Journal of Biomedical Informatics 60 (2016) 120–131

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Software tools for simultaneous data visualization and T cell epitopes and disorder prediction in proteins

Davorka R. Jandrlić [a],*, Goran M. Lazić [b], Nenad S. Mitić [b], Mirjana D. Pavlović [c]

[a] *University of Belgrade, Faculty of Mechanical Engineering, Kraljice Marije 16, Belgrade, Serbia*
[b] *University of Belgrade, Faculty of Mathematics, P.O.B. 550, Studentski trg 16/IV, Belgrade, Serbia*
[c] *University of Belgrade, Institute of General and Physical Chemistry, Studentski trg 12/V, Belgrade, Serbia*

ARTICLE INFO

Article history:
Received 19 July 2015
Revised 15 January 2016
Accepted 28 January 2016
Available online 3 February 2016

Keywords:
Disorder prediction
T cell epitope prediction
Visualization
Massive parallel prediction

ABSTRACT

We have developed EpDis and MassPred, extendable open source software tools that support bioinformatic research and enable parallel use of different methods for the prediction of T cell epitopes, disorder and disordered binding regions and hydropathy calculation. These tools offer a semi-automated installation of chosen sets of external predictors and an interface allowing for easy application of the prediction methods, which can be applied either to individual proteins or to datasets of a large number of proteins. In addition to access to prediction methods, the tools also provide visualization of the obtained results, calculation of consensus from results of different methods, as well as import of experimental data and their comparison with results obtained with different predictors. The tools also offer a graphical user interface and the possibility to store data and the results obtained using all of the integrated methods in the relational database or flat file for further analysis. The MassPred part enables a massive parallel application of all integrated predictors to the set of proteins. Both tools can be downloaded from http://bioinfo.matf.bg.ac.rs/home/downloads.wafl?cat=Software. Appendix A includes the technical description of the created tools and a list of supported predictors.

1. Introduction

Bioinformatics research often includes large-scale analyses of different macromolecular attributes which provide for an increased understanding of complex cellular processes. The aim of these *in silico* approaches is to reduce research time and cost by developing predictive methods to assist in improving experimental approaches in biological research. Predicting protein structure and function are one of the central themes of proteomics. It has been suggested that the major factors of epitope immunodominance in the CD4+ T cell response are either the site within the 3D structure of the protein antigen where the peptide antigenic determinant (epitope) is located, or the amino acid (AA) composition of the epitope (reviewed in [65]. There is also a strong correlation between T-cell epitopes and structured (ordered) regions, suggesting that nearby regions of structural instability (disorder) define the ends of T-cell epitopes in experimental model antigens [27,38]. Cytotoxic T-lymphocyte (CTL) epitopes were found to be highly concentrated in the hydrophobic [33] and α-helical [63] regions of HIV-1 proteins. In addition, selfpeptides with a high frequency of hydrophobic residues, presented by the major histocompatibility complex (MHC) class I molecules, were found to be concentrated in highly conserved regions of human proteins [19]. The observed positional biases of T-cell epitopes could originate from cellular mechanisms involved in epitope processing, transport and presentation [19,55], as well as unequal distribution of bulky hydrophobic, polar and charged AAs in structured regions as compared to unstructured protein regions [54]. Several attempts aimed at predicting vaccine antigens have been based on the physical properties and composition of proteins, such as isoelectric point, molecular weight, hydrophobicity, flexibility, mutability and bulkiness [37]. Such attempts have also combined different strategies for identifying T-cell epitopes, as the physical properties of AAs and the chemical and molecular properties of peptides [20]. Dynamic conformational transitions in flexible proteins or protein assemblages containing long intrinsically disordered protein regions (IDPRs) are factors that potentially influence protein degradation [58]; however, their relationship to epitope processing remains unknown. Disorder predictions can help to improve the recognition of a binding partner of certain proteins. Specialized methods, [9,39], have been developed to identify protein regions

* Corresponding author.
E-mail addresses: djandrlic@mas.bg.ac.rs (D.R. Jandrlić), chupcko@alas.matf.bg.ac.rs (G.M. Lazić), nenad@matf.bg.ac.rs (N.S. Mitić), mpavlovic@iofh.bg.ac.rs (M.D. Pavlović).

http://dx.doi.org/10.1016/j.jbi.2016.01.016
1532-0464/© 2016 Elsevier Inc. All rights reserved.

that are unstructured in the unbound state and adopt a stable conformation in their bound state, referred to as molecular recognition features (elements) (MORF, MORFs) or disordered binding regions, respectively. Overlap between these regions and short eukaryotic linear motifs (3–10 AA residues) revealed the correspondence to functionally relevant interaction sites which have been shown to play vital roles in eukaryotic regulation and signaling [41,5].

It is, therefore, of practical value for the research in disorder and epitope prediction to have at one's disposal different programs for the prediction and comparative presentation of results. In this paper, we describe two software tools, EpDis (Epitope in Disorder) and MassPred (Massive Prediction) that have been developed to support research examining the correlation between epitope positioning and protein disorder. In this chapter we will present a short overview of categories of computational prediction methods supported by EpDis-MassPred tools, and our motivation for constructing such tools; chapter 2 includes a description of the EpDis-MassPred architecture; chapter 3 provides an example of its application; chapter 4 presents a short overview of existing tools and chapter 5 presents discussion and conclusion.

1.1. Categories of computational predictions supported by EpDis-MassPred tools

1.1.1. Protein disorder prediction

The identification of intrinsically unstructured (disordered) proteins (IDPs) among molecules that fail to crystallize has challenged the classical protein structure-to-function paradigm according to which protein function depends on a well-defined three-dimensional structure. Despite the lack of defined tertiary and/or secondary structure under *in vitro* physiological conditions, proteins that are fully or partially disordered play crucial functional roles, acquiring a defined structure only when bound to other molecules [61]. Due to their conformational plasticity and flexibility, disordered protein regions (IDRs) are engaged in high specificity/low affinity interactions with multiple unrelated partners by utilizing different folding-induced secondary structural elements. The significance of disordered proteins as major regulators of cell functions is reflected in their involvement in the pathogenesis of human diseases, such as cancer, cardiovascular diseases, amyloidosis and neurodegenerative diseases [60]. The identification of disordered regions from protein primary sequences reveals protein domains that can form crystal structures and facilitates functional annotation and further analysis of proteins. To this end, many approaches to disorder prediction based on different concepts, have been developed.

Propensity-based disorder predictors rely on the physicochemical properties of amino acids, or on the concept of the physical background of disorder, i.e. the CH plot, FoldIndex, PreLink, and GlobPlot. Machine learning-based disorder predictors are founded on training on data sets of disordered regions, characterized by missing densities in protein X-ray datasets from PDB, or the flexible regions obtained from nuclear magnetic resonance (NMR) studies. PONDR VSL2B, DisEMBL, DISOPRED2, OnD-CRF and RONN, which are included in the EpDis and MassPred tools, belong to this category of predictors. Methods that rely on the propensity of protein regions to fold or unfold are based on the potential of amino acids to establish or avoid contact with each other. Our software includes the IUPred predictor which evaluates the energy resulting from interactions between amino acid residues based on the inter-residue contacts in globular proteins [15]. Predictors can also combine different approaches. For example, the IsUnstruct method, also included in our software, has adapted the Ising model from statistical mechanics. The method applies dynamic programming to the Ising model, using data from PDB files of proteins sorted in the Disordered Residues Database, [32]. A more accurate sequence-based classification of IDRs is a major challenge for linking IDRs to their biological roles from the molecular to the systems level.

1.1.2. Prediction of disordered binding regions

IDRs could act as flexible linkers or could have binding activity through coupled folding and binding of short binding regions (5–25 residues) located in longer IDRs, named (MoRFs [50] or disordered protein-binding regions (disordered binding regions) [39] that function via binding to other macromolecules adopting a rigid conformation upon binding. According to their structures in the bound state, at least three types of MoRFs can be defined: α-MoRFs, β-MoRFs, and ι-MoRFs, which form α-helices, β-strands, and structures without a regular pattern of the protein backbone. Experimental methods for identifying MoRFs are costly and time consuming, which makes computational methods useful for guiding experimental analysis. Several tools for predicting MoRFs, are available for download or online use: α-MoRF-PredI [50], α-MoRF-PredII [4], ANCHOR [39,12], MoRFpred [9], MFSPSSMpred [13] and DISOPRED3 [22]. α-MoRF-PredII is a neural network-based predictor which is limited to prediction of α-MoRFs. MoRFpred combines annotation transfers by similarity to the output of a support vector machine (SVM) model, that examines sequence conservation data, AA physicochemical properties and predictions of intrinsic disorder, relative solvent accessibility and residue flexibility. MFSPSSMpred considers only sequences, which are preprocessed to enhance the signal of local conservation within the IDRsequences, and is based on position specific scoring matrices. DISOPRED3 is a program for a precise disordered region prediction and annotation of protein-binding sites within disordered regions. The first step is identification of disordered residues through a consensus of the output generated by DISOPRED2 and two additional machine-learning classifiers trained on large IDRs, and the second annotates them as protein binders or non binders through an additional SVM classifier trained on experimental data. The ANCHOR method for the prediction of disordered binding regions, included in the EpDis and MassPred tools, is focused on the prediction of MoRFs which bind to globular proteins [39]. It follows the same idea that underlies the IUPred disorder predictor [11], i.e., the unfavorable intrachain interaction energies of disordered binding segments, combined with the high energetic gain as a result of interaction with a globular protein partner. The ANCHOR statistical model was learnt from a small set of chains, found to be unstructured in isolation, but structured in complex with their partners. ANCHOR attained the highest level of sensitivity in the benchmark, performed by Jones and Cozzetto. The other programs in the benchmark were MoRFpred, MFSPSSMpred and DISOPRED3 [22]. The overlap of disordered protein-binding region prediction with predictions of linear interaction motifs (LM) or linear peptide docking/binding sites help to improve recognition of protein partners, which interact via disorder-to-order transition sites [41,12], or potential phosphorylation sites [18]. Correspondence of epitope and disordered binding region predictions with LM and experimentally verified epitopes has been found in certain autoimmune and tumor-proteins, preferentially associated with flanking regions of disordered binding regions and LM [52]. The importance of flanking regions of linear motifs is further supported by the observation that they are often structurally conserved [5].

1.1.3. T cell epitope prediction

Cellular immunity, mediated by T-lymphocytes, is a central mechanism of the adaptive immune response. T-cell epitopes are peptides that have been released from antigenic molecules through proteolytic mechanisms, and subsequently transported to the cell surface, bound to chaperone-like receptors known as major histocompatibility complex (MHC) molecules, which present epitopes to T cells. Computer models of T-cell epitope prediction, based on MHC-peptide binding, a fundamental step in epitope selection, have been developed to complement expensive and time-consuming experimentation. The large-scale computational screening for T-cell epitopes, which are crucial components of modern vaccines and immunotherapeutics, is the standard procedure employed in the discovery of novel T-cell epitopes in infectious diseases, cancer, autoimmunity and allergy [2].

T-cell epitope prediction methods can be broadly divided into two categories: protein sequence-based and protein structure-based [3,14,56]. Protein sequence-based methods use patterns in peptide sequences with known binding affinities to certain MHC types, together with a variety of fitting techniques [34,2,59,29,30,18]. The structure-based methods use structural information on epitope interactions with MHC molecules, such as peptide-MHC docking of all possible core segments of the peptide into the MHC protein, in order to predict the structures of the bound complexes, and allow for machine-learning-based scoring to predict the peptide binding affinities (reviewed in [3]). In contrast with sequence-based methods, comparatively little work has been undertaken to explore the structure-based methods for predicting peptide binding affinities to MHC molecules. Structure-based methods can currently be applied to allotypes that significantly differ from allotypes which have been identified in experimental peptide-MHC binding data, but have so far not been able to gain an accuracy that is comparable to the top-ranked sequence-based methods (reviewed in [35]). However, certain sequence-based methods use some structural information, such as peptide-contacting polymorphic MHC residues (MHC “pseudo-sequences”), used in NetMHCpan and NetMHCIIpan [45–47,16,1], or MHC pockets in TEPITOPE [57] predictors.

Two large-scale benchmark studies performed by Zhang et al. [64] and Karosiene et al. [23] have demonstrated that NetMHCpan and NetMHC, respectively, are among the best publicly available T cell epitope predictors. Furthermore, NetMHCII and NetMHCIIpan have been identified as the best predictors of single T-cell epitopes within MHC class II epitope predictors [8]. The advantage of pan-specific predictors, which makes them potentially suitable for epitope-based vaccine design, is that they cover alleles for which experimental data are unknown, and because they are capable of dealing with HLA class polymorphisms. Given that peptide vaccines incorporating a 'promiscuous' T-cell epitope provide broad spectrum immunogenicity, the pan-specific predictors are the best choice since they cover the largest number of alleles among all existing predictors.

A number of web-accessible integrated methods for T-cell epitope presentation have been developed, as for example methods that combine MHC class I binding prediction with proteasomal cleavage data and results of TAP transport, but a large-scale benchmark by independent groups has not yet been performed for these methods, [35].

However, the development of a tool for the systematic comparison of different methods is necessary. Standardized data representation [14] or large and diverse training and blind datasets [36] are needed to provide reliable performance benchmarks for epitope-MHC binding predictors. Selection of epitopes predicted by different methods increases the probability that the identified peptide is a true epitope.

1.1.4. Hydropathy calculation

Separating the intervals of hydropathy (hydrophobicity/hydro philicity) and plotting these intervals along a protein sequence facilitates the identification of putative structural features, such as membrane-spanning regions, antigenic sites, exposed loops or buried residues. For the computation of hydropathy values of protein regions we have implemented standard hydropathy plots of proteins with some improvements regarding data storage and manipulation. Hydropathy is calculated by scanning with a sliding window, which can be varied according to the expected size of the structural motif under investigation. At each position in the protein, the sum of hydropathy values (indexes) of all the AAs in the window, is divided by the number of AAs in the window (average hydrophilic or hydrophobic index). Implemented methods for hydropathy index calculation, in EpDis-MassPred system, are based the two hydrophobicity AAs scales, commonly used in bioinformatic research, the Kyte–Doolittle (KD) [26] and Hopp–Woods (HW) [17].

Introducing those methods in EpDis/MassPred tools enables massive complex analysis of a large number of proteins, segregation of protein regions, visualization, overlapping with other chosen motifs and keeping as intervals in a database, or result storage in a format suitable for further analysis.

1.2. Motivation

In some fields of scientific enquiry, it is necessary to combine results from different areas of research. For example, predictions obtained by the use of different parameters and methods can greatly improve the accuracy of predictions of the epitope, disorder and disordered binding regions. Some disorder predictors, trained on a data set of short segments of disordered regions, perform better on shorter than on long disordered regions. To improve the reliability of predictions and avoid overfitting of results, predictors based on different principles should be combined. Therein lay many problems for researchers: the use of various predictors is difficult due to inconsistent representation of results. This renders them unsuitable for further analysis or comparison, even in the case of a single protein.

Alternatively, one could use publicly available servers that allow multiple predictions. Web-based predictions are easy to use on a small scale, while large-scale predictions and a direct comparison of different methods are difficult. In addition, some predictors are time consuming. Furthermore, if access to a predictor is possible only through web applications, it is problematic for application on a large number of proteins. On the other hand, even predictors that are available as stand-alone applications, are not always suitable for application on several hundred or more proteins. Moreover, they frequently require certain programming skills for their installation, that most researchers do not have. In research that includes the previously described characteristics of proteins, it is common practice to combine the results related to two or more required characteristics. Additionally, in order to obtain quality data (e.g. through data mining), it is necessary to apply a selected method to a large number of candidate proteins. The inclusion of a large number of proteins slows down result acquisition, complicates result visualization and analysis, necessitating a capability to rapidly access all generated results for the purpose of switching between proteins that are under analysis.

To overcome the above-mentioned obstacles and decrease research time, we have developed two tools: EpDis (Epitope in Disorder) and MassPred (Massive Prediction). Our tools enable easy access and use of disorder, disordered binding, and T-cell epitopepredictors, and methods for hydropathy calculation, input data processing, uniform display and storage of results, and result preparation for further processing and analysis. The tools do not favor one predictor over another. They allow for the simultaneous use of a number of predictors.

The EpDis tool also provides a visual presentation of the results, obtained either from a single method or as a comparative display of results obtained by any of the supported methods. The parallel display of different characteristics can help in determining correlations between characteristics, as for example the relation of T-cell epitopes and ordered or disordered regions of proteins, the extent of hydropathy of epitopes and certain regions in a protein, etc. It is possible to include in the display experimental data about secondary structures and T-cell epitopes. As a support for further analysis, the developed system enables storage/retrieval of values obtained from (predictors') predictions in the relational database or files.

2. System architecture

The developed system consists of two components, EpDis and MassPred, which can be used either independently or as integrated subsystems offering a higher level of functioning. Both components have been developed for the Linux operating system. The complete system provides, among other things, semi-automated installation of a selected set of predictors, the possibility to apply selected predictors to an individual protein or to an arbitrary set of proteins (a mass application), storing of the obtained data in a variety of formats suitable for further processing, and a visual representation of the results. The system components can be accessed via GUI or through the command interface. The organization of the system is shown in Fig. 1.

Both components of the system can be relatively easily expanded to support additional predictors or user-defined functions. The list of predictors included in the current version (V4.0) is presented in Appendix A and html documentation. All predictors can be downloaded from the internet, while methods for calculating the hydropathy index have been written by the authors of this paper. The first criterion for predictor selection was its availability as a stand-alone application. The selected disorder predictors were mainly those that performed well in CASP experiments (Critical Assessment of Protein Structure Prediction experiments). The epitope predictor system supports various versions of NetMHC/NetMHCPan and NetMHCII/NetMHCIIPan, mainly because they are freely available and have a proven accuracy. Both Masspred and EpDis components of the system support the same set of predictors, but differ in their functions and objectives for which they were designed.

2.1. MassPred

MassPred is a set of tools that provides easy predictor installation, application of predictors to input data and filtering of the results of predictor action. In the preparatory step for predictor installation, the user must manually download the selected predictors from the internet. After this point, MassPred provides scripts for automated installation of the desired predictors and their preparation for automatic execution. Predictors (any of the previously mentioned four types) are applied to the protein dataset, which can be stored in one or more files or directories. Each file can include one or more proteins in fasta format. MassPred takes the contents of input files or directories, extracts every single protein and applies the desired predictors to the extracted proteins, creating separate jobs for every pair (protein, predictor). The created jobs can be simultaneously executed on a symmetric multiprocessor computer, or on computers with a multicore/multithread processor architecture. MassPred itself does not perform a parallel execution of a single predictor application to a single protein. After finishing the generated jobs, MassPred collects the results and filters them in a TSV file format in order to prepare the results in a form that can be used as an input in a *load* utility program for loading results in RDBMS tables. By default, the results are filtered for IBM DB2 RDBMS. MassPred can optionally produce files with a set of SQL INSERT statements for loading data to (DB2 or some other) RDBMS. Additional filters are based on the inclusion of only specific regions (ordered/disordered) and epitope (strong or weak binding) types in the output, or the omission of binding affinity, disorder probability and/or hydropathy level of each amino acid in the processed protein. The MassPred mechanism for generating jobs for simultaneous execution is not restricted to supported predictors. In fact, any program that takes a single protein sequence in fasta format can be massively applied to set of proteins.

MassPred is a command line (shell) oriented system, without a GUI interface. It also allows remote execution.

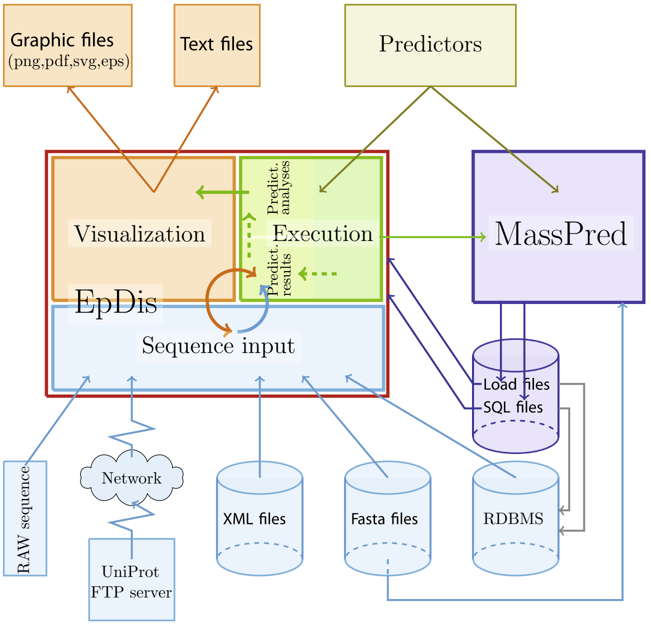


Fig. 1. Global architecture of the EpDis-MassPred system.

2.2. EpDis

EpDis is an open source software tool, published under the MIT license, and developed to facilitate research for biologists and immunologists. It offers simultaneous access to different prediction methods, and combines and compares results obtained by their application to the same input data. In addition to its ability to access predictors, the tool offers the possibility of including experimental data. Experimental data can be compared with predicted results, drawn and displayed. Their overlap with the results of prediction is presented in the form of intervals.

EpDis is composed of components (as shown in Fig. 1) that provide:

(a) an interface for data entry. Input data can be proteins or experimental data (secondary structure, T cell epitopes, and MHC binders). Entering the protein sequences can be performed in several ways: from the fasta file containing single or multiple proteins; from the relational database table where each protein is in AA format along with its unique identifier; by specifying a raw AA sequence, and after acquiring a protein sequence with the specified identifier from the UniProt database through a web interface. Entering experimental data is based on the addition of an appropriate XML file whose structure is in accordance with the predefined XML scheme;

(b) processing of protein sequences by determining T-cell epitopes of different lengths, using one of the supported MHC binding predictors;

(c) processing of protein sequences with determined disordered/ordered regions, using one of the supported disorder predictors;

(d) hydropathy index calculation for peptides of different lengths;
(e) processing of protein sequences by determination of disorderbinding regions using the ANCHOR predictor.

The main characteristics of the EpDis tools are as follows:

- It enables the execution of all supported prediction methods, serial or parallel. For the rapid and efficient execution of predictors, the tool provides two predefined strategies for caching the results of prediction. The choice of strategy for caching depends on the (user configurable) selected configuration. Local caching is applied when the component of the tool that works with the database is disabled, otherwise database caching is applied. The benefit of caching is particularly evident in situations where predictor execution takes a very long time or in cases where a long protein sequence is the input. In subsequent requests for an identical action, the results are obtained from the cache without repetition of calculation.
- Modularity, configurability and extensibility. The addition and use of new prediction methods, result analyzers and result plotters are simple and straightforward. When properly added and configured, the predictor is discovered automatically by the tool on startup and is offered later for use.
- It enables visualization of the predictions in the form of charts and their export to various graphic formats, such as PNG, PDF, SVG, and EPS in high quality. Charts are customizable; therefore, the colors and annotations can be adjusted either in the configuration files or during runtime.
- It enables export of the predictions into flat files in raw format as the original output form of the used prediction method.
- It enables storing of the results of the predictions directly into the relational database tables using commands generated by the MassPred tool through the MassPred component. Tables are designed to allow subsequent efficient querying, search and comparisons of predictions, or the application of different data mining and machine-learning techniques (see SQL DDL in Appendix A and html documentation).
- All DML commands are initially generated according to the IBM DB2 syntax, which is by default RDBMS in EpDis-MassPred systems. For rapid import, in the case of mass insertion MassPred offers the use of the LOAD database utility which provides the best performance results.
- It supports massive execution of the predictors on a large set of proteins by use of the MassPred tool. EpDis utilizes MassPred through the MassPred component, providing an easier specification of input parameters and their forwarding to the MassPred tool.
- Although EpDis supports combining simultaneous execution of all supported methods, any of these methods can be used individually. It is also possible to combine predictors from two arbitrary prediction areas of the four available (disorder prediction, epitope prediction, disordered-binding prediction and, hydropathy calculation), and to visualize the obtained results.

The Epdis tool is written in Java and consists of five basic modules (see in Appendix A).

Some examples of MassPred/EpDis applications are presented in the next section.

3. Examples of application

In the examples of application of the MassPred and EpDis systems, the proteins EBNA1 (UniProt Acc: P03211) and p53 (UniProt Acc: P04637), served as the input.

3.1. MassPred

MassPred, is a script oriented system. It is started by running the script *work.sh* with two input parameters: name of the MassPred application configuration file and the name of the fasta file or directory which includes fasta files. For example,

```
/usr/local/masspred/work.sh configuration.ish
  p03211.faa
```

The configuration file (configuration.ish) includes parameters related to the current computer system (for example, the number of concurrent CPUs), and parameters related to supported predictors. An example of a configuration file is

```
CPU_NUMBER=8
WORK_HYDRO=yes
WORK_ISUNSTRUCT=yes
WORK_VSL2=yes
WORK_NETMHC_1_34A=yes
NETMHC_1_34A_ALLELE_FILE=NetMhc.3.4a.pseudo
NETMHC_1_34A_LENGTH_FROM=8
NETMHC_1_34A_LENGTH_TO=9
...
```

File p03211.faa includes the fasta version of protein(s) on which the predictors are to be applied. The file can include one or more proteins in fasta format. For example,

```
>gi|119110|sp|P03211.1|EBNA1_EBVB9 RecName:
  Full=Epstein-Barr nuclear
antigen 1; Short=EBNA-1; Short=EBV nuclear antigen 1
MSDEGPGTGPGNGLGEKGDTSG
  PEGSGGSGPQRRGGDNHGRGRGRGRGRGGGRPGAPGGSGSGPRHRDGV
RRPQKRPSCIGCKGTHGGTGAGAGAGGAGAGGAGAGGGAGAGGGAGGAG
  GAGGAGAGGGAGAGGGAGGAG
GAGAGGGAGAGGGAGGAGAGGGAGGAGGAGAGGGAGAGGGAGGAGAGG
  GAGGAGGAGAGGGAGAGGAGGA
GGAGAGGAGAGGGAGGAGGAGAGGAGAGGAGAGGAGAGGAGGAGAGGAG
  GAGAGGAGGAGAGGGAGGAGA
GGGAGGAGAGGAGGAGAGGAGGAGAGGAGGAGAGGGAGAGGA
  GAGGGGRGRGGSGGRGRGGSGGRGRGGS
GGRRGRGRERARGGSRERARGRGRGR
  GEKRPRSPSSQSSSSGSPPRRPPPGRRPFFHPVGEADYFEYHQE
GGPDGEPDVPPGAIEQGPADDP
  GEGPSTGPRGQGDGGRRKKGGWFGKHRGQGGSNPKFENIAEGLRALLA
RSHVERTTDEGTWVAGVFVYGGSKTSLYNLRRGTA
  LAIPQCRLTPLSRLPFGMAPGPGPQPGPLRESIVC
YFMVFLQTHIFAEVLKDAIKDLVMTKPAPTC
  NIRVTVCSFDDGVDLPPWFPPMVEGAAAEGDDGDDGDEG
GDGDEGEEGQE
>gi|89902357|ref|YP_524828.1| 50S ribosomal protein
  L7/L12 [Rhodoferax ferrireducens T118]
MAFDKDAFLTALDSMTVMELNDLVKAIEEKFGVSAAAMSAPAAGGAVAA
  VAEEKTEFNVVLLEAGAAKVS
VIKAVREITGLGLKEAKDMVDGAPKNVKEGVSKVDAEAALKKLLDAGA
  KAELK
...
```

After execution, Masspred creates a directory with the added suffix . *out* to the name of the input file. The directory contains gzipped files with results for every required prediction. For example, for the part of the configuration file shown above, a directory p03211.faa.out was created with the following files:

- epitope_success.load.gz - gzipped file with the predicted data from the epitope predictors,
- hydro.load.gz - gzipped file with calculated hydrophobicity values for every AA,
- region_success.load.gz - gzipped file with the predicted data from the disorder predictors.

All files are in TAB separated format and prepared for loading into RDBMS. For example, the content of region_success.load file is

```
119110   P03211.1    p03211 1   413 D IsUnstruct
119110   P03211.1    p03211 414 414 O IsUnstruct
119110   P03211.1    p03211 415 480 D IsUnstruct
119110   P03211.1    p03211 481 541 O IsUnstruct
119110   P03211.1    p03211 542 554 D IsUnstruct
119110   P03211.1    p03211 555 613 O IsUnstruct
119110   P03211.1    p03211 614 641 D IsUnstruct
89902357 YP_524828.1 p03211 1   55  D IsUnstruct
89902357 YP_524828.1 p03211 56  84  O IsUnstruct
89902357 YP_524828.1 p03211 85  123 D IsUnstruct
119110   P03211.1    p03211 1   479 D VSL2b
119110   P03211.1    p03211 480 497 O VSL2b
119110   P03211.1    p03211 498 498 D VSL2b
119110   P03211.1    p03211 499 541 O VSL2b
119110   P03211.1    p03211 542 553 D VSL2b
119110   P03211.1    p03211 554 606 O VSL2b
119110   P03211.1    p03211 607 641 D VSL2b
89902357 YP_524828.1 p03211 1   5   D VSL2b
89902357 YP_524828.1 p03211 6   36  O VSL2b
89902357 YP_524828.1 p03211 37  44  D VSL2b
89902357 YP_524828.1 p03211 45  84  O VSL2b
89902357 YP_524828.1 p03211 85  123 D VSL2b
```

3.2. EpDis

The EpDis user interface panel is divided into two parts, referred to as Execution (displayed in Fig. 2) and MassPred (displayed in Fig. 3). The generated screen with visualized results is presented in Fig. 4.

3.2.1. Execution module

The execution module offers loading of protein sequences in four ways, as described above and presented in Fig. 2, part A; accessing eight, currently supported, disorder prediction methods and one disordered binding method (Fig. 2, part B); accessing seven T-cell prediction methods (Fig. 2, part C), and accessing methods that calculate peptide hydropathy (Fig. 2, part D). Disorder predictors, apart from the protein sequence, do not require specification of other parameters. Prediction methods for T-cell epitope detection (based on MHC binding affinities) require parameter specification that depends on the prediction method. For each epitope predictor there is a form with a corresponding drop-down list of available alleles, peptide length, binding threshold and window length. The user can choose only those parameter values for which the selected predictor can carry out predictions. For hydropathy scoring, it is also possible to choose the window length, which can be the same as peptide length (in that way it is possible to compare the hydropathy score with the binders and non binders), or any length between 8 and 19 AA. From the menu item 'Experimental', a user can view the existing experimental data or add a new set of data (explained in details in the documentation). New sets can be added by selecting the existing XML file using File Browser. The structure of the XML file must be compliant with the predefined XML scheme. The user can observe the structure of the added experimental results and delete them. Deletion of a selected set will remove the XML file containing it from the file system.

3.2.2. MassPred module

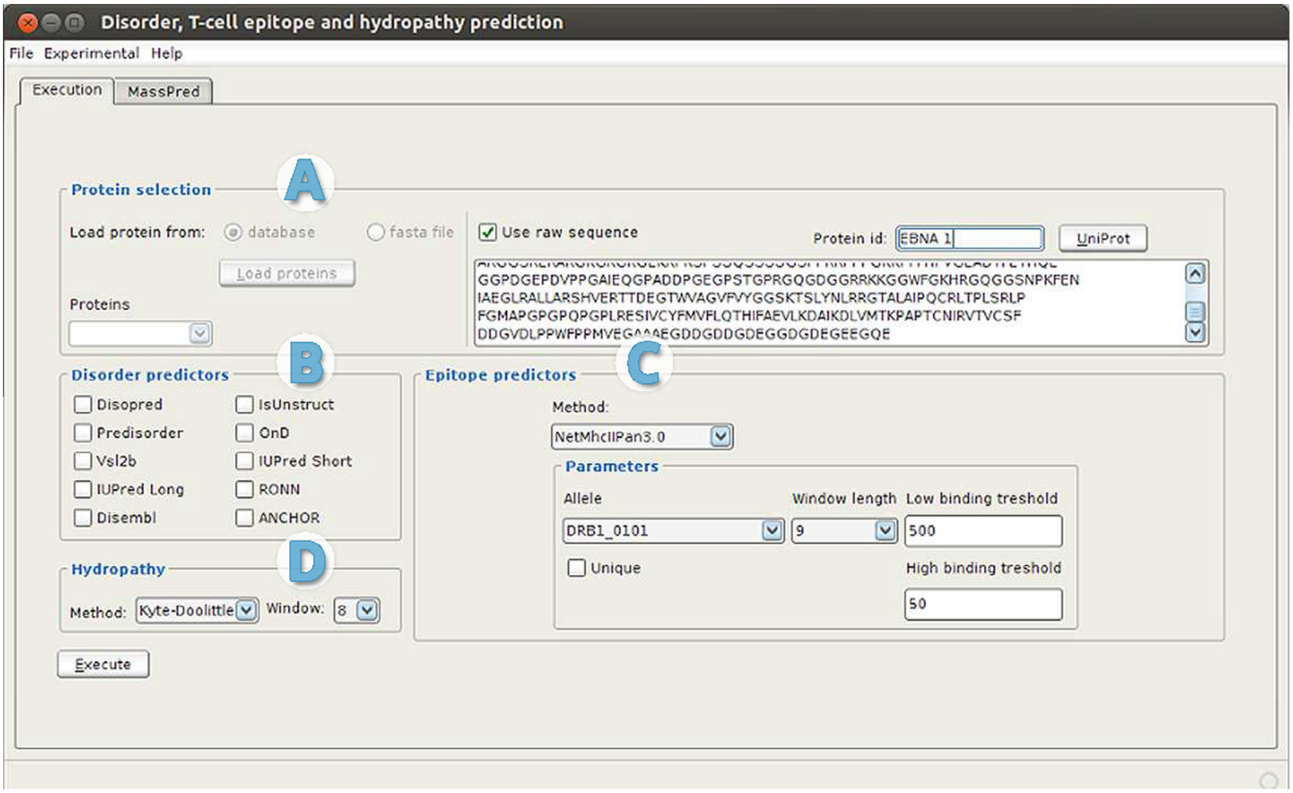
This module enables the running of the MassPred tool from graphic mode. Turning on this component is optional and configurable. Apart from extending the functionalities of EpDis, the main purpose of this module is to facilitate the use of the MassPred tool itself, by allowing the user to specify its parameters simply by selecting the target predictors. The MassPred module has a predefined default configuration, but it allows storing and loading of a custom configuration as well. After execution of MassPred, the output is redirected and displayed to the user in the separate window. The interface for the MassPred module is shown in Fig. 3.

3.2.3. Visualization of the result

A special part of the execution module is the visualization submodule. For each prediction method there are corresponding classes whose instances are responsible for interpreting and displaying the prediction results. The visual presentation of the output after applying selected predictors on the individual protein is shown in Fig. 4 through an example of the human tumor suppressor proteinp53, (UniProt Acc: P04637), involved in the control of cell-cycle and apoptosis. The first panel (PART A in the figure) represents the output of all chosen disorder predictors along with the experimentally determined structure, if defined for a certain protein. For disorder prediction, the output is the probability score, indicating the likelihood of the residue to be a part of the disorder region along each position in the protein sequence. Residues with a score above the horizontal threshold line are predicted to be disordered and with a score below the threshold line are predicted to be ordered. The tool provides the utility of identifying a consensus among all of the chosen predictions. The comparison of predictions helps to identify the possible omission and mismatch or the most accurate prediction if the experimental data are included. The output of the disorder binding region predictor (ANCHOR) could also be represented in the PART A in the figure. Adding experimental data enables the evaluation of the quality of the results obtained by the selected predictor, or analysis of the position of the elements of an experimentally verified structure within a protein. The elements of secondary or disordered structure (experimentally verified) in this example are shown as a horizontal line in three colors for each type of structure: red represents disorder, black represents experimentally determined order and an unknown structure is shown in gray (Fig. 4A).

On the right side of this graph is a table (Fig. 4B) of exported intervals for each prediction. It is possible to choose what is to be presented, whether intervals of ordered or of disordered parts of a protein. The table also contains intervals of matched predictions (consensus) of all predicted disordered and ordered regions, as well as intervals overlapping with experimentally proven secondary structural elements presented as intervals, if available. Below the disorder predictions there are intervals that contain predicted MHC binding peptides (T cell epitopes), experimentally verified naturally processed T-cell epitopes or experimentally verified MHC binding peptides, and hydrophobic intervals. Finally, an intersection of all intervals is presented from which we can see whether the epitopes are located in the corresponding region of the protein. One can simply click on a predictor's name in the table display and see the results of the predictor in the original form in textual format.

The visualization also contains a representation of hydropathy for each peptide (of selected size) in sequence, calculated using



Protein selection

Load protein from: ● database ○ fasta file

Proteins

☑ Use raw sequence

Protein id: EBNA 1

GGPDGEPDVPPGAIEQGPADDPGEGPSTGPRGQGDGGRRKKGGWFGKHRGQGGSNPKFEN
IAEGLRALLARSHVERTTDEGTWVAGVFVYGGSKTSLYNLRRGTALAIPQCRLTPLSRLP
FGMAPGPGPQPGPLRESIVCYFMVFLQTHIFAEVLKDAIKDLVMTKPAPTCNIRVTVCSF
DDGVDLPPWFPPMVEGAAAEGDDGDDGDEGGDGDEGEEGQE

Disorder predictors

□ Disopred □ IsUnstruct
□ Predisorder □ OnD
□ Vsl2b □ IUPred Short
□ IUPred Long □ RONN
□ Disembl □ ANCHOR

Epitope predictors

Method: NetMhcIIPan3.0

Parameters

Allele: DRB1_0101

Window length: 9

Low binding treshold: 500

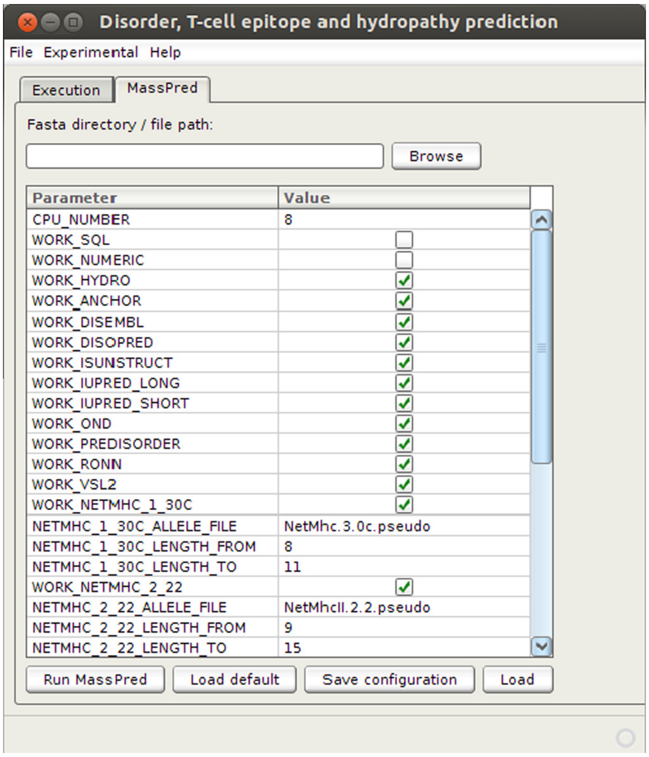
□ Unique

High binding treshold: 50

Hydropathy

Method: Kyte-Doolittle Window: 8

Fig. 2. Execution tab of the EpDis tool.



Fasta directory / file path:

Parameter	Value
CPU_NUMBER	8
WORK_SQL	□
WORK_NUMERIC	□
WORK_HYDRO	☑
WORK_ANCHOR	☑
WORK_DISEMBL	☑
WORK_DISOPRED	☑
WORK_ISUNSTRUCT	☑
WORK_IUPRED_LONG	☑
WORK_IUPRED_SHORT	☑
WORK_OND	☑
WORK_PREDISORDER	☑
WORK_RONN	☑
WORK_VSL2	☑
WORK_NETMHC_1_30C	☑
NETMHC_1_30C_ALLELE_FILE	NetMhc.3.0c.pseudo
NETMHC_1_30C_LENGTH_FROM	8
NETMHC_1_30C_LENGTH_TO	11
WORK_NETMHC_2_22	☑
NETMHC_2_22_ALLELE_FILE	NetMhcII.2.2.pseudo
NETMHC_2_22_LENGTH_FROM	9
NETMHC_2_22_LENGTH_TO	15

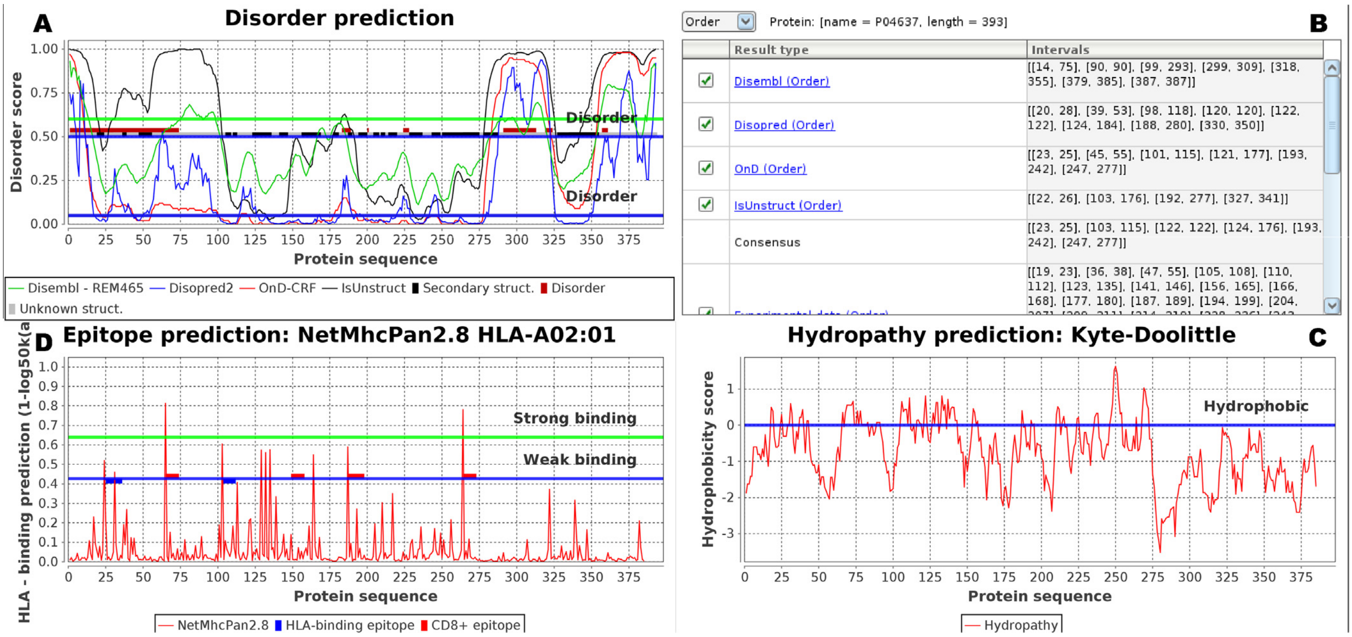
Fig. 3. An interface for MassPred module.

the Kyte–Doolittle or Hopp–Woods scales (Fig. 4C). Peptide hydropathy, is calculated as the average value of hydropathy of each amino acid in the peptide according to selected scale.

In the case of the MHC binding peptides (T-cell epitopes) prediction, the predictions are shown for all peptides of chosen size, where the basic value that is taken into consideration is 1 – log 50 k (aff) (normalized value for the binding affinity of a peptide to specific molecules of MHC classes I or II). Programs perform binding-affinity predictions for peptides of various sizes. Depending on the predicted affinity value, the peptides are classified as non-epitopes (below the blue line), weak (above the blue and below the green line) or strong epitopes (above the green line) (Fig. 4D). Experimentally verified T-cell epitopes are represented as red lines in the actual length of the epitope, along the protein sequence (Fig. 4, part D), while the blue line stands for experimentally determined MHC binding peptides. Each graphic can be individually exported. Experimentally validated epitopes (from Tantigen database), binding to A∗0201 allele or A2 serotype were found to be prevalently hydrophobic, as shown in Fig. 4D.

3.2.4. The combined usage of methods for the disorder, disordered binding region and T-cell epitope prediction

The principles of the combined usage of different methods in the field of protein structure/function analysis are described by various authors, frequently using the example of human p53, which is one of the most complex IDPs. Comparing several disorder predictors, optimized for various typical lengths of disorder, and disordered binding region – prediction by ANCHOR, Meszaros and colleagues gave a profile of p53 disordered binding regions encompassing short linear binding motifs [40]. Huart and Hupp have shown that p53 long disordered regions were enriched in potential disorder-based binding motifs which overlap numerous sites of post-translational modifications, and suggested how amino acid modifications evolved to regulate dynamically the p53 interactome [18]. One mechanism by which the p53 protein exerts its antiproliferative activity is by inducing the transcription of genes that control cell growth through the interaction with transcription



Order ▾ Protein: [name = P04637, length = 393]

	Result type	Intervals
☑	Disembl (Order)	[[14, 75], [90, 90], [99, 293], [299, 309], [318, 355], [379, 385], [387, 387]]
☑	Disopred (Order)	[[20, 28], [39, 53], [98, 118], [120, 120], [122, 122], [124, 184], [188, 280], [330, 350]]
☑	OnD (Order)	[[23, 25], [45, 55], [101, 115], [121, 177], [193, 242], [247, 277]]
☑	IsUnstruct (Order)	[[22, 26], [103, 176], [192, 277], [327, 341]]
	Consensus	[[23, 25], [103, 115], [122, 122], [124, 176], [193, 242], [247, 277]]
☑	Experimental data (Order)	[[19, 23], [36, 38], [47, 55], [105, 108], [110, 112], [123, 135], [141, 146], [156, 165], [166, 168], [177, 180], [187, 189], [194, 199], [204,

Fig. 4. An example of the EpDis-MassPred application for human tumor suppressor protein p53 (UniProt Acc No: P04637). The application displays data on: (A) Disorder/order prediction, obtained using DISOPRED2, OnD-CRF, DisEMBL_Remark465 and IsUnstruct predictors. Residues with a score above the horizontal threshold line are predicted to be disordered and with a score below the threshold line are predicted to be ordered (shown in different colors for different predictors). The red boxes above the middle line represent experimentally characterized regions of disorder (from DisProt database), while order is defined as determined secondary structure (from UniProt database) is represented by black boxes indicating both α-helical and β structures. (B) Intervals of predicted ordered regions, epitopes and hydrophobic regions (according to Kyte–Doolittle AA scale). (C) Hydropathy calculation, according to Kyte–Doolittle AA scale. (D) Nonamer epitope prediction for A∗0201 allele, using NetMhcpan T cell epitope predictor. The experimentally validated epitopes, were represented as HLA I ligands (blue bars) and as CD8+ T cell-inducing (red bars). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

factors or activators or inhibitors of p53's transactivation function. Many of these interactions have been mapped in the p53 sequence (reviewed in [51]), Fig. 5A. Fig. 5B shows EpDis-MassPred application of the outputs of several disorder prediction methods, while Fig. 5C and D display EpDis-MassPred application of the output of disordered binding region – predictor ANCHOR, and experimentally verified T-cell epitopes (Tantigen database) [6] [21], respectively.

The p53N-terminal region (residues 1–101) contains two transcriptional activation domains (TAD1, residues 1–42 and TAD2, residues 43–63) and the polyproline subdomain, important for the apoptotic activity of p53 (PP, residues 62–91). TAD interacts with TFIID, TFIIH, Mdm2, RPA, CBP/p300 and CSN5/Jab1 among many other proteins. The segment between residues 17–27 binds to p53 ligand ubiquitin ligase (MDM2), residues 33–56 binds to replication protein A (RPA 70N), residues 47–55 binds to Tfb1 subunit of transcription activation II H (TFIIH) and residues 45–58 binds to the B subunit of RNA polymerase II, as represented in the structure complexes in PDB (Fig. 5A). All analyzed disorder-prediction methods, (Fig. 4B and C) exhibit a lower score (a dip in the prediction profile) in these binding regions, however to a varying degree, as observed also by Meszaros and coworkers. Some methods, such as VSL2B (trained on both long and short disorder sequences) predict the whole interacting region to be disordered, giving one extended dip covering approximately residues 17–70, while RONN and OnD-CRF react to the presence of transient structure by assigning a score close to a threshold of 0.5, meaning that they cannot correctly classify these regions as ordered or disordered. DISOPRED2, IUPred-L, IsUnstruct or DisEMBL_Remark465 and PONDR VL-XT give two distinct dips in this region – first, corresponding to the MDM2 binding region and the second, corresponding to other three, overlapping ligand-binding regions. TAD was shown by NMR to be populated with preformed structures which fold upon binding to their ligands. MDM2, RPA70N [18], and Tfb1 [7] binding sites form amphiphatic α-helices via the interaction with their structured partners. The high-confidence predictions of ANCHOR covering these binding regions in the N-terminal domain, Fig. 5C, is the strongest prediction-level evidence for the presence of disordered binding regions (as opposed to coiled-coil region or a short collapsed structure) [40].

All analyzed disorder prediction methods assign a relatively low score to the majority of the central DNA binding domain (DBD) of p53 (spanning residues 102–292), Figs. 4A and 5B. The validity of the prediction can be ascertained through the solved secondary structure, with numerous β-strands in the domain region, Fig. 5B (UniProt database). Segment around residues 150–190, with disorder scores above the threshold line, according to the majority of predictors, could be a disordered region connecting two structured domains within DBD. The high-confidence predictions of ANCHOR points towards two putative disordered binding regions at the borderlines of DBD, Fig. 5C, overlapping the protein kinase CHK2 motifs [18] and secondary structural elements. The regions in the DNA binding domain, represented by structure complexes in the PDB, include interactions with p53 endogenous partners: DNA, the BRCT domain of 53BP1 and the SH3 domain of 53BP2, Fig. 5A.

The C-terminal region (residues 293–393), which is prevalently disordered, has a high consensus between different prediction methods concerning the non-interacting disordered regions. In the tetramerization region (TD, residues 325–356) and the regulatory domain (RD), residues 363–393), that is able to bind to a multitude of different partners, all methods react with a lower score, (a dip in the disorder prediction profile). The presence of disordered binding regions is supported by the high confidence of ANCHOR prediction, and the presence of β-strand (residues 326–333) and an α-helix (residues 335–354) structure in TD. The overlapping p53 regions that mediate interactions with several functional

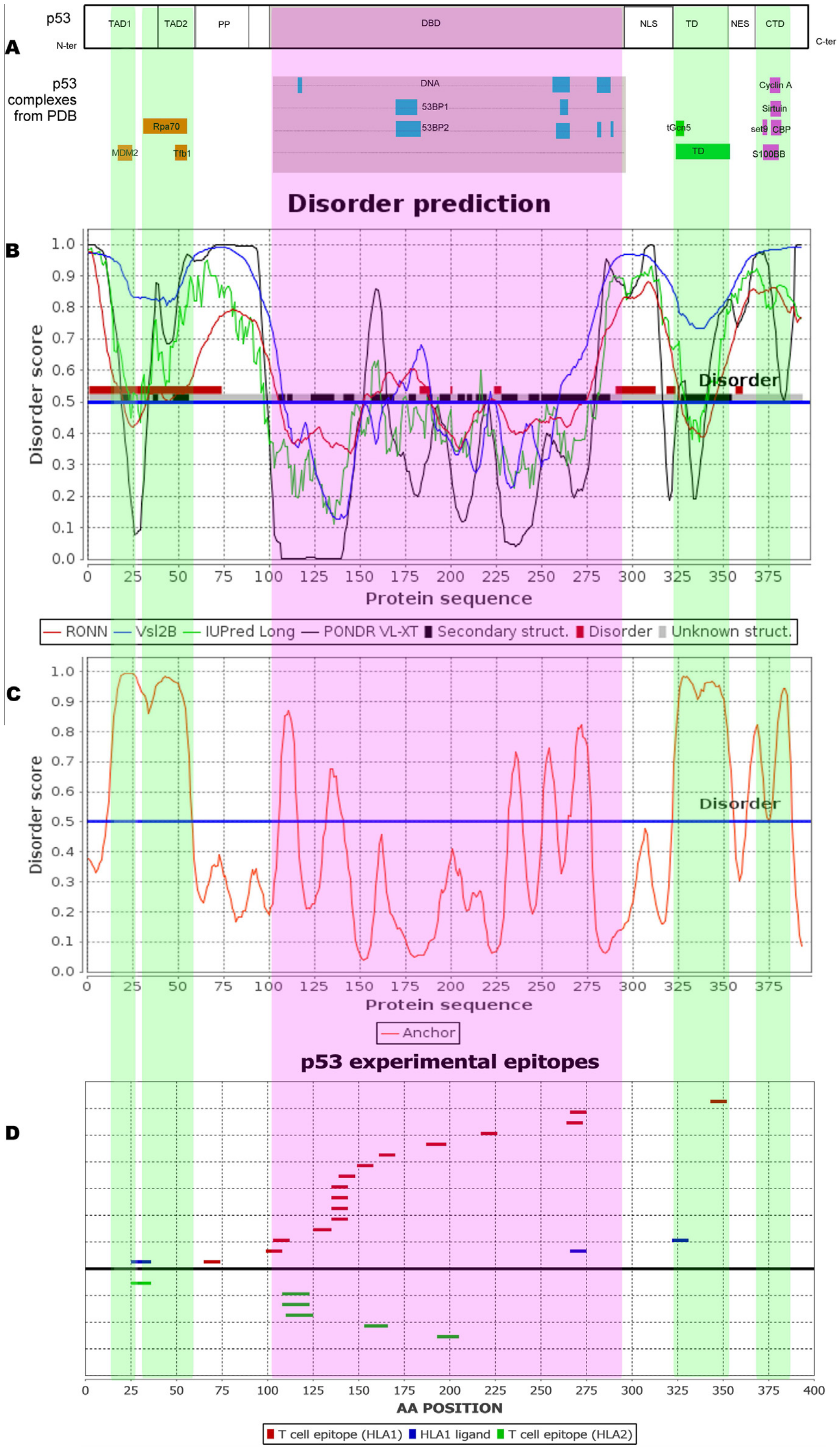


Fig. 5. Summary of p53 interactions, structure, predicted disordered binding sites and experimentally validated T-cell epitopes. (A) TAD1 and TAD2 – the N-terminal transcriptional activation domains; PP – the polyproline subdomain; DBD - the central DNA binding domain; NLS – the nuclear localization signal, TD – the tetramerization domain; NES – the nuclear export signal, and RD – the regulatory domain. The known biologically relevant binding sites on human p53, represented in structure complexes in the PDB, (from [51]), are shown with colored boxes, labeled with the name of the binding partner. The red boxes above the middle line represent experimentally characterized regions of disorder (from DisProt database), while the order (determined secondary structure from UniProt database) is represented by black boxes. (B) The disorder prediction outputs from VSL2B, IUPred-L, RONN and PONDR VL-XT predictors (generated with EpDis-MassPred tools). (C) The disordered binding regions, predicted by ANCHOR (generated with EpDis-MassPred tools). (D) The experimentally validated epitopes, represented above the middle line, are HLA-I epitopes. HLA-I ligands are shown with blue bars, while CD8+ T cell-inducing epitopes are shown with red bars. CD4+ T cells-inducing HLA-II epitopes are represented with green bars below the middle line (from Tantigen database). The central, ordered DNA binding domain is shown in transparent pink and experimentally verified disordered binding regions are shown in transparent green while the rest of the protein is disordered and is shown in white. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

regulators, predicted as disordered binding regions by ANCHOR, were mapped to the C-terminal end of p53 from residues 374 to 388, including the protein kinase cyclin A, deacetylase sirtuin, the activator CBP, or repressor S100ββ. This p53 region displays all three major secondary structure types, becoming a helix when binding to S100ββ, a sheet when binding to sirtuin, and a coil with two distinct backbone trajectories when binding to CBP and cyclin A2 [51].

Naturally processed and presented wt p53 HLA-I epitopes, known, so far, to induce CTL in preclinical and clinical studies (reviewed in [48] and in Tantigen database), were found in ordered or order/disorder transition regions, defined by the majority of analyzed disorder predictors and ANCHOR, respectively, Fig. 5B–D. Notably, epitopes are concentrated in the DBD and disordered binding regions and linear motifs, and absent in the extremely disordered regions: 1–24 and 58–99 in TAD, or 272–321 in the C-terminal domain. An exception is naturally processed epitope 68–73, on the borderline of the disordered PP domain. However, this epitope partly overlaps the PXXP motif in PP domain, which binds directly to the transcriptional coactivator p300 [10]. High-affinity binding HLA-A2 peptide, 25–35, located at the borderline of the MDM2 binding region (residues 15–29), was not found to be naturally processed and presented to CTL (Chikamatsu and DeLeo, unpublished results, cit. [21], although experiments with synthetic peptide 25–35 have shown that this epitope is not deleted from the normal immune repertoire [49].

All naturally processed CD4+ T helper determinants (Tantigen database; [6] (except $p53_{25-35}$ epitope) are located in the DNA binding domain, defined as ordered by the majority of disorder predictors. p53 CD4+ epitopes were also found to overlap the picks in the ANCHOR profile. As for $p53_{25-35}$, previous studies strongly discounted the possibility that this peptide might act as naturally processed cytotoxic or helper T cell epitope, although it was found to be a high-affinity binding HLA-A2 peptide. This epitope is on the borderline of the region 15–29, with high confidence of ANCHOR prediction Fig. 5C, which undergoes disorder-to-order transition upon binding to MDM2 [62]. The reason why epitope 25–35 was not found to be naturally generated in several experimental studies, could be because poorly expressed under natural conditions. The $p53_{25-35}$ peptide was found to be naturally presented, in autologous dendritic cells transfected with wt or mutant p53 cDNA, inducing CD4+ T cells, restricted, at a minimum, with HLA-DR7 and -DR11 alleles [21]. Immunization with a p53 overlapping synthetic long peptides (SLP vaccine), which are supposed to be exogenously processed and presented by APC, induced p53-specific Th immune responses (although dominated by Th2 cytokines) in ovarian cancer patients [28]. The most immunogenic was the central part 116–248. According to the report by van der Burg and colleagues (citated by [21] for colon cancer subjects, CD4+ T cell proliferative response to 30-mer peptides grouped in pools, corresponding to residues 1–142, 129–270 and 257–393, was maximal against 129–270 pool, whereas unresponsive to 1–142 peptide pool. These results confirmed that the most immunogenic part of the wt p53, inducing HLA-II specific response in humans, is the central, DNA binding domain with predominantly ordered structure, as discussed previously for HLA-I immune response to wt p53 epitopes. The positional biases of T-cell epitopes in ordered protein regions and borderlines of predicted disordered binding sites or protein-binding linear motifs, observed for p53, is reminiscent of the results obtained using an EpDis tool on several tumor-associated antigens and systemic nuclear autoantigens [52]. These positional biases could originate from a higher percentage of bulky hydrophobic, polar and charged amino acids in structured regions and epitopes as compared to unstructured regions of proteins and non-epitopes or to a high proteolytic sensitivity of disordered regions, and might influence the capture of antigens, their processing and subsequent epitope presentation and immunodominance.

4. Related work

We have already discussed different (single) predictors for the disorder, disordered binding, T cell epitope prediction and hydropathy calculation. However, there are tools that combine different predictors. To our knowledge, there are several disorder meta predictors (which combine more disorder predictors) like GeneSilicoMetaDisorder [25], MFDp [43], and MeDor [31]. The first one does not have a stand alone application and cannot be used for a large scale analysis. The other two include a subset of predictors that are already involved in EpDis-MassPred system. For the prediction of disorder-binding regions, there is an ANCHOR server. The server includes IUPred disorder predictor for comparative analysis of disorder and disordered binding regions. Both predictors are included in EpDis-MassPred system. Moreover, EpDis-MassPred system enables comparative analyses of the ANCHOR predictions with wider sets of disorder predictors. For MHC binding prediction there are very valuable tools: IEDB MHC I and IEDB MHC II binding prediction tools (http://tools.immuneepitope.org/main/tcell/) which offer a combination of the most widely used MHC binding predictors. The tools are available as servers, as well as stand alone applications. Models in these tools are regularly updated with new data. However, these tools are intended for MHC binding prediction only and the comparison of different MHC binding predictors. The outputs of these tools are suitable for import into the database and thus a possible large-scale analysis, but for further research certain computer and programming skills are required. EpDis-MassPred system and IEDB tools, both involve overlapping set of MHC binding predictors. In fact, to expand a set of MHC binding predictors we included models from these tools in EpDis-MassPred tools that were not available as a stand alone application elsewhere (models to predict MHC class I binding epitope: ann [44], smm [53], smmpmbec [24]).

None of the described tools offer the possibility of combining various characteristics of the protein, or a comparative overview of the different characteristics. The tools which include MHC binding predictors do not offer a visual representation of the results. Furthermore, none of the existing tools offer the possibility of entering a known (experimentally determined) characteristic (not necessarily related to the same characteristics they are determining). Finally, none of the existing tools enable massive parallel application of any integrated predictors, or storing results into relation database.

5. Discussion/Conclusion

At present, there is a large number of T-cell epitope or disordered region prediction softwares which, although of high accuracy, cannot be considered as fully reliable. To overcome the problem of overfitting to certain data, several predictors based on different methodologies should be used. The potential of prediction software is considerable when there are enough data to build a good model, which increases the need for improved tools for large scale analyses as well as further refining of existing methods for their combined use.

In this paper, we described a system which consists of two tools suitable for answering important questions through a bioinformatics approach. The developed system is intended to be used by researchers who study the structure of proteins, the properties of disordered proteins and the characteristics and position of T-cell epitopes in a protein. The system currently includes support and use of predictors for disordered regions, disordered-binding regions, MHC binders and T-cell epitope identifiers, and methods for calculating hydropathy. The system is flexible and can be easily extended with new functionalities, and is very useful for preparing data for large-scale analysis. For all available predictors outputs/results are displayed and stored in a uniform format suitable for further analysis. The system allows for easy input of experimental data linked to various structural aspects of proteins, such as disordered/ordered regions, T-cell epitopes, hydrophobic/hydrophilic regions, or any other structural features. This allows for a comparatively effortless comparison of results obtained from different predictors and with experimental data, parallel use different predictors and assessment of the quality of results acquired by different predictors. Our system can serve as a basis for the development of novel methods or meta-predictors, can reduce the time required for data collection and analysis, and provides a platform that makes the addition of new methods and the creation of combinations with existing methods straightforward. The system allows for multiple (mass) launching of predictors on more fasta files with an arbitrary number of proteins. The developed tools simplify and enhance the time-consuming task of assessing and comparing different types of prediction methods. The only shortcoming of the tools, presented here, is related to difficulties in installing some of the predictors, as the installation of external predictors is required. Due to licensing issues, we could not include the stand-alone versions of these methods in our tools; however, this problem is circumvented and semi-automated by the use the MassPred tool (a detailed explanation is provided in the manual, along with all of the difficulties that were encountered when installing the predictors, together with explanations on how to overcome them). EpDis comes with an easy installer and it is enough to simply follow all of the steps for installation (detailed instructions for setting up the configuration file are provided with the tool).

Over the past 3 years we have used these tools exhaustively for various tasks. The complete system has proven to be robust, scalable and fault free. It has been used in different research scenarios with different input data on different Ubuntu Linux machines. The largest task was to determine disordered regions in the dataset with more than 8,500,000 proteins downloaded from NCBI, on 48 CPU core computers. The investigation of the correlation between predicted disordered protein regions and the occurrence of epitopes in these regions, hydropathy properties of epitopes and nonepitopes within the ordered and disordered regions was presented in Mitić et al. [42]. Our research was conducted on 619 proteins and a total of 1986 alleles of both MHC classes (1469 for MHC class I, and 517 for MHC class II), which would not have been feasible without the tools described in this paper. Comparing results of epitope and disorder, and disordered binding region predictions with experimentally verified epitopes and secondary structural elements, T cell epitope instances could be enriched with information of structural instances, which could facilitate the understanding of the role that protein structure and function may have on the development of immune response [52].

In future we plan to expand the system with new functionalities, and to enable downloading proteins from the NCBI database through a web interface as we have done for the UNIPROT database.

Conflict interest

None declared.

Acknowledgment

The work presented has been financially supported by the Ministry of Education, Science and Technological Development, Republic of Serbia, Project Nos. 174021 and 174002.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jbi.2016.01.016.

References

[1] M. Andreatta, E. Karosiene, M. Rasmussen, A. Stryhn, S. Buus, M. Nielsen, Accurate pan-specific prediction of peptide-MHC class II binding affinity with improved binding core identification, Immunogenetics 67 (11–12) (2015) 641–650, http://dx.doi.org/10.1007/s00251-015-0873-y.

[2] V. Brusic, V.B. Bajic, N. Petrovsky, Computational methods for prediction of T-cell epitopes–a framework for modelling, testing, and applications, Methods 34 (4) (2004) 436–443, http://dx.doi.org/10.1016/j.ymeth.2004.06.006.

[3] A.J. Bordner, Towards universal structure-based prediction of class II MHC epitopes for diverse allotypes, PLoS One 5 (12) (2010) e14383, http://dx.doi.org/10.1371/journal.pone.0014383.

[4] Y. Cheng, C.J. Oldfield, J. Meng, P. Romero, V.N. Uversky, A.K. Dunker, Mining alpha-helix-forming molecular recognition features with cross species sequence alignments, Biochemistry 46 (47) (2007) 13468–13477, http://dx.doi.org/10.1021/bi7012273.

[5] C. Chica, F. Diella, T.J. Gibson, Evidence for the concerted evolution between short linear protein motifs and their flanking regions, PLoS One 4 (7) (2009) e6052, http://dx.doi.org/10.1371/journal.pone.0006052.

[6] K. Chikamatsu, A. Albers, J. Stanson, W.W. Kwok, E. Appella, T.L. Whiteside, A.B. DeLeo, P53 (110–124) -specific human CD4+ T-helper cells enhance in vitro generation and antitumor function of tumor-reactive CD8+ T cells, Cancer Res 63 (13) (2003) 3675–3681.

[7] P. Di Lello, L.M. Jenkins, T.N. Jones, B.D. Nguyen, T. Hara, H. Yamaguchi, J.G. Omichinski, Structure of the Tfb1/p53 complex: insights into the interaction between the p62/Tfb1 subunit of TFIIH and the activation domain of p53, Mol. Cell 22 (6) (2006) 731–740, http://dx.doi.org/10.1016/j.molcel.2006.05.007.

[8] I. Darren, R.F. Darren, I. Doytchinova, Improving in silico prediction of epitope vaccine candidates by union and intersection of single predictors, World J. Vacc. 01 (02) (2011) 15–22, http://dx.doi.org/10.4236/wjv.2011.12004.

[9] F.M. Disfani, W.L. Hsu, M.J. Mizianty, C.J. Oldfield, B. Xue, A.K. Dunker, V.N. Uversky, L. Kurgan, MoRFpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins, Bioinformatics 28 (12) (2012) i75–i83, http://dx.doi.org/10.1093/bioinformatics/bts209.

[10] D. Dornan, H. Shimizu, L. Burch, A.J. Smith, T.R. Hupp, The Proline Repeat Domain of p53 Binds Directly to the Transcriptional Coactivator p300 and Allosterically Controls DNA-Dependent Acetylation of p53, Mol. Cell. Biol. 23 (23) (2003) 8846–8861, http://dx.doi.org/10.1128/mcb.23.23.8846-8861.2003.

[11] Z. Dosztanyi, V. Csizmok, P. Tompa, I. Simon, IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content, Bioinformatics 21 (16) (2005) 3433–3434, http://dx.doi.org/10.1093/bioinformatics/bti541.

[12] Z. Dosztanyi, B. Meszaros, I. Simon, Bioinformatical approaches to characterize intrinsically disordered/unstructured proteins, Brief Bioinf. 11 (2) (2010) 225–243, http://dx.doi.org/10.1093/bib/bbp061.

[13] C. Fang, T. Noguchi, D. Tominaga, H. Yamana, MFSPSSMpred: identifying short disorder-to-order binding regions in disordered proteins based on contextual local evolutionary conservation, BMC Bioinf. 14 (2013) 300, http://dx.doi.org/10.1186/1471-2105-14-300.

[14] P. Gilchuk, C.T. Spencer, S.B. Conant, T. Hill, J.J. Gray, X. Niu, M. Zheng, J.J. Erickson, K.L. Boyd, K.J. McAfee, C. Oseroff, S.R. Hadrup, J.R. Bennink, W. Hildebrand, K.M. Edwards, J.E. Crowe Jr, J.V. Williams, S. Buus, A. Sette, T.N. Schumacher, A.J. Link, S. Joyce, Discovering naturally processed antigenic determinants that confer protective T cell immunity, J. Clin. Invest. 123 (5) (2013) 1976–1987, http://dx.doi.org/10.1172/JCI67388 (Epub 2013).

[15] J. Habchi, P. Tompa, S. Longhi, V.N. Uversky, Introducing protein intrinsic disorder, Chem. Rev. 114 (13) (2014) 6561–6588, http://dx.doi.org/10.1021/cr400514h (Epub 2014 April 17).

[16] I. Hoof, B. Peters, J. Sidney, L.E. Pedersen, A. Sette, O. Lund, M. Nielsen, NetMHCpan, a method for MHC class I binding prediction beyond humans, Immunogenetics 61 (1) (2009) 1–13, http://dx.doi.org/10.1007/s00251-008-0341-z.

[17] T.P. Hopp, K.R. Woods, A computer program for predicting protein antigenic determinants, Mol. Immunol. 20 (4) (1983) 483–489.

[18] A.-S. Huart, T. Hupp, Evolution of conformational disorder & diversity of the P53 interactome, BioDiscovery (8) (2013) 5, http://dx.doi.org/10.7750/BioDiscovery.2013.8.5.

[19] A.L. Hughes, M.K. Hughes, Self peptides bound by HLA class I molecules are derived from highly conserved regions of a set of evolutionarily conserved proteins, Immunogenetics 41 (1995) 257–262.

[20] J.H. Huang, H.L. Xie, J. Yan, H.M. Lu, Q.S. Xu, Y.Z. Liang, Using random forest to classify T-cell epitopes based on amino acid properties and molecular features, Biochimie 103 (2014) 1–6, http://dx.doi.org/10.1016/j.biochi.2014.03.016 (Epub 2014 April 8).

[21] D. Ito, A. Albers, Y.X. Zhao, C. Visus, E. Appella, T.L. Whiteside, A.B. DeLeo, The wild-type sequence (wt) p5325-35 peptide induces HLA-DR7 and HLA-DR11-restricted CD4+ Th cells capable of enhancing the ex vivo expansion and

function of anti-wt p53264–272 peptide CD8+ T cells, J. Immunol. 177 (10) (2006) 6795–6803, http://dx.doi.org/10.4049/jimmunol.177.10.6795.

[22] D.T. Jones, D. Cozzetto, DISOPRED3: precise disordered region predictions with annotated protein-binding activity, Bioinformatics 31 (6) (2015) 857–863, http://dx.doi.org/10.1093/bioinformatics/btu744.

[23] E. Karosiene, C. Lundegaard, O. Lund, M. Nielsen, NetMHCcons: a consensus method for the major histocompatibility complex class I predictions, Immunogenetics 64 (3) (2012) 177–186, http://dx.doi.org/10.1007/s00251-011-0579-8.

[24] Y. Kim, J. Sidney, C. Pinilla, A. Sette, B. Peters, Derivation of an amino acid similarity matrix for peptide: MHC binding and its application as a Bayesian prior, BMC Bioinf. 10 (2009) 394, http://dx.doi.org/10.1186/1471-2105-10-394.

[25] L.P. Kozlowski, J.M. Bujnicki, MetaDisorder: a meta-server for the prediction of intrinsic disorder in proteins, BMC Bioinf. 13 (2012) 111, http://dx.doi.org/10.1186/1471-2105-13-111.

[26] J. Kyte, R.F. Doolittle, A simple method for displaying the hydropathic character of a protein, J. Mol. Biol. 157 (1) (1982) 105–132.

[27] S.J. Landry, Helper T-cell epitope immunodominance associated with structurally stable segments of hen egg lysozyme and HIV gp120, J. Theor. Biol. 203 (3) (2000) 189–201, http://dx.doi.org/10.1006/jtbi.1999.1056.

[28] N. Leffers, M.J. Gooden, R.A. de Jong, B.N. Hoogeboom, K.A. ten Hoor, H. Hollema, H.W. Nijman, Prognostic significance of tumor-infiltrating T-lymphocytes in primary and metastatic lesions of advanced stage ovarian cancer, Cancer Immunol. Immunother. 58 (3) (2009) 449–459, http://dx.doi.org/10.1007/s00262-008-0583-5.

[29] H.H. Lin, S. Ray, S. Tongchusak, E.L. Reinherz, V. Brusic, Evaluation of MHC class I peptide binding prediction servers: applications for vaccine research, BMC Immunol. 9 (2008) 8, http://dx.doi.org/10.1186/1471-2172-9-8.

[30] H.H. Lin, G.L. Zhang, S. Tongchusak, E.L. Reinherz, V. Brusic, Evaluation of MHC-II peptide binding prediction servers: applications for vaccine research, BMC Bioinf. 9 (suppl. 12) (2008) S22, http://dx.doi.org/10.1186/1471-2105-9-S12-S22.

[31] P. Lieutaud, B. Canard, S. Longhi, MeDor: a metaserver for predicting protein disorder, BMC Genomics 9 (suppl. 2) (2008) S25, http://dx.doi.org/10.1186/1471-2164-9-S2-S25.

[32] M.Y. Lobanov, O.V. Galzitskaya, The Ising model for prediction of disordered residues from protein sequence alone, Phys. Biol. 8 (3) (2011) 035004, http://dx.doi.org/10.1088/1478-3975/8/3/035004.

[33] M. Lucchiari-Hartz, V. Lindo, N. Hitziger, S. Gaedicke, L. Saveanu, P.M. van Endert, G. Niedermann, Differential proteasomal processing of hydrophobic and hydrophilic protein regions: contribution to cytotoxic T lymphocyte epitope clustering in HIV-1-Nef, Proc. Natl. Acad. Sci. USA 100 (13) (2003) 7755–7760, http://dx.doi.org/10.1073/pnas.1232228100.

[34] O. Lund, M. Nielsen, C. Kesmir, A.G. Petersen, C. Lundegaard, P. Worning, S. Brunak, Definition of supertypes for HLA molecules using clustering of specificity matrices, Immunogenetics 55 (12) (2004) 797–810, http://dx.doi.org/10.1007/s00251-004-0647-4.

[35] C. Lundegaard, O. Lund, S. Buus, M. Nielsen, Major histocompatibility complex class I binding predictions as a tool in epitope discovery, Immunology 130 (3) (2010) 309–318, http://dx.doi.org/10.1111/j.1365-2567.2010.03300.x.

[36] Y. Kim, J. Sidney, S. Buus, A. Sette, M. Nielsen, B. Peters, Dataset size and composition impact the reliability of performance benchmarks for peptide-MHC binding predictions 15:241, 2014 <http://www.biomedcentral.com/1471-2105/15/241>.

[37] C^1. Mayers, M. Duffield, S. Rowe, J. Miller, B. Lingard, S. Hayward, R.W. Titball, Analysis of known bacterial protein vaccine antigens reveals biased physical properties and amino acid composition, Comp. Funct. Genomics 4 (5) (2003) 468–478, http://dx.doi.org/10.1002/cfg.31.

[38] S.J. Melton, S.J. Landry, Three dimensional structure directs T-cell epitope dominance associated with allergy, Clin. Mol. Allergy 6 (2008) 9, http://dx.doi.org/10.1186/1476-7961-6-9.

[39] B. Meszaros, I. Simon, Z. Dosztanyi, Prediction of protein binding regions in disordered proteins, PLoS. Comput. Biol. 5 (5) (2009) e1000376, http://dx.doi.org/10.1371/journal.pcbi.1000376.

[40] B. Mészáros, Z. Dosztányi, C. Magyar, I. Simon, Bioinformatical approaches to unstructured/disordered proteins and their interactions, in: A. Liwo (Ed.): Computational Methods to Study the Structural & Dynamics of Biomolecules, SSBN 1, Springer-Verlag, Berlin Heidelberg, 2014, pp. 525–556, doi:http://dx.doi.org/10.1007/978-3-642-28554-7_13_c.

[41] B. Meszaros, Z. Dosztanyi, I. Simon, Disordered binding regions and linear motifs–bridging the gap between two models of molecular recognition, PLoS One 7 (10) (2012) e46829, http://dx.doi.org/10.1371/journal.pone.0046829.

[42] N.S. Mitić, M.D. Pavlović, D.R. Jandrlić, Epitope distribution in ordered and disordered protein regions – part A. T-cell epitope frequency, affinity and hydropathy, J. Immunol. Meth. 406 (2014) 83–103, http://dx.doi.org/10.1016/j.jim.2014.02.012.

[43] M.J. Mizianty, W. Stach, K. Chen, K.D. Kedarisetti, F.M. Disfani, L. Kurgan, Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources, Bioinformatics 26 (18) (2010) i489–i496, http://dx.doi.org/10.1093/bioinformatics/btq373.

[44] M. Nielsen, C. Lundegaard, P. Worning, S.L. Lauemoller, K. Lamberth, S. Buus, O. Lund, Reliable prediction of T-cell epitopes using neural networks with novel sequence representations, Protein Sci. 12 (5) (2003) 1007–1017, http://dx.doi.org/10.1110/ps.0239403.

[45] M. Nielsen, S. Justesen, O. Lund, C. Lundegaard, S. Buus, NetMHCIIpan-2.0 – improved pan-specific HLA-DR predictions using a novel concurrent alignment and weight optimization training procedure, Immunome Res. 6 (2010) 9, http://dx.doi.org/10.1186/1745-7580-6-9.

[46] M. Nielsen, O. Lund, NN-align. An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction, BMC Bioinf. 10 (2009) 296, http://dx.doi.org/10.1186/1471-2105-10-296.

[47] M. Nielsen, C. Lundegaard, T. Blicher, K. Lamberth, M. Harndahl, S. Justesen, S. Buus, NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence, PLoS One 2 (8) (2007) e796, http://dx.doi.org/10.1371/journal.pone.0000796.

[48] H.W. Nijman, A. Lambeck, S.H. van der Burg, A.G. van der Zee, T. Daemen, Immunologic aspect of ovarian cancer and p53 as tumor antigen, J. Transl. Med. 3 (2005) 34, http://dx.doi.org/10.1186/1479-5876-3-34.

[49] H.W. Nijman, S.H. Van der Burg, M.P.M. Vicrboom, J.G.A. Houbiers, W. Martin Kast, C.J.M. Melief, P53, a potential target for tumor-directed T cells, Immunol. Lett. 40 (2) (1994) 171–178, http://dx.doi.org/10.1016/0165-2478(94)90189-9.

[50] C.J. Oldfield, Y. Cheng, M.S. Cortese, P. Romero, V.N. Uversky, A.K. Dunker, Coupled folding and binding with alpha-helix-forming molecular recognition elements, Biochemistry 44 (37) (2005) 12454–12470, http://dx.doi.org/10.1021/bi050736e.

[51] C.J. Oldfield, J. Meng, J.Y. Yang, M.Q. Yang, V.N. Uversky, A.K. Dunker, Flexible nets: disorder and induced fit in the associations of p53 and 14–3-3 with their partners, BMC Genomics 9 (suppl. 1) (2008) S1, http://dx.doi.org/10.1186/1471-2164-9-S1-S1.

[52] M.D. Pavlović, D.R. Jandrlić, N.S. Mitić, Epitope distribution in ordered and disordered protein regions. Part B – ordered regions and disordered binding sites are targets of T- and B-cell immunity, J. Immunol. Meth. 407 (2014) 90–107, http://dx.doi.org/10.1016/j.jim.2014.03.027.

[53] B. Peters, A. Sette, Generating quantitative models describing the sequence specificity of biological processes with the stabilized matrix method, BMC Bioinf. 6 (2005) 132, http://dx.doi.org/10.1186/1471-2105-6-132.

[54] P. Radivojac, L.M. Iakoucheva, C.J. Oldfield, Z. Obradovic, V.N. Uversky, A.K. Dunker, Intrinsic disorder and functional proteomics, Biophys. J. 92 (5) (2007) 1439–1456, http://dx.doi.org/10.1529/biophysj.106.094045.

[55] J. Sidney, B. Peters, N. Frahm, C. Brander, A. Sette, HLA class I supertypes: a revised and updated classification, BMC Immunol. 9 (2008) 1, http://dx.doi.org/10.1186/1471-2172-9-1.

[56] R.E. Soria-Guerra, R. Nieto-Gomez, D.O. Govea-Alonso, S. Rosales-Mendoza, An overview of bioinformatics tools for epitope prediction: implications on vaccine development, J. Biomed. Inform. 53C (February) (2015) 405–414, http://dx.doi.org/10.1016/j.jbi.2014.11.003 (Epub 2014 November 10).

[57] T. Sturniolo et al., Generation of tissue-specific and promiscuous HLA ligand database using DNA microarrays and virtual HLA class II matrices, Nat. Biotechnol. 17 (1999) 555–561.

[58] M.J. Suskiewicz, J.L. Sussman, I. Silman, Y. Shaul, Context-dependent resistance to proteolysis of intrinsically disordered proteins, Protein Sci. 20 (8) (2011) 1285–1297, http://dx.doi.org/10.1002/pro.657.

[59] J.C. Tong et al., Methods and protocols for prediction of immunogenic epitopes, Brief. Bioinform. 8 (2) (2006) 96–108.

[60] V.N. Uversky, C.J. Oldfield, A.K. Dunker, Intrinsically disordered proteins in human diseases: introducing the D2 concept, Ann. Rev. Biophys. 37 (2008) 215–246.

[61] V.N. Uversky, A.K. Dunker, Understanding protein non-folding, Biochim. Biophys. Acta 1804 (6) (2010) 1231–1264, http://dx.doi.org/10.1016/j.bbapap.2010.01.017.

[62] V.N. Uversky, Intrinsically disordered proteins may escape unwanted interactions via functional misfolding, Biochim. Biophys. Acta 1814 (5) (2011) 693–712, http://dx.doi.org/10.1016/j.bbapap.2011.03.010.

[63] K. Yusim, C. Kesmir, B. Gaschen, M.M. Addo, M. Altfeld, S. Brunak, A. Chigaev, V. Detours, B.T. Korber, Clustering patterns of cytotoxic T-lymphocyte epitopes in human immunodeficiency virus type 1 (HIV-1) proteins reveal imprints of immune evasion on HIV-1 global variation, J. Virol. 76 (17) (2002) 8757–8768.

[64] H. Zhang, C. Lundegaard, M. Nielsen, Pan-specific MHC class I predictors: a benchmark of HLA class I pan-specific prediction methods, Bioinformatics 25 (1) (2009) 83–89, http://dx.doi.org/10.1093/bioinformatics/btn579.

[65] J.M. Weaver, A.J. Sant, Understanding the focused CD4 T cell response to antigen and pathogenic organisms, Immunol. Res. 45 (2–3) (2009) 123–143, http://dx.doi.org/10.1007/s12026-009-8095-8.