJ, Xue B, Mizianty MJ, Dunker AK, Kurgan L, Uversky VN (2014) A creature with a hundred waggly tails: intrinsically disordered proteins in the ribosome. Cell Mol Life Sci 71 (8):1477-1504.
33. Peng Z, Mizianty MJ, Xue B, Kurgan L, Uversky VN (2012) More than just tails: intrinsic disorder in histone proteins. Mol Biosyst 8 (7):1886-1901.
34. Tompa P, Csermely P (2004) The role of structural disorder in the function of RNA and protein chaperones. Faseb Journal 18 (11):1169-1175.
35. Wu Z, Hu G, Yang J, Peng Z, Uversky VN, Kurgan L (2015) In various protein complexes, disordered protomers have large per-residue surface areas and area of protein-, DNA- and RNA-binding interfaces. FEBS Lett 589 (19 Pt A):2561-2569.
36. Wang C, Uversky VN, Kurgan L (2016) Disordered nucleiome: Abundance of intrinsic disorder in the DNA- and RNA-binding proteins in 1121 species from Eukaryota, Bacteria and Archaea. Proteomics 16 (10):1486-1498.
37. Chowdhury S, Zhang J, Kurgan L (2018) In Silico Prediction and Validation of Novel RNA Binding Proteins and Residues in the Human Proteome. Proteomics:e1800064.
38. TOMPA P, CSERMELY P (2004) The role of structural disorder in the function of RNA and protein chaperones. The FASEB Journal 18 (11):1169-1175.
39. Ivanyi-Nagy R, Davidovic L, Khandjian EW, Darlix J-L (2005) Disordered RNA chaperone proteins: from functions to disease. Cellular and Molecular Life Sciences CMLS 62 (13):1409-1417.
40. Liu ZP, Wu LY, Wang Y, Zhang XS, Chen LN (2010) Prediction of protein-RNA binding sites by a random forest method with combined features. Bioinformatics 26 (13):1616-1622.
41. Wang L, Huang C, Yang MQ, Yang JY (2010) BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features. BMC Systems Biology 4 (1):S3.
42. Walia RR, Xue LC, Wilkins K, El-Manzalawy Y, Dobbs D, Honavar V (2014) RNABindRPlus: a predictor that combines machine learning and sequence homology-based methods to improve the reliability of predicted RNA-binding residues in proteins. PloS one 9 (5):e97725.
43. Wang L, Brown SJ (2006) BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. Nucleic acids research 34 (Web Server issue):W243-248.
44. Kumar M, Gromiha MM, Raghava GP (2008) Prediction of RNA binding sites in a protein using SVM and PSSM profile. Proteins 71 (1):189-194.
45. Yang X, Wang J, Sun J, Liu R (2015) SNBRFinder: A Sequence-Based Hybrid Algorithm for Enhanced Prediction of Nucleic Acid-Binding Residues. PloS one 10 (7):e0133260.
46. Walia RR, Caragea C, Lewis BA, Towfic F, Terribilini M, El-Manzalawy Y, Dobbs D, Honavar V (2012) Protein-RNA interface residue prediction using machine learning: an assessment of the state of the art. BMC bioinformatics 13:89.
47. Yan J, Kurgan L (2017) DRNApred, fast sequence-based method that accurately predicts and discriminates DNA- and RNA-binding residues. Nucleic Acids Res 45 (10):e84.
48. Yan J, Friedrich S, Kurgan L (2016) A comprehensive comparative review of sequence-based predictors of DNA- and RNA-binding residues. Brief Bioinform 17 (1):88-105.
49. Meszaros B, Simon I, Dosztanyi Z (2009) Prediction of protein binding regions in disordered proteins. PLoS Comput Biol 5 (5):e1000376.
50. Khan W, Duffy F, Pollastri G, Shields DC, Mooney C (2013) Predicting binding within disordered protein regions to structurally characterised peptide-binding domains. PloS one 8 (9):e72838.
51. Disfani FM, Hsu WL, Mizianty MJ, Oldfield CJ, Xue B, Dunker AK, Uversky VN, Kurgan L (2012) MoRFpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins. Bioinformatics 28 (12):i75-83.
52. Meng F, Kurgan L (2018) High-throughput prediction of disordered moonlighting regions in protein sequences. Proteins.

53. Meng F, Kurgan L (2016) DFLpred: High-throughput prediction of disordered flexible linker regions in protein sequences. Bioinformatics 32 (12):i341-i350.
54. Oldfield CJ, Uversky VN, Kurgan L (2018) Predicting Functions of Disordered Proteins with MoRFpred. Methods Mol Biol 1851.
55. Yan J, Dunker AK, Uversky VN, Kurgan L (2016) Molecular recognition features (MoRFs) in three domains of life. Mol Biosyst 12 (3):697-710.
56. Peng Z, Kurgan L (2015) High-throughput prediction of RNA, DNA and protein binding regions mediated by intrinsic disorder. Nucleic Acids Res 43 (18):e121.
57. Peng Z, Wang C, Uversky VN, Kurgan L (2017) Prediction of Disordered RNA, DNA, and Protein Binding Regions Using DisoRDPbind. Methods Mol Biol 1484:187-203.
58. Gawlik K, Gallay PA (2014) HCV core protein and virus assembly: what we know without structures. Immunologic research 60 (1):1-10.
59. Ivanyi-Nagy R, Lavergne J-P, Gabus C, Ficheux D, Darlix J-L (2008) RNA chaperoning and intrinsic disorder in the core proteins of Flaviviridae. Nucleic acids research 36 (3):712-725.
60. Sharma K, Didier P, Darlix JL, de Rocquigny H, Bensikaddour H, Lavergne JP, Penin F, Lessinger JM, Mely Y (2010) Kinetic analysis of the nucleic acid chaperone activity of the hepatitis C virus core protein. Nucleic acids research 38 (11):3632-3642.
61. Piovesan D, Tabaro F, Mičetić I, Necci M, Quaglia F, Oldfield CJ, Aspromonte MC, Davey NE, Davidović R, Dosztányi Z, Elofsson A, Gasparini A, Hatos A, Kajava AV, Kalmar L, Leonardi E, Lazar T, Macedo-Ribeiro S, Macossay-Castillo M, Meszaros A, Minervini G, Murvai N, Pujols J, Roche DB, Salladini E, Schad E, Schramm A, Szabo B, Tantos A, Tonello F, Tsirigos KD, Veljković N, Ventura S, Vranken W, Warholm P, Uversky VN, Dunker A K, Longhi S, Tompa P, Tosatto SCE (2017) DisProt 7.0: a major update of the database of disordered proteins. Nucleic acids research 45 (Database issue):D219-D227.
62. Wootton JC, Federhen S (1993) Statistics of Local Complexity in Amino-Acid-Sequences and Sequence Databases. Comput Chem 17 (2):149-163.
63. McGuffin LJ, Bryson K, Jones DT (2000) The PSIPRED protein structure prediction server. Bioinformatics 16 (4):404-405.
64. Kawashima S, Ogata H, Kanehisa M (1999) AAindex: Amino Acid Index Database. Nucleic acids research 27 (1):368-369.
65. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25 (17):3389-3402.
66. World Health Assembly (2010) Viral hepatitis: report by the Secretariat. vol A63/15. World Health Organization, Geneva.