

University of Dundee

The Dundee Resource for Sequence Analysis and Structure Prediction

MacGowan, Stuart; Madeira, Fabio; Britto-Borges, Thiago ; Warowny, Mateusz ; Drozdetskiy, Alexey; Procter, James

Published in:
Protein Science

DOI:
[10.1002/pro.3783](https://doi.org/10.1002/pro.3783)

Publication date:
2019

Document Version
Peer reviewed version

[Link to publication in Discovery Research Portal](#)

Citation for published version (APA):

MacGowan, S., Madeira, F., Britto-Borges, T., Warowny, M., Drozdetskiy, A., Procter, J., & Barton, G. (2019). The Dundee Resource for Sequence Analysis and Structure Prediction. *Protein Science*.
<https://doi.org/10.1002/pro.3783>

General rights

Copyright and moral rights for the publications made accessible in Discovery Research Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from Discovery Research Portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain.
- You may freely distribute the URL identifying the publication in the public portal.

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

The Dundee Resource for Sequence Analysis and Structure Prediction

Stuart A. MacGowan¹, Fábio Madeira^{1,2}, Thiago Britto-Borges^{1,3}, Mateusz Warowny¹, Alexey Drozdetskiy^{1,4}, James B. Procter¹ and Geoffrey J. Barton^{1,*}

¹Division of Computational Biology, College of Life Sciences, University of Dundee, UK. Present addresses²⁻⁴: ²European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Cambridge, UK. ³Section of Bioinformatics and Systems Cardiology, Department of Internal Medicine III and Klaus Tschira Institute for Integrative Computational Cardiology, University of Heidelberg, Germany. ⁴Altius UK, London, UK. *Correspondence: g.j.barton@dundee.ac.uk.

Abstract

The Dundee Resource for Sequence Analysis and Structure Prediction (DRSASP; <http://www.compbio.dundee.ac.uk/drsasp.html>) is a collection of web services provided by the Barton Group at the University of Dundee. DRSASP's flagship services are the JPred4 webserver for secondary structure and solvent accessibility prediction and the JABAWS 2.2 webserver for multiple sequence alignment, disorder prediction, amino acid conservation calculations and specificity determining site prediction. DRSASP resources are available through conventional web interfaces and APIs but are also integrated into the Jalview sequence analysis workbench which enables the composition of multitool interactive workflows. Other existing Barton Group tools are being brought under the banner of DRSASP, including NoD (Nucleolar localisation sequence detector) and 14-3-3-Pred. New resources are being developed that enable the analysis of population genetic data in evolutionary and 3D structural contexts. Existing resources are actively developed to exploit new technologies and maintain parity with evolving web standards. DRSASP provides substantial computational resources for public use and since 2016 DRSASP services have completed over 1.5 million jobs.



Figure 1: The Dundee Resource for Sequence Analysis and Structure Prediction.

1 Introduction

The flood of sequence data across all species continues to grow in rate and volume. While there are many challenges in managing these large data sets, the major hurdle is to use the raw sequence data to inform our knowledge and understanding of biological systems. In order to achieve this goal, accurate and reliable software tools are required to make structural and functional predictions from the sequence data. Over 30 years, our group has developed innovative software packages, web servers and databases that allow the structure and function of protein sequences to be probed and has used these in conjunction with experiments to improve understanding of specific biological systems.

The Dundee Resource for Sequence Analysis and Structure Prediction (DRSASP; Figure 1) encapsulates many of these methods with techniques developed by other groups as a collection of publicly available protein sequence analysis web services. The resource provides convenient access through websites, APIs and the Jalview¹ analysis workbench to a variety of algorithms including secondary structure prediction, disorder prediction, multiple sequence alignment, evolutionary conservation calculations and other functional site predictions²⁻¹⁰. DRSASP helps to translate Barton Group research into new web services accessible to a wide community as well as ensuring the sustainability of the popular JPred² and JABAWS¹⁰. Initially, DRSASP comprised JPred3¹¹, JABAWS:MSA¹² and Kinomer⁵. Over the last few years new services have been added such as NoD⁹ and 14-3-3 Pred³ and our main services have undergone significant updates. The sustained contribution and relevance of DRSASP has been recognised in the granting of Elixir-UK Tier 1 Resource status¹³. This signifies Elixir-UK's view that DRSASP is an important contributor in the strategic area of Protein Structure and Function. In this article we summarise the current DRSASP (August 2019) and look forward to new resources that will be added in the near future.

2 The DRSASP Toolbox

Table 1 presents an overview of the DRSASP tools and categorises their application, availability and technology. The tools address a range of general biological questions: What is the structure of the protein? Will the protein crystallise? Which amino-acid residues are conserved across a set of homologues and what type of conservation is present (e.g. identity, hydrophobicity, charge)? Which residues are important for functional specificity? Where does the protein localise in the cell? Are any residues likely to be involved in protein-protein interactions? In terms of technology, 14-3-3-Pred, NoD and the XTal suite are implemented by sequence trained machine learning algorithms; Kinomer is a profile HMM (Hidden Markov model) based method; JPred is a multiple neural network method trained from sequence alignment profiles and AACon and AMAS¹⁴ contain a variety of residue set-based calculations. JABAWS itself is a web services framework with which DRSASP serves a range of sequence analysis methods. Most DRSASP services are accessible via web forms, which are mainly suitable for small-scale analyses. For bulk analyses, some services provide programmatic-APIs and/ or pre-computed datasets. Many services are available directly from Jalview¹ or provide results in Jalview compatible format. In the following sections we provide a concise description of each tool covering what it does, how it works, how it can be applied through research examples and how it is used.

Table 1: Summary of the DRSASP tools

Service ^b	Application	Tech	Availability ^a				Released	Ref
			HTML	API	Jalview ¹	Dataset		
JPred4	Protein 2° structure and solvent accessibility prediction	ANN	✓	✓	✓	✓ ^c	2015	2
JABAWS 2.2	Bioinformatics tool web services framework. Provides: MSA, conservation analysis, disorder prediction and RNA 2° structure prediction	Multiple		✓	✓		2018	10
Slivka ^d	Successor to JABAWS (see above)		(✓)	(✓)	(✓)		-	
ProteoCache ^d	DRSASP data warehouse			(✓)		(✓)	-	
pyDRSASP ^d (ProteoFAV) (ProIntVar) (VarAlign)	Variant and structure analyses			(✓)		(✓)	-	
AACon	Conservation			✓	✓			
14-3-3-Pred	PPIs	ANN SVM PSSM	✓	✓	✓ ^e		2015	3
NoD	Subcellular localisation	ANN	✓			✓	2011	9
Kinomer	Sequence classification	pHMM	✓			✓	2009	5
XTal	Crystallisation propensity		✓			✓ ^f	2008	6
(OB-score)	Construct design	Z-score						8
(XANNPred)		ANN						7
(ParCrys)		Parzen window						
AMAS	Functional residues		✓				1993	14

Columns - "Service": The name of the DRSASP service; "Application": The application area; "Tech": implementation technology – ANN (Artificial Neural Network – machine learning), SVM (Support Vector Machine), pHMM (Profile Hidden Markov Model), PSSM (Position Specific Scoring Matrix). "HTML": Tick means service has a web page form interface; "API": Indicates the service has an Application Programming Interface; "Jalview": shows services that are directly accessible from Jalview. "Dataset": Indicates availability of datasets associated with the method. "Released": First release date of the service. Footnotes - a) Parentheses denote that the availability is in development. b) Parentheses denote subcomponents of a service. c) The JPred training and test datasets are available for download from the website. Access to pre-computed JPred predictions for certain proteomes (e.g. Human) will be available from ProteoCache in future but for now are obtainable from the API or upon request from the authors. d) service is in development. e) 14-3-3-Pred provides Jalview compatible feature files. f) OB-Score predictions are available for Pfam 31.0.

2.1 JABAWS: Java Bioinformatics Analysis Web Services

One of our objectives for DRSASP is to deliver resources via a common interface and to make it easy for others to deploy the same services on their own computing infrastructure. With this in mind we developed the JABAWS^{10,12} framework. JABAWS simplifies the provision of bioinformatics tools as web services by abstracting web interfaces, tool wrapping, wrapper execution and data models. The DRSASP instance of JABAWS provides access to multiple sequence alignment methods, disorder predictors, an RNA secondary structure predictor and methods for conservation calculation from multiple sequence alignments.

For multiple sequence alignment, JABAWS includes Clustal Omega¹⁵, Clustal W¹⁶, Mafft^{17,18}, Muscle¹⁹, T-coffee²⁰, Probcons²¹, MSAProbs²² and GLProbs²³. The availability of these varied multiple sequence alignment programs allows the user to select the best tool for the sequences they wish to align or to compare the results from different algorithms interactively in Jalview or programmatically using the JABAWS client. This approach can also be taken with the multiple options JABAWS provides for residue conservation scoring and disorder prediction. For disorder prediction we have DisEMBL²⁴, IUPred²⁵, Jronn²⁶ and GlobPlot²⁷ and there are examples where users report the results from two or more of these options²⁸. For MSA interpretation, 17 conservation scores and the SMERFS score⁴ for functional site prediction are available through JABAWS, implemented with our AACon software discussed further in §2.3. For RNA secondary structure prediction, JABAWS provides the RNAalifold method from the ViennaRNA package²⁹.

JABAWS allows the specification of command line parameter presets. For example, in addition to the default settings, MUSCLE¹⁹ is configured with separate presets that are suitable for protein alignments and nucleotide alignments whilst MAFFT¹⁸ presets are configured to implement the NW-NS-PartTree-1, FFT-NS-i, FFT-NS-1, L-INS-i, E-INS-i and G-INS-i strategies. For maximum flexibility, command-line options are exposed via the JABAWS interface allowing users to run tools with options suitable for their own needs.

Most Jalview¹ users will access the Dundee JABAWS instance as this is pre-configured by default Jalview installations. This makes JABAWS functions accessible immediately after installing Jalview. If a user prefers to keep their data local, work without access to the internet or tackle very large problems, they may wish to install JABAWS on their personal computer or site-wide computing resource at their institution. The simplest way to create a JABAWS instance is with the JABAWS virtual appliance or Docker container (see <http://www.compbio.dundee.ac.uk/jabaws22/archive/docker/Dockerfile>) but a WAR file (Web Application Archive) is provided that is better suited for institutional installations. Jalview can be configured to use the alternative JABAWS instance via *Tools* → *Preferences* → *Web Services*. JABAWS services can also be accessed programmatically via a downloadable command line client. Alternatively, users may interface with the JABAWS SOAP API with their own preferred SOAP client. These modes are best suited to users who wish to use JABAWS service for high-throughput analyses or as part of computational pipelines. The public JABAWS service at www.compbio.dundee.ac.uk/JABAWS/ currently has no fair usage policies imposed, but public jobs are restricted to defined maxima for the number of submitted sequences and average sequence length. These restrictions are applied on a tool/preset specific basis and are obtained via SOAP operations, for example with the -limits argument to the JABAWS command line client. Limits vary from 500 – 2,000 sequences for sequence alignment, 2,000 – 5,000 sequences for disorder prediction and 2,000 – 10,000 sequences for disorder calculations. Additionally, all jobs are limited to one hour of compute time. Jobs of larger than the relevant size limits will not be accepted and long running jobs are terminated.

Figure 2 illustrates how to run MAFFT¹⁸ on an alignment using the L-INS-i presets in Jalview. Jalview has a sophisticated yet intuitive interface to JABAWS. Jalview permits custom tool parameters, alignment or realignment of alignment subsets and automatically displays

results from JABAWS appropriately. In this example, the result is a new MSA and is displayed in a new alignment window. The JABAWS protein disorder or conservation tools create annotation tracks on the alignment on which they are run. Jalview also allows custom parameters to be set for a JABAWS tool via a dialog accessed under the appropriate *Web Service* sub-menus.

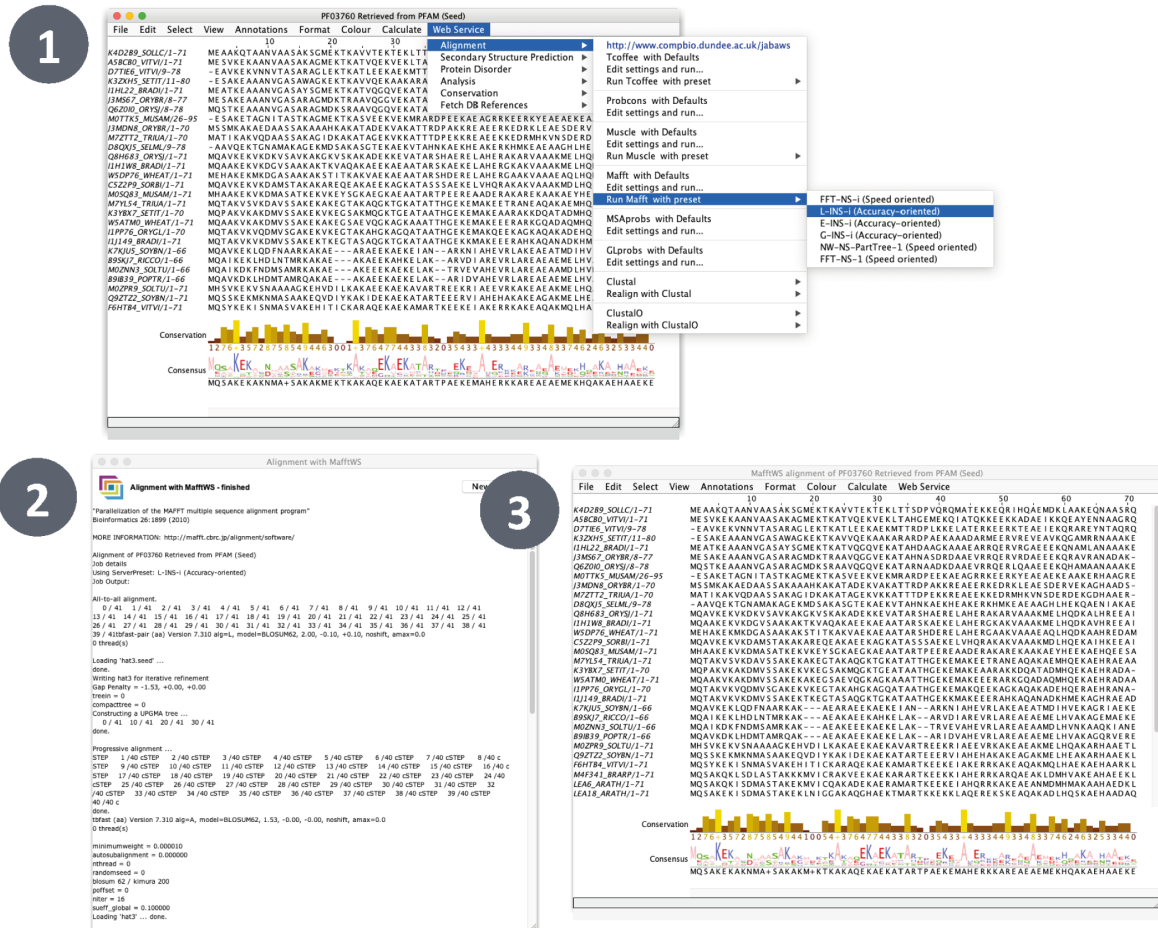


Figure 2: Running MAFFT¹⁸ L-INS-i alignment with Jalview's¹ default JABAWS¹⁰ configuration. 1) Web Service → Alignment → Run MAFFT with preset → L-INS-i. If custom parameters are desired they can be set in the dialog available through "Edit settings and run..." 2) A new window reports the job arguments and its progress. 3) The resulting alignment opens in a new window (NB. the results MSA can be reopened with the "New Window" in the progress window).

We have found the convenience of JABAWS beneficial in our own research. An analysis of all four disorder predictions in JABAWS in a set of known O-linked β -N-acetylglucosamine transferase (OGT; 620 proteins) compared to a negative control set (1,164 proteins) showed that disorder was likely to be an element of OGT substrate recognition, despite the absence of clear sequence motifs³⁰. High-throughput disorder predictions were tried as features in the prediction of 14-3-3 protein binding-sites (see §2.4)³. JABAWS also simplified the calculation of conservation scores for several thousand Pfam³¹ alignments³².

2.2 JPred4: A Protein Secondary Structure Prediction Server

The JPred4² web server predicts secondary structure and solvent accessibility for a given protein sequence or multiple sequence alignment with the JNet 2.3.1 algorithm. A predicted protein secondary structure is useful in many ways when experimentally determined structures are unavailable. For example, secondary structure predictions can be used to

improve multiple sequence alignments, as a starting point for 3D structure prediction or to interpret patterns of conservation in an alignment.

Statistical and machine learning based approaches have proven effective at predicting protein secondary structure from sequence³³⁻³⁵. JNet 2.3.1 has a secondary structure prediction three-state accuracy (Q_3 ; α -helix, β -strand and coil) of 82.0 %², which was as good as the PSIPRED³⁶ and PredictProtein³⁷ self-reported blind test accuracies at the time of development. Since then, Xu and co-workers³⁸ reported Q_3 accuracies for JPred of 80-83 % across a series of five other test datasets, values which were comparable to the other algorithms they tested and only slightly below the authors' DeepCNF-SS program (82-85 % across the five datasets)³⁸. JPred4 solvent accessibility predictions are 90.0 %, 83.6 % and 78.1 % accurate for buried, part-exposed and surface residues, respectively².

JPred4 can make predictions for a single sequence, a batch of single sequences or a pre-computed multiple sequence alignment. The sequence pipeline begins by searching the PDB for homologues and will advise the user of any matches that are found since if the 3D-structure of a homologue is known, this provides a strong guide to the secondary and tertiary structure of the protein and secondary structure prediction is less useful. The sequence is then checked against the DRSASP ProteoCache (see §3.2) and if found, the full JPred results are retrieved from the datastore within a few seconds. The sequence is queried against Uniref90 with PSI-BLAST and a non-redundant multiple sequence alignment is constructed from the matches. From here, JPred generates a profile HMM with HMMER and passes this profile HMM and the PSSM from PSI-BLAST to JNet and the Lupas coiled-coil predictor³⁹. In the MSA pipeline, the profile HMM and PSSM are generated directly from the user supplied MSA and these are fed to JNet without any PSI-BLAST search. Figure 3 illustrates JPred results visualised with Jalview and UCSF Chimera. The JPred predicted secondary structure is shown in Jalview as an annotation track where green indicates strand and red indicates predicted helical regions. This colouring is then transferred to the mapped PDB structure 3axm⁴⁰ through the Jalview-UCSF Chimera interface to illustrate the accuracy of the prediction. JPred4 returns results in several formats: graphically by generating an SVG with Jalview; HTML formatted alignment with prediction tracks; PDF generated with Alscript⁴¹ and in Jalview¹ via a JVL file (Jalview Launch file; requires Jalview ≥ 2.11 installed locally).

JPred4 can be accessed in multiple ways. The website provides a convenient interface to allow users to make secondary structure predictions for a single sequence, a batch of sequences or for a user-provided MSA. JPred4 predictions for a sequence or MSA can also be obtained from directly within Jalview. Alternatively, JPred4 can be accessed programmatically via its REST API and a Perl command-line client is available as the recommended interface. This allows users to submit, monitor and retrieve JPred4 predictions *en masse* or as part of computational pipelines. The API client is a suitable means to obtain whole proteome scale JPred prediction sets without overloading the JPred4 server.

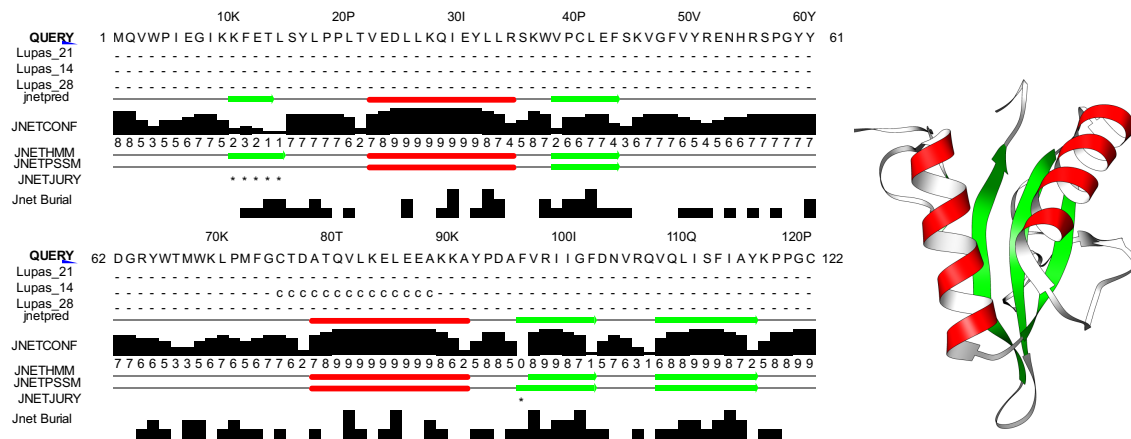


Figure 3: Illustration of a JPred⁴² secondary structure prediction displayed in Jalview¹ (left) and UCSF Chimera⁴² (right). Below the query sequence, JPred provides several annotation tracks for visualisation in Jalview. These are the Lupas³⁹ Coil predictions with varying window sizes ("-" = no coil; "c" = likely coil; "C" = coil); the final JNet prediction (red, helix; green, strand) followed by a confidence score for the prediction (0-9; least to highest confidence). These are followed by separate predictions where JNet is given only the profile HMM or PSSM and the JNETJURY track that indicates positions where these predictions differ (indicated by '*'). Finally, burial predictions are represented by a histogram of values ranging 0-3, representing no burial and burial at 25%, 5% and 0% thresholds, respectively. The query sequence and structure illustration are derived from PDB ID: 3AXM⁴⁰.

A good way to understand JPred's relevance is to see how others have applied JPred predictions to address problems. JPred can be applied in analyses involving a few proteins, whole proteomes or other large sets of proteins or as part of new computational pipelines. An example of the application of JPred to guide experimental work is the identification of the paired amphipathic helix protein Sin3a interaction domain in the methylcytosine dioxygenases TET1 and TET3⁴³. The authors identified a common helical region in TET1 and TET3 outside of the known oxygenase and Zinc finger domains that was absent in TET2. The putative TET1-Sin3A interaction helix was confirmed experimentally with co-immunoprecipitation, site-directed mutagenesis and NMR. JPred predictions were also used to assist the Cryo-EM structure determination of the DNA-bound PolD complex⁴⁴. High-throughput applications of JPred include structurally rationalising the distribution of aspirin mediated lysine acetylations in the human proteome⁴⁵; determining the factors affecting heterologous protein solubility⁴⁶ and identifying kinases with a helix present in their activation loop across the human kinome⁴⁷. Lastly, JPred is an essential part of the QuanTest⁴⁸ method for MSA benchmarking that compares MSAs containing sequences of known structure by assessing the accuracy of the JPred secondary structure predictions made from them.

2.3 ACon

ACon is a Java implementation of 18 methods of scoring amino acid residue conservation in multiple sequence alignments. The majority of the methods are described in Valdar's 2002 review⁴⁹ with additional algorithms that were developed in the Barton group. The methods include the symbol frequency based Shenkin score⁵⁰, the physicochemical property based Zvelebil score⁵¹, the redundancy aware Valdar score⁵² and the specificity sensitive SMERFS score⁴. These examples illustrate how different scoring algorithms consider residue conservation as characterised by different features of the alignment. This point is demonstrated in a real-world example in Figure 4, which compares five different conservation scores for an excerpt of the Pfam³¹ WD40 repeat family MSA. In this example,

the scores do not all concur on what positions are most conserved in this alignment. Jalview's physicochemical conservation score highlights the consensus Asp and Val/Ile as the two most physicochemically conserved in contrast with the consensus, Valdar and Shenkin scores that all include the His and Trp consensus positions amongst the most conserved. Indeed, even the physicochemical-based Zvelebil score identifies very different positions as the most conserved due to different treatments of gaps and aberrant or atypical residues.

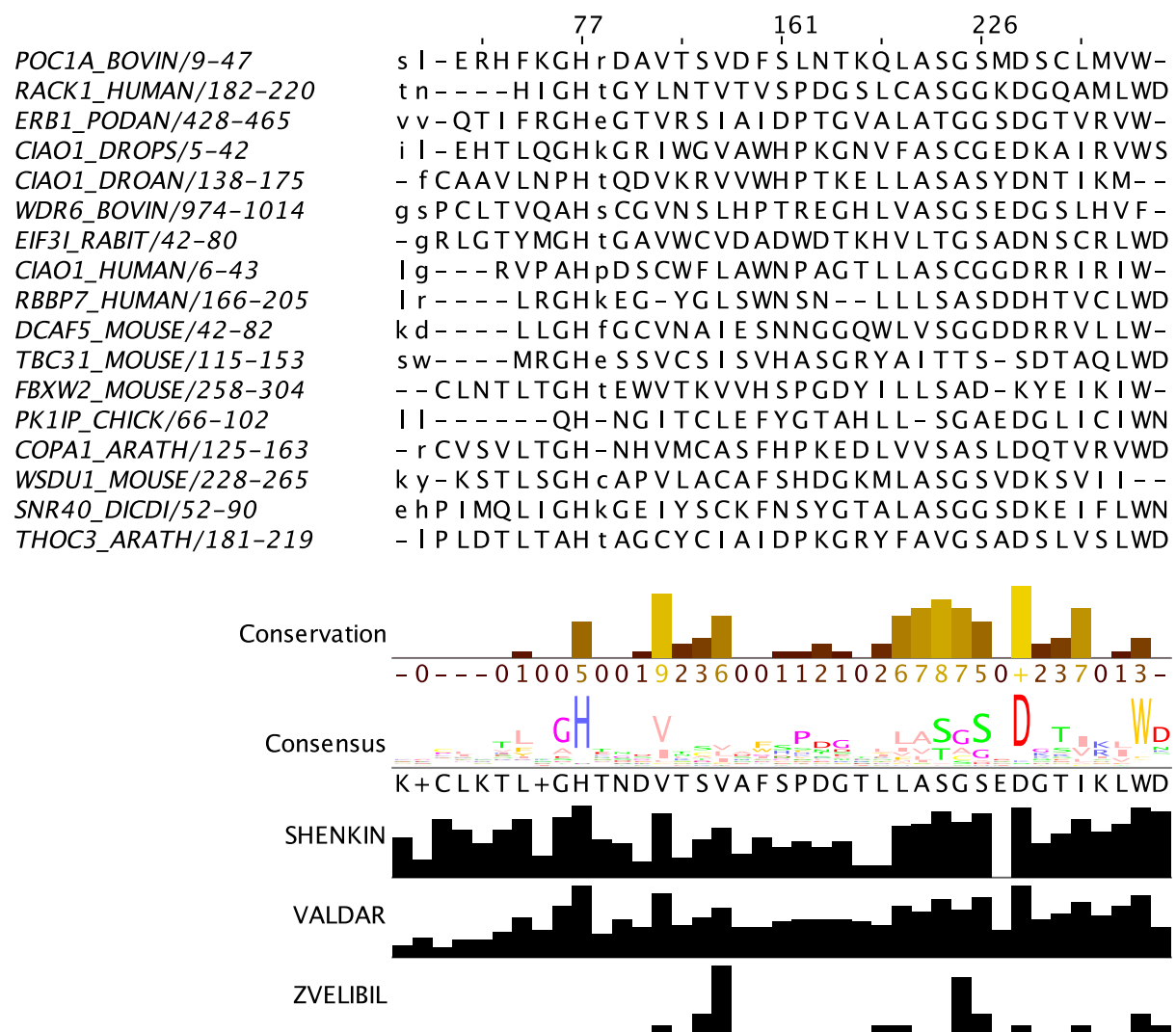


Figure 4: Comparison of evolutionary conservation scores. An excerpt of the Pfam³¹ WD40 repeat family (PF00400) is displayed together with Jalview¹ annotation tracks representing five different conservation metrics (the scores were calculated for the first 89 SwissProt sequences in this Pfam, only the first 17 are shown). The Conservation and Consensus tracks are calculated by Jalview whilst the Valdar, Shenkin and Zvelebil tracks are calculated with AACon via JABAWS called from the Jalview webservices menu.

AACon is accessible via the JABAWS¹⁰ web service, which as described §2.1, is available via Jalview or the JABAWS CLI client. AACon is also available as an executable JAR file, Java library or its own web service. Users interested in analysing conservation in only a few MSAs will probably find the Jalview-JABAWS interface sufficient for their needs. Studies that require high throughput conservation calculations or where a numerical comparison of different conservation scores is desired will best be served by either JABAWS Client or

AACon executable. In this case, the user should determine whether remote execution would be advantageous and check if their alignments are within the Dundee JABAWS service sequence limits. The precise limits vary depending on what conservation scores are requested but range between 2,000 – 10,000 sequences of average length 1,000 – 10,000 residues for the most intensive to least intensive requests; the precise limits can be queried with the JABAWS client. If these conditions are met then the JABAWS Client is suitable, otherwise it is recommended to use the AACon executable locally (<https://github.com/bartongroup/aacon>).

2.4 14-3-3-Pred

14-3-3-Pred³ is a webserver that predicts 14-3-3-binding sites. 14-3-3 proteins regulate a variety of cellular processes by binding pairs of phosphorylated Ser/Thr residues on its target substrates⁵³. 14-3-3-Pred combines predictions from PSSM, SVM and ANN models, which were trained on a gold standard set of 14-3-3 binding sites created by a modest extension of the ANIA⁵⁴ database and curated negative set, into a consensus predictor. Recent applications of 14-3-3-Pred include a screen of 106 putative substrates in tomato⁵⁵; the localisation of the 14-3-3 target residues in the Nuclear receptor subfamily 1 group I member 2 protein⁵⁶ and a target residue in the inactive tyrosine-protein kinase transmembrane receptor ROR1⁵⁷.

Figure 5 displays the 14-3-3-Pred web interface where proteins of interest can be queried using single UniProt accession identifiers or as sequences in FASTA format. The results page displays a table with the site scores as well as information on the phosphorylation state of the respective Ser/Thr for each queried protein. Alternatively, a file containing up to 100 protein sequences in FASTA format can be uploaded. 14-3-3-Pred then generates comma/tab-separated output results files that can be used to compare predictions, elaborate hypotheses, and prioritise laboratory experiments to investigate the predicted sites. Results can be accessed using single UniProt IDs ('pid=<identifier>') and specifying the output format ('out=<format>') as either JSON, CSV or TSV. An example query is <http://www.compbio.dundee.ac.uk/1433pred/pid=O96013&out=json>

Welcome to 14-3-3-Pred: A webserver to predict 14-3-3-binding sites in proteins

Use the input fields below to get candidate 14-3-3-binding sites.

1a

Input a single UniProtKB ID
 Example: Try UniProt ID [O96013](#), [Q96E09](#) or [Q7Z](#)

1b

```
>sp|Q16613|SNAT_HUMAN Serotonin N-acetyltransferase OS=Homo sapiens
OX=9606 GN=AANAT PE=1 SV=1
MSTQSTHPLKPEAPRLPPGIPESPSCQRRHTLPASEFRCLTP
```

Example: Load the FASTA sequence for UniProt ID [Q16613](#)

1c

Example: Upload multiple Protein Sequences in FASTA format

2

3

Input FASTA sequence

```
>sp|Q16613|SNAT_HUMAN Serotonin N-acetyltransferase OS=Homo sapiens
OX=9606 GN=AANAT PE=1 SV=1
MSTQSTHPLKPEAPRLPPGIPESPSCQRRHTLPASEFRCLTP
```

4

Candidate 14-3-3-binding sites

Position	Peptide [-6:4]	ANN	PSSM	SVM	Consensus	phosphoS/T
3	---MS[T]QSTH	0.050	-0.643	-1.782	-0.792	-
5	--MSTQ[S]THPL	0.110	-0.158	-1.001	-0.350	-
6	-MSTQS[T]HPLK	0.302	0.305	-0.589	0.006	-
23	PPGIPE[S]PSCQ	0.022	-0.554	-2.065	-0.866	-
25	GIPESP[S]CQRR	0.321	0.185	-0.611	-0.035	-
31	SCQRRH[T]LPAS	0.968	1.819	1.826	1.538	Yes
35	RHTLPA[S]EFRCL	0.192	-0.108	-1.228	-0.381	-
41	SEFRCL[T]P---	0.295	0.618	-0.533	0.127	-

Legend
 ANN - Artificial Neural Network (cut-off = 0.55)
 PSSM - Position-Specific Scoring Matrix (cut-off = 0.80)
 SVM - Support Vector Machine (cut-off = 0.25)
 Consensus - Average of the scores provided by the three methods (cut-off = 0.50)
 pSer/Thr sites are obtained from UniProt and PhosphoSitePlus

Highly scored by three methods
 Highly scored by two methods only
 Highly scored by one method only

5

Results overview with Jalview annotated sequence (SVG)

Q16613/1-42 MSTQSTHPLKPEAPRLPPGIPESPSCQRRHTLPASEFRCLTP

6

Protein sequence

Jalview annotation

Jalview SVG

Figure 5: 14-3-3-Pred³ submission page (back). The website presents a form where you can enter either a UniProt accession (1a), a FASTA sequence (1b) or upload a set of sequences in a FASTA file (1c). The prediction is started by clicking "Submit" (2). 14-3-3-Pred results page (front). The results indicate the query sequence with S/T sites highlighted (3); a table showing the query motifs, the prediction scores and whether the site is known to be phosphorylated (4); a sequence view of the predictions (5) and download links including Jalview feature file format (6).

Figure 6 illustrates the results of a 14-3-3-Pred analysis on sheep serotonin N-acetyltransferase. The prediction was run via the webserver and the results downloaded as Jalview features. These were then loaded into Jalview and Jalview's PDB lookup identified the structure 1ib1⁵⁸ and this was opened in UCSF Chimera via Jalview. Out of 22 Ser/Thr sites, 14-3-3 Pred correctly identifies pThr 31 as a 14-3-3 binding-site with high-confidence (i.e., all method concordance) whilst Ser 118 is falsely predicted to be a 14-3-3 binding-site

albeit with low-confidence. A third high-confidence positive prediction is found for pSer 205, which is not resolved in this structure.

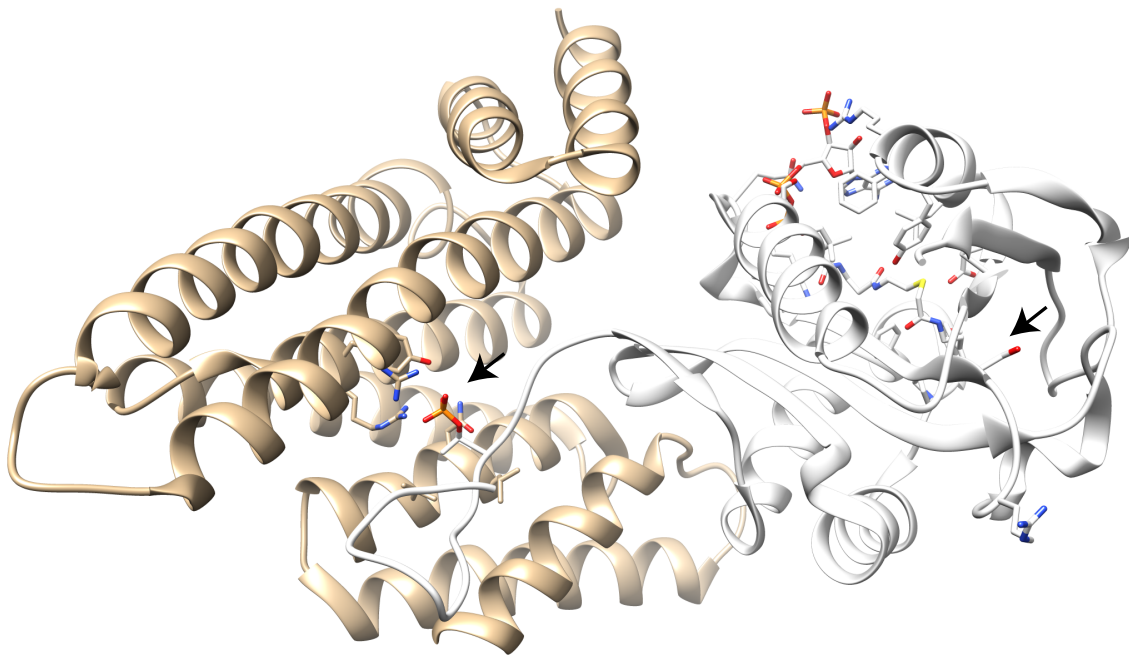


Figure 6: Illustration of Serotonin N-acetyltransferase (right; white) in complex with 14-3-3 zeta (left; tan) showing the interaction of pThr31 with 14-3-3 zeta. The 14-3-3-Pred predicted 14-3-3 targets pThr 31 and Ser 118 in Serotonin N-acetyltransferase are indicated with black arrows. Figure adapted from PDB ID: 1ib1⁵⁸ chains A and E, with UCSF Chimera and Jalview.

2.5 NoD

NoD^{9,59} is a predictor of nucleolar localization sequences (NoLSs) in proteins. NoLSs are short basic motifs that localize proteins to the nucleolus. The NoD algorithm is an artificial neural network (ANN) that was trained using 3-fold cross-validation on 46 experimentally validated NoLSs and negative sequences representing non-NoLS nuclear localisation sequences and randomly selected non-nucleolar cytoplasmic and nucleoplasmic sequences. NoD predictions were computed for the human proteome and 10 of the top scoring NoLSs were experimentally confirmed⁵⁹.

Figure 7 illustrates the NoD submission and results pages. You can search the set of NoLSs predicted in 9,531 human proteins out of the 43,534 human proteins considered from IPI⁶⁰ (version 3.40). NoLS predictions for an arbitrary protein sequence in FASTA format can be obtained via the text input box. If possible, full-length protein sequences should be used to obtain maximum prediction accuracy. Optionally, users can decide to include JPred3¹¹ secondary structure prediction in the prediction of NoLSs. This results in more accurate predictions but requires more computation time (usually around 10 minutes but up to 6 hours is known). Once the protein sequence has been submitted, a waiting page is displayed providing users with a link to the output page. This link can be bookmarked and consulted later. The results page indicates the positions and sequences of any predicted NoLS. A graph of the predictor score along the length of the sequence is also shown. NoD can also be downloaded and run locally, in which case tabular output can be obtained more amenable to high-throughput analyses.

What Does NoD Do ...

NoD is a predictor of nucleolar localization sequences (NoLSs) in proteins. NoLSs are short basic motifs that localize proteins to the [nucleolus](#), a sub-compartment of the nucleus. You can search the set of NoLSs predicted in 9531 human proteins out of the 43534 human proteins considered (from [IP1](#) version 3.40). For more information, visit the [Help Document](#).

1

Accession-based search

Search whether a human protein accession (eg Human protein accession (eg (Currently supported identifier

You can download the complete

2a

Protein Sequence Prediction

Please paste your protein sequence in the text box. Non-standard amino acids are not supported. If you wish to use a command-line version of the program, please see the [Help Document](#).

>NOL12
MGRNKKKKRDGDDRRPRLVLSFDEE
TTISDLDSGARLLGLTPPEGAG

☐ Use Jpred secondary structure prediction but up to 6 hours occasionally

2b

3

6

NoLS predictions for protein NOL12

(these predictions are based only on sequence)

2 NoLSs are predicted in this protein:

MGRNKKKKRDGDDRRPRLVL (between positions 1 and 20)

TASLHAHSRKKVKRKHPRRAQDSKKPPRAPRTSKAQRRLTGKARHSGE (between positions 165 and 213)

Position in full-length protein (NoLSs shown in red):

MGRNKKKKRDGDDRRPRLVLSFDEEKRREYLTFGFKRKKVERKKAIEEIKQRLKEEQRL
REERHQEYLKMLAEREAELEADELDRLVTAKTESVQYDHPNHTVTVTISDLDSGARL
LGLTPPEGAGDRSEEEASSTKPTKALPRKSRDPLLSQRISL TASLHAHSRKKVKRKH
PRRAQDSKKPPRAPRTSKAQRRLTGKARHSGE

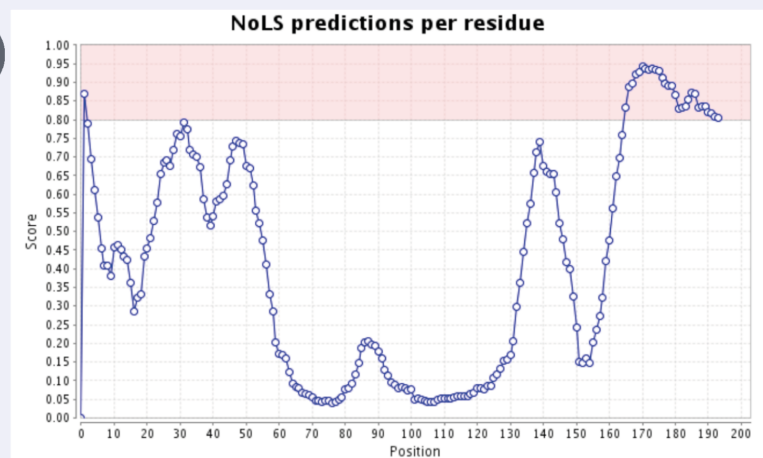


Figure 7: NoD⁹ input form (back). The user can input either a protein accession to query a pre-computed set of results (1) or paste a FASTA sequence (2a) to run an ab initio prediction. If a sequence prediction is requested this can be done with or without using a JPred prediction as a feature (2b; n.b. NoD uses JPred3). The prediction is started by clicking "Submit" (3). NOD output form (front). Any predicted nucleolar localisation sequences are shown both in isolation (4) and in context of the query sequence (5) and a line plot of the indicating the average score of 20 residue segments (6; see online help for more info).

The NoD server has been in continual use since its creation. A recent study employed NoD to scan for nucleolar localisation motifs in Fbw7 α , - β and - γ isoforms⁶¹. NoD correctly identified the nucleolar localisation signal in Fbw7 γ , suggested the presence of a weak signal for the nucleoplasmic Fbw7 α and reported no signal for the cytoplasmic Fbw7 β . The NoLS in Fbw7 γ was also shown to be the binding epitope for nucleophosmin (NPM1). Predicted

NoLS in the CENP-W and Tat proteins were also experimentally verified by the authors to bind NPM1⁶¹. Predicted NoLS were subsequently found in p14arf, another NPM1 interactor⁶². Mitrea *et al.*⁶³ found that, 63 % of a curated list of 83 NPM1 interactors had NoLS predicted by NoD and many of these NoLS overlapped with the so-called “multivalent R-motifs” the authors hypothesised. In a separate study, Duan *et al.*⁶⁴, used NoD to locate a suspected NoLS in the C-terminal domain of poly(A)-specific ribonuclease (PARN), which they then demonstrated experimentally was essential for nucleolar localisation.

2.6 Kinomer

The Kinomer^{5,65} webserver allows accurate identification of protein kinases (PKs) and their classification into kinase families. Kinomer also includes a browsable database of pre-computed predictions of PKs in 43 eukaryotic genomes organised in kinase classes. Kinomer works by scanning sequences against a library of PK multilevel profile HMMs. The Kinomer profile HMM library comprises 38(+1) profile HMMs and is known as “KinaseLib2” (KL2). KL2 was developed by iteratively sub-dividing the known PK families by sequence similarity and testing the performance of profile HMMs built from these subgroups to recall and classify other known PKs. KL2 was determined to be more accurate than an alternatively trialled KinaseLib1 (KL1), which contained 12 profile HMMs, one for each of the eight known conventional eukaryotic protein kinase (ePK) and four atypical protein kinase (aPK) families. The ePKs are AGC, CAMK, CK1, CMGC, RGC, STE, TK and TKL. The aPKs are Alpha, PIKK, PDHK and RIO. The Kinomer database was built by scanning whole proteomes against the Kinomer’s multilevel profile HMM library. Recent applications include the classification of kinases in the fungal pathogen *Cryptococcus neoformans*⁶⁶. They compared the proportions of kinase classes in the fungal pathogens, *C. neoformans*, *Candida albicans* and *Aspergillus fumigatus*. The *C. neoformans* Kinase Phenome Database contains Kinomer annotations.

Users are also able to browse the Kinomer database or classify a sequence by scanning against the Kinomer profile HMMs. Figure 8 displays the Kinomer sequence classification submission page. From here, a single sequence can be input via the text box or uploaded in FASTA format. The results of previous jobs can also be retrieved via the job ID. The Kinomer results page reports the best classification for the input sequence along with high-scoring alternative matches. Scores for all potential matches are also shown as well as the alignments corresponding to each match.

Classify a protein sequence

Sequences should be in any of the many [sequence formats](#) recognised by [EMBOSS](#). Sequence files should be in plain text format, NOT Microsoft Word.

1

My sequence is No file chosen

2

Notes:

3

Retrieve results from a previous job

Job ID:

4

Download raw results file

Kinase matches for AAK1_Hsap/52-313 in HMM library

KinaseLib2

The best match for your sequence is a kinase domain of group [CAMK](#)

Alternative matches above threshold are to groups [STE](#), [AGC](#), [CMGC](#)

5

Detailed Results for hits above threshold

ID	Score	E-value
CAMK_sub5	22.2	5e-16
CAMK_sub3	17.9	2.3e-14
STE_sub3	15.8	1.9e-14
CAMK_sub7	8.5	7.1e-15
AGC_sub1	7.0	1.6e-15
AGC_sub4	-3.4	4.3e-17
AGC_sub3	-13.3	2.2e-13
AGC_sub5	-20.7	1.5e-12
STE_sub4	-42.7	4e-11
STE_sub2	-45.2	1.8e-10
CMGC_sub3	-46.0	8.2e-13
STE_sub1	-47.7	1.4e-12
AGC_sub2	-48.0	1.5e-13
CMGC_sub1	-85.0	1.7e-09
CMGC_sub2	-109.5	6.6e-10

6

Alignments

CAMK_sub5 with score 22.2 E= 5e-16

```

1 ----LAEGGFAIVFLVR-TSNGMKCALKRMFVNNE-HdlQVCKREIQIMR 51
   + G F+ V +++ +++g k+AlK   + h+ +++++rE++i++
1 FikkiGrGnFgvVkecehraTgqklAlKIiklpkklh...everEvsilk 51

52 DLSGHNIVGYIDSSINNVSSGDVWEVLILMDFCRGGQVNNLMNQLQt 102
   l+ H ni++ +++ ++ + + + +++++ M ++ GG + ++ ++
52 glrvHpniiqLye....vfehtkk...rLyLVMElasGGELFDriiskgs. 102

103 gFTENEVLQIFCDTCEAVARLHQCKtpIIHRDLKVENILL--HGRH-YV 153
   +fTE+e+ +++ + ++Av LH+ ++I HRDLK EN+Ll hd+
103 .fTEreakavlkqilsAvrYLHsln..lIHRDLKPENLLlCehgdgadiK 153

154 LCDFGSATNkfnpqtegvnavEDEIKKYTTLSYRAPEMVNL-----YSg 204
   ++DFG A+ k +++++++ + ++ + + t+ Y APE ++
154 itDFGfAkik.....gelktfcGtPeYvAPEvlgkRRrhqkek. 204

205 KIIT--KA----DIWALGCLLYKLCYFTLPF----GESQVA---ICDGN 255
   ii t ++ DiW+LG +Y L ++PF g++ + i G
205 GiipTptPYgsvDiWslGViaYiLLsGspPFskntgdnlaedLriLeGk 255

256 FTI--PDNSRYSQDMHCLIRYMLEPDPDKRPDIYQV----- 256
   + + + + S+d +Ir +L dP +R+ + q+
256 yrfpseewaeiSedAKdfIrrlLkvdPeaRlTasqiLsHPWl 267

```

Figure 8: The Kinomer⁵ search input (left) and output forms (right). The user can paste a FASTA sequence (1) and start the classification by clicking “Submit” (2) or retrieve the results from a previously submitted job using the Kinomer job ID (3). If there are any hits to the Kinomer profile HMM library above Kinomer’s thresholds then the best matching kinase group (4) and alternative matches are reported (5). Alignments for each hit are shown below (6) and can be downloaded from the top of the page.

2.7 Xtal

Xtal⁶⁻⁸ is a collection of methods that predict the likelihood of a protein succeeding in a crystallisation experiment. Predicting the crystallisation propensity is useful for construct design and prioritising targets for structural genomics projects. The algorithms within Xtal are the OB-Score⁶, ParCrys⁷ and XANNPred⁸. The Xtal algorithms were developed over several years and each represented an improvement over the previous in terms of predictive performance as a result of improved algorithms and training data. Despite the precedence of XANNpred, which in our hands is the most accurate of the three, we provide and maintain the OB-Score and ParCrys since they remain useful and display their own

strengths. For example, although it was our first crystallisation propensity predictor, OB-score was one of four algorithms determined to be ideal for fast proteome-wide target selection in a recent review⁶⁷.

The OB-Score⁶ predicts whether a protein is likely to lead to a successful structure determination by calculating and assessing its predicted isoelectric point (pI) and grand average of hydrophobicity (GRAVY)⁶. This is achieved by comparing the pI, GRAVY values to proteins that have been successfully crystallised. This relatively simple approach yielded an accuracy of 69.8 % with AUC 0.711 on an independent test dataset⁷. The OB-Score was calculated for nearly 250 proteomes to compare each organisms' suitability for high-throughput crystallography as well as the sequences in Pfam 17.0⁶⁸ to identify a good candidate template structure for the protein families. These datasets remain available for download from the website for archival reasons but a researcher wishing to conduct a similar analysis is urged to use a recent dataset. For this reason, we recently calculated OB-Scores for 30,498,342 sequences across 16,449 families from Pfam 31.0; this new dataset and future updates can be found at http://www.compbio.dundee.ac.uk/xtal/ob_datasets/. It is also simple to calculate OB-Scores on a large-scale via the distributed Perl application, for example it took less than 30 seconds to calculate OB-Scores for the 42,500 sequences in PF00001.20. The OB-Score webserver returns the raw value of the OB-Score. This is interpreted with the following thresholds: a predictive threshold of 0.809 optimised accuracy over the test dataset; OB-Score ≥ 5 can be considered high-scoring, and 1.5 yields an optimal MCC (Matthews' Correlation Coefficient⁶⁹) on a real-world dataset. The OB-Score was also recently employed to prioritise tractable targets for insecticides against the malaria vector *Anopheles gambiae*⁷⁰.

ParCrys⁷ is a Parzen Window based estimator of crystallisation propensity that uses pI, hydrophobicity and the frequencies of S, C, G, F, Y, M residues only. The sequence is predicted as one of three classes: difficult to crystallise ("recalcitrant"); amenable to crystallisation ("amenable") or very amenable to crystallisation ("high-scoring"). Extensive feature selection was performed during the development of ParCrys. ParCrys surpassed the OB-Score even when using a reduced feature set of only pI and hydrophobicity, indicating that the Parzen Window model itself provided significant advantages. The inclusion of the remaining residue frequency features led to further performance gains compared to the OB-score. Adding other amino acids as features besides S, C, G, F, Y and M led to performance degradation, which was reasoned to be due to correlation between pI and charged residue frequencies and consequently a no-benefit decline in the parameter/ observation ratio. ParCrys achieved an accuracy of 79.1 % with AUC 0.844.

XANNpred-PDB and XANNpred-SG (together XANNpred⁸) are neural networks that predict whether a protein is likely to produce diffraction quality crystals based on amino acid frequencies (including dipeptides), sequence length and molecular weight as well as predicted isoelectric point, hydrophobicity (GES), secondary structure (JPred), transmembrane regions (TMHMM2) and protein disorder (RONN)⁸. The two neural networks differ only in their training where XANNpred-PDB was trained with a positive training set derived from the PDB and XANNpred-SG's positive training set was derived from the now retired PepcDB, which included sequences that were known to crystallise but had

not necessarily been solved at the time. XANNpred achieved AUC 0.854⁸. The XANNpred webserver calculates the required sequence features and runs both neural networks to provide the prediction results. XANNpred also provides predictions for sub-sequences within the query via a sliding window approach. This provides region-specific crystallisation propensities that are particularly useful for construct design. Figure 9 illustrates how the XANNpred windowed predictions vary over the XANNpred demo sequence (PDBT26731). In this example, the windows centred on residues 33-47 are above the threshold for XANNpred-PDB, this suggests that residues 2-78 are more amenable to crystallisation than the remaining sequence (i.e., these residues are in at least one high-scoring 31 residue window).

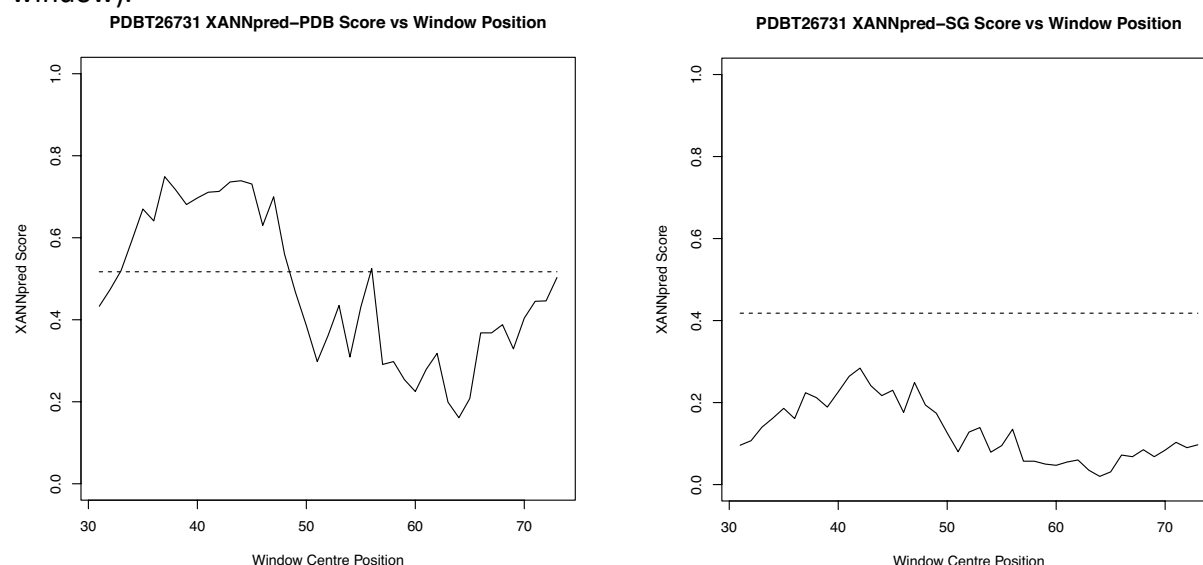



Figure 9: XANNpred⁸ windowed predictions for XANNpred-PDB (left) and XANNpred-SG (right). The prediction threshold is indicated by the dashed line (0.517 for XANNpred-PDB; 0.418 for XANNpred-SG). The windows are 61 residues long and so the first window is centred at residue 31. A relaxed interpretation considers high-scoring regions as those residues that are contained within a high-scoring window (i.e., ± 31 residues of the window centre). A conservative interpretation is restricted to where the window centres are above the prediction threshold. XANNpred provides these figures as attachments in the results email.


All three predictors in Xtal are available via web forms. The OB-Score⁶ and ParCrys⁷ are accessed via a single submission page whilst XANNpred⁸ submissions are made via its own page. Figure 10 illustrates the OB-Score/ ParCrys submission and example results page (NB. the submission form for XANNpred is very similar). A user can submit sequences in FASTA format via a text box or file upload. After a few moments OB-Score and ParCrys predictions are reported via a results page in an HTML table. XANNpred predictions are returned in tabular format via email. If requested, XANNpred windowed predictions are included in the results email as PDF attachments (Figure 9). Alternatively, the OB-Score Perl application and data can be downloaded and run locally after following some minor configuration instructions returning results in TSV format (Tab Separated Values). Pre-calculated OB-Scores are available for Pfam 31.0.



UNIVERSITY OF
DUNDEE

SSPF Crystallisation Propensity Predictors

THE BARTON GROUP



OB-Score
& ParCrys

More Information

Predictions to estimate a protein's propensity to form diffraction-quality crystals are calculated using the methods ParCrys and the OB-Score. ParCrys is a Parzen Window approach based on calculated isoelectric point, hydrophobicity and the frequencies of S, C, G, F, Y, M residues. For more details see [Overton, Padovani, Girolami & Barton \(2008\), Bioinformatics 24:901-907](#) and the SSPF crystallisation propensity predictors [information page](#). The OB-Score is a Z-score scale based on calculated isoelectric point and hydrophobicity. For more details, see [Overton & Barton \(2006\), FEBS Lett. 580, 4005-4009](#) and the SSPF crystallisation propensity predictors [information page](#). Note that results usually take about 3 minutes to return.

The OB-Score algorithm and associated data are available for download from [here](#).

Paste your fasta format sequence(s) here: ([Example of a sequence in fasta format](#))

1a

```
>PDBT26731
ADEEESTPCDNVETDQQTFAAFNKAERELQSAIDELIERMRDQFGDEA
GLMSRIEAAEKVWSQLRDADCKVETHAEQPGSNAYQIAWNSCIAQRSDERA
```

1b

Or upload a fasta file:

2

[Download ParCrys datasets](#)

3

Your Results

Identifier	ParCrys Prediction	ParCrys Score	OB-Score	GRAVY	Isoelectric Point	Sequence Length
PDBT26731	Recalcitrant	2340000	-0.16	-0.92	4.01	103

Please cite: [Overton, Padovani, Girolami & Barton \(2008\) "ParCrys: A Parzen Window Density Estimation Approach to Protein Crystallisation Propensity Prediction," Bioinformatics 24:901-907](#) and/or [Overton & Barton \(2006\) "A normalised scale for structural genomics target ranking: The OB-Score," FEBS Lett. 580, 4005-4009](#), as appropriate.

[More Information](#)

Sequence(s) Analysed

```
>PDBT26731
ADEEESTPCDNVETDQQTFAAFNKAERELQSAIDELIERMRDQFGDEA
GLMSRIEAAEKVWSQLRDADCKVETHAEQPGSNAYQIAWNSCIAQRSDERA
```

Figure 10: XTal input form for OB-Score⁶ and ParCrys⁷. Xtal output form. Users can input a sequence or multiple sequences by pasting FASTA format into the textbox (1a) or uploading a FASTA file (1b). The prediction is then run by clicking the "GO!" button. A link to download the ParCrys datasets is provided at the bottom of the page. Once the calculation is complete, the results page will load and display a table listing the OB-Score and the ParCrys score and prediction for each submitted sequence alongside the GRAVY, pI and the sequence length.

2.8 AMAS

AMAS¹⁴ (Analysis of Multiply Aligned Sequences) is an hierarchical conservation analysis algorithm based on a set representation of amino acid physicochemical properties. The AMAS server has been in operation since 1994. In addition to the standard identification of residues that are conserved in all sequences at a position, AMAS can indicate various types of sub-group conservation. For instance, the AMAS output differentiates columns that are conserved in some but not all subgroups (*conserved and similar*; e.g., where a structural constraint is lost in particular subgroups) from columns that are conserved in most subgroups but where each subgroup conserves a different feature (*conserved but different*; e.g., sites important for specificity). This description is admittedly abstract and a more complete illustration can be found in the AMAS paper.

Figure 11 displays the AMAS submission page where users can run the analysis on their own multiple sequence alignments. FASTA, PFAM or AMPS formatted alignments may be pasted directly into the provided textbox or uploaded from the user's local storage. AMAS also requires the user to provide subgroup classifications. Suitable groups could be derived from overall sequence similarity, functional similarity or taxonomic relationships. Group membership is indicated by lines of comma delimited sequence indexes or ranges as indicated in the paragraph preceding the textbox. Note that the AMAS conservation analysis

can be run with only a single group specified but, in this case, only the standard conservation score can be returned. The AMAS analysis can then be run with default settings by clicking the “Do The Analysis” button at the bottom of the page.

Use Example Files

Clear Example Files

Select the format of your input alignment: FASTA

1

Paste the alignment here:

>DBQXJ5_SELM/9-78

-AAVQEKTNAMAKAGEKMDSAKASGTEKAEKVTAHNKAEKHEAKERKHMKEAEAAAGHLHEKQAEINIAKAE

>QBH683_ORYSJ/1-71

MQAVKEKVKDKVSAVKAGKGVSKAKADEKKEVATARSHAERELAHERAKARVAAKMEHLHQDKALHREEAI

>IHT1W8_BRADI/1-71

MQAAKEKVKDGVSAAKAKTKVAQAKAEKAEAAATARSKAEKELAHERGKAKVAAKMEHLHQDKAVHREEAI

>CSZZP9_SORBI/1-71

MQAVKEKVKDAMSTAKAKAREQEAKAEKAGKATASSAEKELVHQRAKAKVAAKMDLHQEKAIHKEEI

>WSDP76_WHEAT/1-71

MEHAKEKMKDGASAAKAKSTITTKAKVAEKAEAAATARSHDERELAHERGAAKVAAAEQLHQDKAAHREDAM

or upload an alignment file: Choose file No file chosen

2

Enter "Sensible Groups" here:

YOU MUST DEFINE SOME GROUPS!

2-15

16-20

21-24

25-36

37-39

3a

Select property table: extra_ul

extra_ul is for extracellular proteins, intra_ul is for intracellular. This just affects the treatment of Cys. extra is a good default. There are other property tables supplied with AMAS and you can devise your own, but these features aren't available on this server.

3b

Select the conservation threshold: 7

This is the most useful parameter to fiddle with. Higher numbers highlight more highly conserved positions, lower numbers highlight less conserved positions. Values between 5 and 10 are worth trying with the property tables available here; 7 is a good default starting point.

3c

You can set some of the many AMAS options here. See the [AMAS manual](#) for details.

Font size for PostScript output

6

Output orientation

landscape

Shading

colour

Ignore atypical residues (percentage)

0

Number of gaps to ignore per sub group

0

Frequency histogram, or similarity/difference report

histogram

If histogram, enter height for bars

10

Show identifiers or numbers in report file

numbers

Mask amino acids at unconserved positions

☒ No. ☐ Yes.

Only produce the highlighted alignment

☒ No. ☐ Yes.

4

The results of your analysis will persist for 30 MINUTES before being deleted.

Do The Analysis

Reset

amas@compbio.dundee.ac.uk

Figure 11: The AMAS¹⁴ input form. 1) AMAS accepts FASTA formatted alignments (also AMPS or Pfam) input via the textbox or file upload. 2) Groups are defined via a textbox with one line per group and sequences referred to by their row index (e.g. “1-5” on a line defines a group of the first five sequences). Advanced options: 3a) the property table is selected (default: extra_ul); 3b) the conservation threshold is set (default: 7) and 3c) this section allows several to be set. The first few relate to alignment formatting with Alscript, “atypical residues” and “gaps to ignore” are described in the text, and other display options. “Only highlighted alignment” will prevent the histogram or difference table being printed (e.g., 2 in Figure 12).

Figure 12 illustrates the AMAS output visualisation. The block colouring indicates the subgroup conservation, distinguishing identity in all subgroups (red), identity within a subgroup (blue) and conservation within a subgroup (green). The histograms summarise the AMAS comparison of the subgroups. The upper histograms show the overall conservation (red) and subgroup similarities (pink) whilst the lower histogram (orange) shows the average of the subgroup differences. The most dissimilar sites in terms of subgroup-

subgroup comparison (i.e., large values on the inverted histogram) are most likely to be important for specificity and are worth closer inspection. AMAS results are also available in text format.

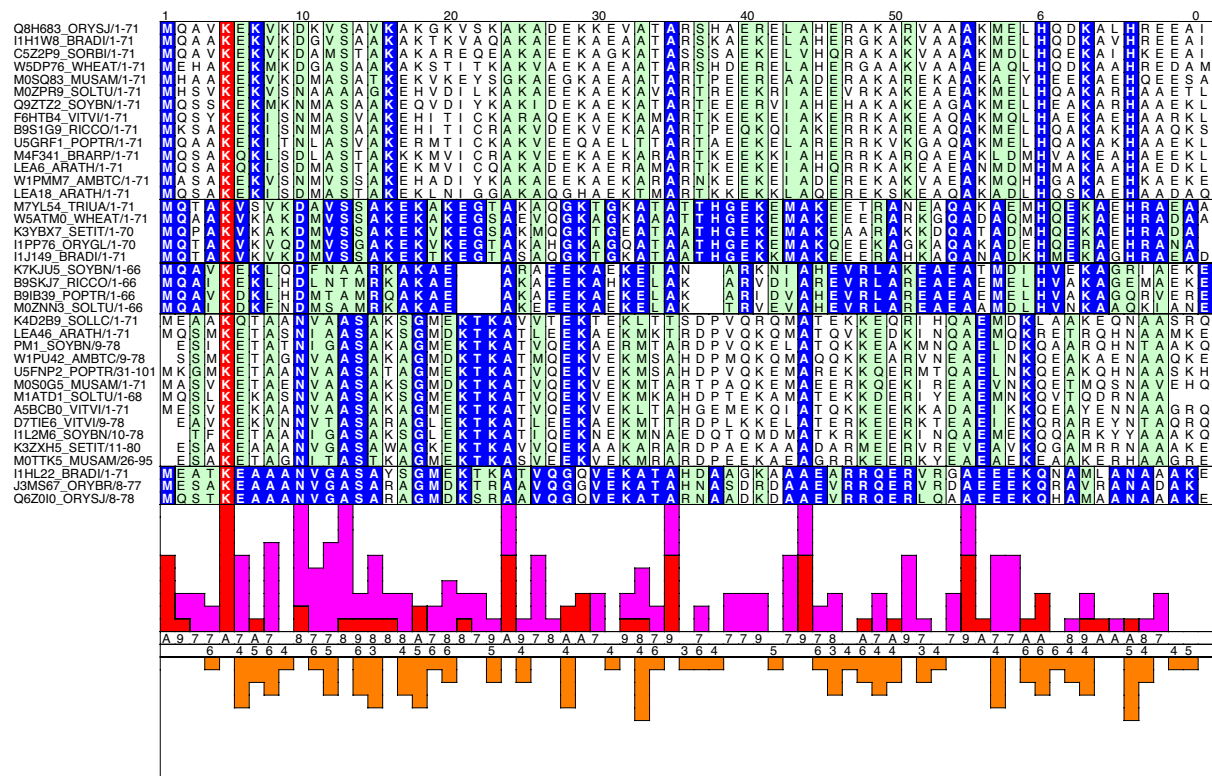


Figure 12: AMAS¹⁴ results visualisation of an illustrative analysis upon Pfam PF03760. The alignment illustrates within group and between group conservation. Within group conservation is illustrated by block shading within the subgroups: blue indicates subgroup identity whilst green indicates property conservation. Additionally, red shading indicates total conservation across all groups. The histogram displays the similarities (orange) and differences scores (violet). The visualisation is generated with Alscript⁴¹.

Several important settings can be adjusted. The *property table* selection defines the amino acid physicochemical set memberships used to define the properties that can be conserved. Three options are available via the web interface: *extra_ul* is recommended for extracellular proteins where Cys is assumed to form disulphide bonds whilst *intra_ul* is recommended for intracellular proteins where Cys are assumed to be present as free thiols. The third option available *ch* is specifically for detecting conserved charges and changes in the polarity of conserved charges in certain subgroups and defines positive (His, Arg and Lys), negative (Glu and Asp) and charged (His, Arg, Lys, Glu and Asp) sets (NB. This last group is documented in the AMAS manual). The *conservation threshold*, T defines what AMAS will consider to be a conserved position in a subgroup or subgroup pair. Higher (T) values will also result in a more specific analysis since only subgroup pairs where both subgroups have individual conservation scores $> T$ are evaluated. Note that T must be less than the maximum possible conservation score (C_{max}), which is determined by the number of properties in the property table; the server will error and report the allowed values if this rule is broken. The parameters labelled “Ignore atypical residues” and “Number of gaps to ignore per subgroup” influence how sensitive the conservation score is to gaps and potentially aberrant residues. The remaining parameters in the lower options section control the formatting of the Alscript⁴¹ output alignment. Of particular note is the “Frequency histogram, or similarity/difference report” option, which controls whether summaries of the subgroup pairwise comparisons are shown (*histogram*; best when there are many subgroups) or the

individual subgroup pair conservations are shown (*differences*; clearest when there are only a few groups).

3 New Services under Development

We are currently developing several new services for DRSASP. Slivka is an evolution of the JABAWS¹⁰ concept written in Python that is designed to improve upon JABAWS' limitations. ProteoCache is the DRSASP "data warehouse", at its core it is an Apache Cassandra database designed to hold and return pre-calculated results for all DRSASP tools, accelerating performance and providing a means to perform integrated analyses across our resources. ProteoFAV, ProIntVar and VarAlign are Python packages that we created to meet our own research requirements for carrying out integrated analyses across protein sequences, multiple sequence alignments, 3D structures and human genetic variation. These tools are discussed individually in brief below; further detailed discussion of their capabilities will be published upon each tool's release.

3.1 Slivka

Slivka is a new web services framework currently in development that will supersede the JABAWS framework. Slivka is implemented in Python with Flask, ZeroMQ and MongoDB. Key advantages of Slivka compared to JABAWS are significantly simplified tool configuration, better facilities for tool chaining and the capability of Slivka to generate tool specific web forms. Tool configuration in Slivka requires just two YAML files: a run configuration file to specify the command line interface of the tool and a form configuration that specifies the parameters exposed through web API. Files uploaded to or generated by (i.e., results) the Slivka server for analysis (e.g., sequence files, MSAs) can be referenced via a *uuid*, which facilitates tool chaining since results can be referenced server-side. Slivka is currently in advanced testing stages and we expect to deploy a public production server early in 2020.

3.2 ProteoCache

ProteoCache is a database containing pre-computed results of DRSASP and other applications for whole proteomes built with Apache Cassandra together with a Node.js API based on DataStax's cassandra-driver. Apache Cassandra is a scalable and robust NoSQL database. At the time of writing, the database contains JPred4 predictions (Including full alignments and PSI-BLAST profiles) for most of the Human (57,823 sequences; 78 %), *S. cerevisiae* S288C (5,049; 83 %) and *E. coli* K-12 (4,144; 94 %) UniProt reference proteomes as well as disorder predictions for 79,513 sequences from the four disorder predictors provided by JABAWS. Tables in ProteoCache are indexed by sequence to allow fast lookup of new DRSASP queries. Currently, JPred4 interfaces with ProteoCache to improve the performance of JPred4 for previously run sequences. Our goal is that all DRSASP applications will similarly interface with the ProteoCache to improve performance of our web services. The ProteoCache itself will in the future be able to serve bulk downloads of whole proteomes or other large selections of sequences and also permit complex queries over the data.

3.3 ProteoFAV, ProIntVar & VarAlign

Over the last few years we have been researching how human genetic variants are distributed in proteins with respect to protein structure and conservation³². This has led to

the development of software that simplifies the complex task of connecting the heterogenous data derived from variants, protein sequences, protein structures and multiple sequence alignments. Our approach is to represent the data as Pandas DataFrames. Once all these data are harmonised, we can conduct complex queries and aggregations. For example, “return all missense variants at residues where the position is conserved and involved in a hydrogen bond with a ligand in Pfam PF00001” and “count all missense variants in each alignment column of PF00017”. This software is being developed as a series of Python modules and we will release the libraries and provide a webservice through Slivka upon journal publication of an updated version of our analysis of human variation in Pfam alignments³².

Figure 13 illustrates one view of these data in Jalview. gnomAD⁷¹ variants were mapped to the residues in the Pfam³¹ SH2 domain alignment and are formatted as Jalview features with VarAlign. The sequences shown in Figure 13 are amongst the most missense depleted (constrained) human sequences in this family and were identified in Jalview by (View → Feature settings... → Sequence sort by Density) when only the missense variants were shown. VarAlign also fetched protein-ligand interaction data for all sequences in the alignment from the PDBe with ProIntVar. Rendering these data as features in Jalview allows the identification of co-located missense variants at these sites if there are any. Jalview allows quick visualisation of these features on a mapped protein structure through its integration with UCSF Chimera.

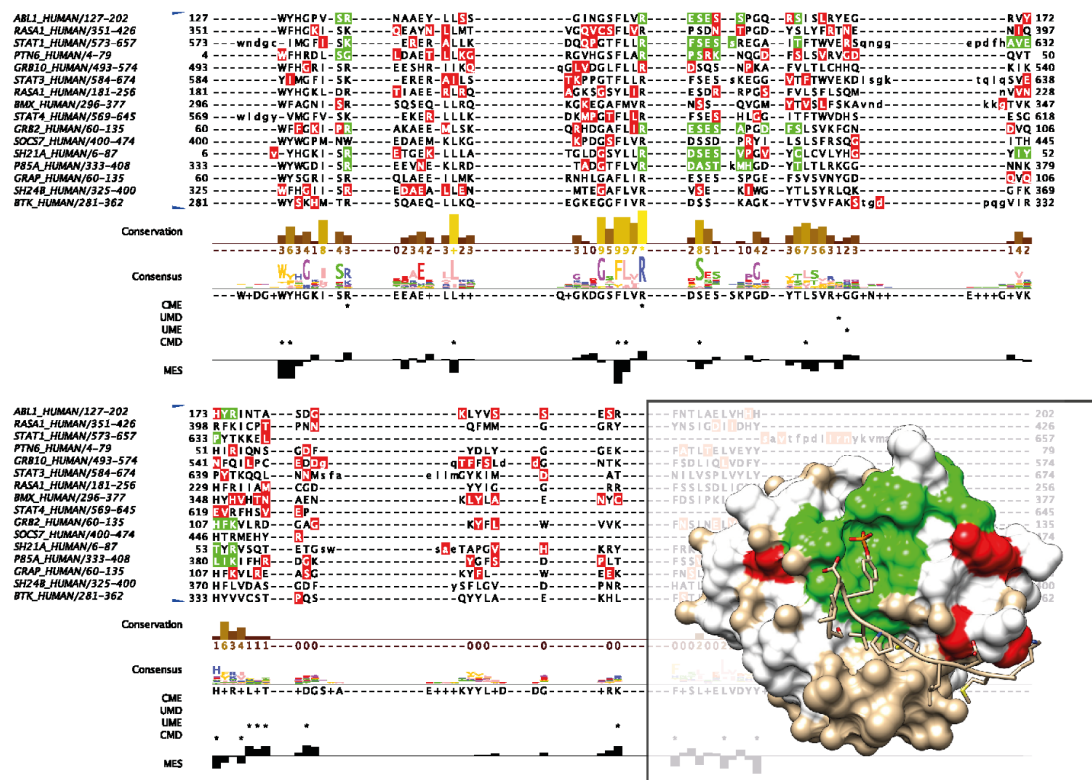


Figure 13: Example output from VarAlign and ProIntVar analysis³² of SH2 domains from Pfam³¹ PF00017 visualised with Jalview¹ and UCSF Chimera⁴². In the alignment, nine of the most missense depleted SH2 domains are shown. The locations of missense variants from the gnomAD⁷¹ dataset are shown as semi-transparent red features. The locations of residue-ligand interactions by ligands that bind in the SH2 canonical binding-site are shown in semi-transparent green. In these proteins,

no missense variants occur at these positions (i.e., these features do not overlap). Four annotation tracks are shown, from top to bottom: Jalview calculated consensus; whether positions are classified as unconserved-missense depleted (UMD), unconserved-missense enriched (UME), conserved-missense enriched (CME) or conserved-missense depleted (CMD) by VarAlign default thresholds and the missense enrichment score. The structure shows the interaction between the SH2 domain of Phosphatidylinositol 3-kinase regulatory subunit alpha and the Platelet-derived growth factor receptor beta phosphotyrosyl peptide in PDB ID: 2IUI. The locations of missense variants from the gnomAD dataset are shown red. The locations of residue-ligand interactions by ligands that bind in the SH2 canonical binding-site – in any structure that maps to this protein – are shown in green.

4 DRSASP Workflows in Jalview

Jalview¹ – a program for multiple sequence alignment editing, visualisation and analysis – provides an interface to many of the DRSASP tools. This enables users to carry out sophisticated workflows that combine DRSASP tools and Jalview’s built-in analysis capabilities interactively. For example, a useful Jalview workflow is to cluster sequences iteratively, prune outliers and align the remaining sequences. Once the alignment is judged sufficiently accurate, further DRSASP services can be invoked to calculate residue conservation, predict specificity determining sites (SMERFS; *Calculate* → *Calculate Tree* → AMAS) and predict structural features (solvent accessibility, secondary structure and disorder). This rich annotation set can help interpret experimental observations (e.g. UniProt mutation data) and/ or provide an enhanced understanding of the protein by projecting them onto structure in a Jalview linked MSA-structure session.

5 General Developments

We are committed to improving our software and data practices by working towards implementing OSS recommendations⁷² (see also guidelines cited therein) and the FAIR principles⁷³. In this vein, we will continue to add DRSASP resources to the bio.tools registry⁷⁴, deposit annotations in collaborative repositories (e.g., PDBe-KB, see below) and make datasets and code publicly available. Some DRSASP projects align well with these ideas in their very concept. For instance, the JABAWS and Slivka frameworks will enhance the interoperability (aggregate services) and reproducibility (consistent execution environment) of bioinformatics tools in general whilst the ProteoCache promotes data reuse and integration.

A relevant development is our work to improve DRSASP’s efficiency in whole proteome analyses. The pre-computed data in ProteoCache (§3.2) is one aspect of this. Another direction we have pursued is the annotation of PDB structures in collaboration with the PDBe-KB⁷⁵ project as a data depositor. So far, we have annotated the set of human sequences in PDBe with 14-3-3-Pred predictions. This resulted in 1,941 representative PDB chains receiving at least one positive prediction. These are accessible via PDBe-KB (e.g., <https://www.ebi.ac.uk/pdbe/pdbe-kb/proteins/Q92879>; see in expanded “Predicted PTM sites” track in “Functional Annotations” section). Python scripts to assist running DRSASP tools in a high-throughput manner and generating PDBe-KB compliant JSON output are available from https://github.com/bartongroup/FM_FunPDBe.git.

As part of the PDBe-KB deposition process we were required to conform DRSASP results data to an agreed upon JSON standard. This effort is the beginning of a larger effort to harmonise the data out from DRSASP tools. A further step will be to ensure this work is efficiently translated into better Jalview integration for the DRSASP tools that are not currently well integrated. This might be achieved by introducing PDBe-KB JSON parsing to

Jalview or by converting the JSON to an existing Jalview format. Whilst this does not constitute full Jalview integration (i.e., the services are not called from Jalview) this may prove a useful stopgap and is worthwhile anyway to enable high-throughput data generation where it is advantageous to generate required data in bulk (e.g., an analysis on Pfam might generate Jalview compatible annotations for 1000s of alignments, these annotated families can then be “browsed” with Jalview).

Lastly, we are making improvements in the testing and portability of DRSASP tools. A key priority in the short-term is to improve the deployment of the DRSASP tools with the initial focus being JPred. This will involve applying modern technologies such as containerisation (e.g., Docker) or modern dependency management solutions (e.g., Conda). In addition to simplifying our internal maintenance workflows, this will have the added advantage of simplifying the local installation of JPred so that users will have the option of running a local instance. Moreover, improving the portability of our software is an important component of our efforts to ensure our work is as reproducible as possible. On the technical front, we have also made improvements to DRSASP service reliability through the introduction of continuous monitoring. In addition to standard HTTP checks we now use end-to-end interface tests for JPred and JABAWS services.

6 Conclusion

The Dundee Resource for Sequence Analysis and Structure Prediction provides several bioinformatics web services for the scientific community. The tools address a wide variety of biological questions but are connected by the common themes of protein sequence analysis and structure prediction. The services provide secondary structure prediction, disorder prediction, multiple sequence alignment, functional site prediction and more. DRSASP tools are accessible via web forms, programmatic APIs and some are suitable for local installation. A unique aspect of DRSASP is its tight integration with Jalview.

As well as maintaining and continually developing existing tools DRSASP has several new services that are close to release. Slivka and ProteoCache will improve the delivery of DRSASP services but they will also enable new developments in the future. (e.g., aggregated services and large-scale integrated analyses). ProteoFAV, ProIntVar and VarAlign are new services close to release that will enable new research, especially at the intersection of human genetics and protein structure.

7 Acknowledgements

We are grateful to all authors of methods that have been included in DRSASP. We thank Tom Walsh and the Dundee Research Computing team for supporting our IT infrastructure and James Abbott for his expert sysadmin advice. This work was supported by Biotechnology and Biological Sciences Research Council Grants [BB/J019364/1 and BB/R014752/1] and Wellcome Trust Biomedical Resources Grant [101651/Z/13/Z].

8 References

- 1 Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M. & Barton, G. J. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189-1191, doi:10.1093/bioinformatics/btp033 (2009).

- 2 Drozdetskiy, A., Cole, C., Procter, J. & Barton, G. J. JPred4: a protein secondary structure prediction server. *Nucleic Acids Res* **43**, W389-394, doi:10.1093/nar/gkv332 (2015).
- 3 Madeira, F. *et al.* 14-3-3-Pred: improved methods to predict 14-3-3-binding phosphopeptides. *Bioinformatics* **31**, 2276-2283, doi:10.1093/bioinformatics/btv133 (2015).
- 4 Manning, J. R., Jefferson, E. R. & Barton, G. J. The contrasting properties of conservation and correlated phylogeny in protein functional residue prediction. *BMC Bioinformatics* **9**, 51, doi:10.1186/1471-2105-9-51 (2008).
- 5 Martin, D. M., Miranda-Saavedra, D. & Barton, G. J. Kinomer v. 1.0: a database of systematically classified eukaryotic protein kinases. *Nucleic Acids Res* **37**, D244-250, doi:10.1093/nar/gkn834 (2009).
- 6 Overton, I. M. & Barton, G. J. A normalised scale for structural genomics target ranking: the OB-Score. *FEBS Lett* **580**, 4005-4009, doi:10.1016/j.febslet.2006.06.015 (2006).
- 7 Overton, I. M., Padovani, G., Girolami, M. A. & Barton, G. J. ParCrys: a Parzen window density estimation approach to protein crystallization propensity prediction. *Bioinformatics* **24**, 901-907, doi:10.1093/bioinformatics/btn055 (2008).
- 8 Overton, I. M., van Niekerk, C. A. & Barton, G. J. XANNpred: neural nets that predict the propensity of a protein to yield diffraction-quality crystals. *Proteins* **79**, 1027-1033, doi:10.1002/prot.22914 (2011).
- 9 Scott, M. S., Troshin, P. V. & Barton, G. J. NoD: a Nucleolar localization sequence detector for eukaryotic and viral proteins. *BMC Bioinformatics* **12**, 317, doi:10.1186/1471-2105-12-317 (2011).
- 10 Troshin, P. V. *et al.* JABAWS 2.2 distributed web services for Bioinformatics: protein disorder, conservation and RNA secondary structure. *Bioinformatics* **34**, 1939-1940, doi:10.1093/bioinformatics/bty045 (2018).
- 11 Cole, C., Barber, J. D. & Barton, G. J. The Jpred 3 secondary structure prediction server. *Nucleic Acids Res* **36**, W197-201, doi:10.1093/nar/gkn238 (2008).
- 12 Troshin, P. V., Procter, J. B. & Barton, G. J. Java bioinformatics analysis web services for multiple sequence alignment--JABAWS:MSA. *Bioinformatics* **27**, 2001-2002, doi:10.1093/bioinformatics/btr304 (2011).
- 13 *ELIXIR-UK Node Services*. <https://elixiruknode.org/node-services> (accessed 30th August 2019)."
- 14 Livingstone, C. D. & Barton, G. J. Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. *Comput Appl Biosci* **9**, 745-756 (1993).
- 15 Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* **7**, 539, doi:10.1038/msb.2011.75 (2011).
- 16 Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947-2948, doi:10.1093/bioinformatics/btm404 (2007).
- 17 Katoh, K. & Toh, H. Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinform* **9**, 286-298, doi:10.1093/bib/bbn013 (2008).
- 18 Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**, 772-780, doi:10.1093/molbev/mst010 (2013).

- 19 Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792-1797, doi:10.1093/nar/gkh340 (2004).
- 20 Notredame, C., Higgins, D. G. & Heringa, J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* **302**, 205-217, doi:10.1006/jmbi.2000.4042 (2000).
- 21 Do, C. B., Mahabhashyam, M. S., Brudno, M. & Batzoglou, S. ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Res* **15**, 330-340, doi:10.1101/gr.2821705 (2005).
- 22 Liu, Y., Schmidt, B. & Maskell, D. L. MSAProbs: multiple sequence alignment based on pair hidden Markov models and partition function posterior probabilities. *Bioinformatics* **26**, 1958-1964, doi:10.1093/bioinformatics/btq338 (2010).
- 23 Ye, Y. *et al.* GLProbs: Aligning Multiple Sequences Adaptively. *IEEE/ACM Trans Comput Biol Bioinform* **12**, 67-78, doi:10.1109/TCBB.2014.2316820 (2015).
- 24 Linding, R. *et al.* Protein disorder prediction: implications for structural proteomics. *Structure* **11**, 1453-1459 (2003).
- 25 Dosztanyi, Z., Csizmok, V., Tompa, P. & Simon, I. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* **21**, 3433-3434, doi:10.1093/bioinformatics/bti541 (2005).
- 26 Yang, Z. R., Thomson, R., McNeil, P. & Esnouf, R. M. RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics* **21**, 3369-3376, doi:10.1093/bioinformatics/bti534 (2005).
- 27 Linding, R., Russell, R. B., Neduva, V. & Gibson, T. J. GlobPlot: Exploring protein sequences for globularity and disorder. *Nucleic Acids Res* **31**, 3701-3708, doi:10.1093/nar/gkg519 (2003).
- 28 Bungard, D. *et al.* Foldability of a Natural De Novo Evolved Protein. *Structure* **25**, 1687-1696 e1684, doi:10.1016/j.str.2017.09.006 (2017).
- 29 Bernhart, S. H., Hofacker, I. L., Will, S., Gruber, A. R. & Stadler, P. F. RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics* **9**, 474, doi:10.1186/1471-2105-9-474 (2008).
- 30 Britto-Borges, T. & Barton, G. J. A study of the structural properties of sites modified by the O-linked 6-N-acetylglucosamine transferase. *PLoS One* **12**, e0184405, doi:10.1371/journal.pone.0184405 (2017).
- 31 El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucleic Acids Res* **47**, D427-D432, doi:10.1093/nar/gky995 (2019).
- 32 MacGowan, S. A. *et al.* Human Missense Variation is Constrained by Domain Structure and Highlights Functional and Pathogenic Residues. *bioRxiv*, 127050, doi:10.1101/127050 (2017).
- 33 Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* **292**, 195-202, doi:10.1006/jmbi.1999.3091 (1999).
- 34 Cuff, J. A. & Barton, G. J. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins* **40**, 502-511 (2000).
- 35 Heffernan, R. *et al.* Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. *Sci Rep* **5**, 11476, doi:10.1038/srep11476 (2015).
- 36 Buchan, D. W. *et al.* Protein annotation and modelling servers at University College London. *Nucleic Acids Res* **38**, W563-568, doi:10.1093/nar/gkq427 (2010).

- 37 Yachdav, G. *et al.* PredictProtein--an open resource for online prediction of protein structural and functional features. *Nucleic Acids Res* **42**, W337-343, doi:10.1093/nar/gku366 (2014).
- 38 Wang, S., Peng, J., Ma, J. & Xu, J. Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields. *Scientific Reports* **6**, 18962, doi:10.1038/srep18962 (2016).
- 39 Lupas, A., Van Dyke, M. & Stock, J. Predicting coiled coils from protein sequences. *Science* **252**, 1162-1164, doi:10.1126/science.252.5009.1162 (1991).
- 40 Matsumura, H. *et al.* Crystal structure of rice Rubisco and implications for activation induced by positive effectors NADPH and 6-phosphogluconate. *J Mol Biol* **422**, 75-86, doi:10.1016/j.jmb.2012.05.014 (2012).
- 41 Barton, G. J. ALSCRIPT: a tool to format multiple sequence alignments. *Protein Eng* **6**, 37-40 (1993).
- 42 Pettersen, E. F. *et al.* UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem* **25**, 1605-1612, doi:10.1002/jcc.20084 (2004).
- 43 Chandru, A., Bate, N., Vuister, G. W. & Cowley, S. M. Sin3A recruits Tet1 to the PAH1 domain via a highly conserved Sin3-Interaction Domain. *Scientific Reports* **8**, 14689, doi:10.1038/s41598-018-32942-w (2018).
- 44 Raia, P. *et al.* Structure of the DP1–DP2 PolD complex bound with DNA and its implications for the evolutionary history of DNA and RNA polymerases. *PLOS Biology* **17**, e3000122, doi:10.1371/journal.pbio.3000122 (2019).
- 45 Tatham, M. H. *et al.* A Proteomic Approach to Analyze the Aspirin-mediated Lysine Acetylome. *Mol Cell Proteomics* **16**, 310-326, doi:10.1074/mcp.O116.065219 (2017).
- 46 Pellizza, L., Smal, C., Rodrigo, G. & Arán, M. Codon usage clusters correlation: towards protein solubility prediction in heterologous expression systems in *E. coli*. *Scientific Reports* **8**, 10618, doi:10.1038/s41598-018-29035-z (2018).
- 47 Zhao, Z., Xie, L. & Bourne, P. E. Insights into the binding mode of MEK type-III inhibitors. A step towards discovering and designing allosteric kinase inhibitors across the human kinome. *PLoS One* **12**, e0179936, doi:10.1371/journal.pone.0179936 (2017).
- 48 Le, Q., Sievers, F. & Higgins, D. G. Protein multiple sequence alignment benchmarking through secondary structure prediction. *Bioinformatics* **33**, 1331-1337, doi:10.1093/bioinformatics/btw840 (2017).
- 49 Valdar, W. S. Scoring residue conservation. *Proteins* **48**, 227-241, doi:10.1002/prot.10146 (2002).
- 50 Shenkin, P. S., Erman, B. & Mastrandrea, L. D. Information-theoretical entropy as a measure of sequence variability. *Proteins* **11**, 297-313, doi:10.1002/prot.340110408 (1991).
- 51 Zvelebil, M. J., Barton, G. J., Taylor, W. R. & Sternberg, M. J. Prediction of protein secondary structure and active sites using the alignment of homologous sequences. *J Mol Biol* **195**, 957-961, doi:10.1016/0022-2836(87)90501-8 (1987).
- 52 Valdar, W. S. & Thornton, J. M. Conservation helps to identify biologically relevant crystal contacts. *J Mol Biol* **313**, 399-416, doi:10.1006/jmbi.2001.5034 (2001).
- 53 Bridges, D. & Moorhead, G. B. 14-3-3 proteins: a number of functions for a numbered protein. *Sci STKE* **2005**, re10, doi:10.1126/stke.2962005re10 (2005).
- 54 Tinti, M. *et al.* ANIA: ANnotation and Integrated Analysis of the 14-3-3 interactome. *Database (Oxford)* **2014**, bat085, doi:10.1093/database/bat085 (2014).

- 55 Lu, Y. *et al.* Characterization of ubiquitin ligase SIATL31 and proteomic analysis of 14-3-3 targets in tomato fruit tissue (*Solanum lycopersicum* L.). *J Proteomics* **143**, 254-264, doi:10.1016/j.jprot.2016.04.016 (2016).
- 56 Kim, S. W. *et al.* Role of 14-3-3 sigma in over-expression of P-gp by rifampin and paclitaxel stimulation through interaction with PXR. *Cell Signal* **31**, 124-134, doi:10.1016/j.cellsig.2017.01.001 (2017).
- 57 Yu, J. *et al.* Wnt5a induces ROR1 to associate with 14-3-3 ζ for enhanced chemotaxis and proliferation of chronic lymphocytic leukemia cells. *Leukemia* **31**, 2608, doi:10.1038/leu.2017.132 (2017).
- 58 Obsil, T., Ghirlando, R., Klein, D. C., Ganguly, S. & Dyda, F. Crystal structure of the 14-3-3 ζ :serotonin N-acetyltransferase complex. a role for scaffolding in enzyme regulation. *Cell* **105**, 257-267, doi:10.1016/s0092-8674(01)00316-6 (2001).
- 59 Scott, M. S., Boisvert, F. M., McDowall, M. D., Lamond, A. I. & Barton, G. J. Characterization and prediction of protein nucleolar localization sequences. *Nucleic Acids Res* **38**, 7388-7399, doi:10.1093/nar/gkq653 (2010).
- 60 Kersey, P. J. *et al.* The International Protein Index: an integrated database for proteomics experiments. *Proteomics* **4**, 1985-1988, doi:10.1002/pmhc.200300721 (2004).
- 61 Di Matteo, A. *et al.* Structural investigation of nucleophosmin interaction with the tumor suppressor Fbw7 γ . *Oncogenesis* **6**, e379, doi:10.1038/oncsis.2017.78 (2017).
- 62 Luchinat, E. *et al.* Identification of a novel nucleophosmin-interaction motif in the tumor suppressor p14^{arf}. *FEBS J* **285**, 832-847, doi:10.1111/febs.14373 (2018).
- 63 Mitrea, D. M. *et al.* Nucleophosmin integrates within the nucleolus via multi-modal interactions with proteins displaying R-rich linear motifs and rRNA. *Elife* **5**, doi:10.7554/eLife.13571 (2016).
- 64 Duan, T. L., He, G. J., Hu, L. D. & Yan, Y. B. The Intrinsically Disordered C-Terminal Domain Triggers Nucleolar Localization and Function Switch of PARN in Response to DNA Damage. *Cells* **8**, doi:10.3390/cells8080836 (2019).
- 65 Miranda-Saavedra, D. & Barton, G. J. Classification and functional annotation of eukaryotic protein kinases. *Proteins* **68**, 893-914, doi:10.1002/prot.21444 (2007).
- 66 Lee, K.-T. *et al.* Systematic functional analysis of kinases in the fungal pathogen *Cryptococcus neoformans*. *Nature Communications* **7**, 12766, doi:10.1038/ncomms12766 (2016).
- 67 Wang, H. *et al.* Critical evaluation of bioinformatics tools for the prediction of protein crystallization propensity. *Brief Bioinform* **19**, 838-852, doi:10.1093/bib/bbx018 (2018).
- 68 Bateman, A. *et al.* The Pfam protein families database. *Nucleic Acids Res* **32**, D138-141, doi:10.1093/nar/gkh121 (2004).
- 69 Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta* **405**, 442-451, doi:10.1016/0005-2795(75)90109-9 (1975).
- 70 Adebiyi, M. O., Ogunlana, O. O., Adebiyi, E., Fatumo, S. & Rasgon, J. L. The Anopheles gambiae Insecticidal Targets Made Bare by In-silica Analysis. *International Conference on African Development Issues (CU-ICADI) 2015: Biotechnology and Bioinformatics Track*, 32-39 (2015).

- 71 Karczewski, K. J. *et al.* Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv*, 531210, doi:10.1101/531210 (2019).
- 72 Jimenez, R. C. *et al.* Four simple recommendations to encourage best practices in research software. *F1000Res* **6**, doi:10.12688/f1000research.11407.1 (2017).
- 73 Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* **3**, 160018, doi:10.1038/sdata.2016.18 (2016).
- 74 Ison, J. *et al.* Tools and data services registry: a community effort to document bioinformatics resources. *Nucleic Acids Res* **44**, D38-47, doi:10.1093/nar/gkv1116 (2016).
- 75 PDBe-KB consortium. PDBe-KB: a community-driven resource for structural and functional annotations. *Nucleic Acids Res*, doi:10.1093/nar/gkz853 (2019).