ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/332048741

Protein Structure Annotations

Chapter · March 2019

DOI: 10.1007/978-3-030-02634-9_10

CITATIONS
0

READS
82

2 authors:

Mirko Torrisi
University College Dublin
5 PUBLICATIONS **5** CITATIONS

SEE PROFILE

Gianluca Pollastri
University College Dublin
110 PUBLICATIONS **3,671** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Project Protein Disorder Prediction View project

Project Ab Initio Prediction of 1D Protein Structure Annotations View project

All content following this page was uploaded by Mirko Torrisi on 03 May 2019.

The user has requested enhancement of the downloaded file.

Chapter 2: Protein Structure Annotations

Mirko Torrisi, Gianluca Pollastri

Contents

Mirko Torrisi torrisimirko@yahoo.com

Gianluca Pollastri gianluca.pollastri@ucd.ie

School of Computer Science, University College Dublin, Dublin, Ireland

Abstract

This chapter aims to introduce to the specifics of protein structure annotations and their fundamental position in Structural Bioinformatics, and Bioinformatics in general. Proteins are profoundly characterized by their structure in every aspect of their functioning and, while over the last decades there has been a close to exponential growth of known protein sequences, the growth of known protein structures has been closer to linear because of the high complexity and cost of determining them. Thus, protein structure predictors are among the most thoroughly assessed tools in Bioinformatics (in venues such as CASP or CAMEO) because they allow the structural study of proteins on a large scale. This chapter presents the key types of protein structure annotation and the methods and algorithms for predicting them, with the aim to give both a historical perspective on their development and a snapshot of their current state-of-the-art. From one-dimensional protein annotations – i.e., secondary structure, solvent accessibility and torsion angles – to more complex and informative two-dimensional protein abstractions – i.e., contact maps –, both mature and currently developing methods for protein structure annotations are introduced. The aim of this overview is to facilitate the adoption and development of state-of-the-art protein structural predictors. Particular attention is given to some of the best performing and freely available web servers and standalone programs to predict protein structure annotations.

Keywords: protein structure annotations, secondary structure prediction, solvent accessibility prediction, torsional angles prediction, contact map prediction

Protein Structure Annotations

Proteins hold a unique position in structural Bioinformatics. In fact, more so than other biological macromolecules such as DNA or RNA, their structure is directly and profoundly linked to their function. Their cavities, protuberances, and their overall shapes determine with what and how they will interact, and, therefore, the roles assumed in the hosting organism. Unfortunately, the complexity, wide variability, and ultimately the sheer number of diverse structures present in nature, make the characterization of proteins extremely expensive, and complex. For this reason, considerable effort has been spent on predicting protein structures by computational means, either directly, or in the form of abstractions that simplify the prediction while still retaining structural information. These abstractions, or protein structure annotations, may be one-dimensional when they can be represented by a string or a sequence of numbers, typically of the same length as the protein's primary structure (the sequence of its amino acids). This is the case, for instance, of secondary structure (SS) or solvent accessibility (SA). Another important class of abstractions is composed of two-dimensional properties, that is, features of pairs of amino acids (AA) or SS, such as contact and distance maps, disulphide bonds, or pairings of strands into β-sheets.

Machine Learning (ML) techniques have been extensively used in Bioinformatics, and in Structural Bioinformatics in particular. The abundance of freely available data – such as the Protein Data Bank (PDB)[1], and the Universal Protein Resource[2] –, and their complexity, make Proteins an ideal domain where to apply the most recent, and sophisticated ML techniques, such as Deep Learning[3]. Nonetheless, there are pitfalls to avoid and best practices to follow to correctly train and test any ML method on protein sequences[4].

Deep Learning is a collection of methods and techniques to efficiently train nuanced parametric models such as Neural Networks (NN) with multiple hidden layers [5]. These layers contain hierarchical representations of the features of interest extracted from the input. NN are the de facto standard ML method to predict protein structure annotations. They have a central role at the two most important academic assessments of protein structure predictors: CASP and CAMEO[6]. Thus, they are widely used to predict protein one-dimensional and two-dimensional structural abstractions.

A typical predictor of protein structure annotations will first look for evolutionary information (PSI-BLAST is commonly used for this task), then will encode the information found, following this will run a ML method (usually a NN) on the encoded information and finally will process the output into a human-readable format. Differently from *ab initio* methods, template-based predictors directly exploit structural information of resolved proteins alongside evolutionary information[7].

Position-Specific Iterated Basic Local Alignment Search Tool (PSI-BLAST)[8] is the de facto standard algorithm, released with the BLAST+ suite, to address protein alignment. In particular, it is commonly used in substitution of BLAST, whenever remote homologues have relevance. PSI-BLAST executes a BLAST call to find similar proteins in a given database, then it either uses the resulting Multiple Sequence Alignment (MSA) to construct a Position Specific Score Matrix (PSSM), or outputs the MSA itself. The entire process is usually iterated few times using the last PSSM as query for the next iteration — in order to improve the PSSM, and, thus, maximize the sensitivity of the method. The trade-off for increasing the number of iterations, and the sensitivity of the method, is a higher likelihood of corrupting the PSSM, including false positive queries into it[9]. For this reason, and the nature itself of the tool, it is fundamental to consider PSI-BLAST as a predicting tool and not as an exact algorithm[10].

HHblits[11] is a 2011 algorithm to address protein alignment. It focuses on fast iterations, and high precision and recall. It obtains these gains by adopting Hidden Markov Models (HMM) to represent both query and database sequences. The overall approach resembles the PSI-BLAST one — except that HMM rather than PSSM are the central entity. In fact, the heuristic algorithm looks for similar proteins in the HMM database at first. Then, it either uses the resulting HMM to improve the HMM query, and iterate, or outputs the MSA found with the last HMM. The same trade-off between number of iterations and likelihood of corrupting the HMM stands for HHblits as it does for PSI-BLAST.

In this chapter we review the main abstractions of protein structures. Namely, SS, SA, torsional angles (TA) and distance/contact maps. For each of them we describe an array of ML algorithms that have been used for their characterisation, point to a set of public tools available to the research community, including some that have been developed in our laboratory, and try to outline the state of the art in their prediction. These structure annotations are complementary with one another as they look at proteins from different views. That said, some annotations received far more interest from the bioinformatics community than others, for reasons such as simplicity or the intrinsic nature of the feature itself. We focus more on these well-assessed annotations, keeping in mind that the main function of protein structure annotations is to facilitate the understanding of the very core of any protein: the three-dimensional structure.

The PSSM built by PSI-BLAST, or the HMM built by HHblits, or the encoded MSA built by either PSI-BLAST or HHblits, are generally used as inputs to a protein feature predictor. Different releases of the database used to find evolutionary information may lead to different outcomes. Normally, a computer able to look for evolutionary information (thus, execute PSI-BLAST or HHblits calls successfully) has the right hardware to run the standalones here presented with no problem.

All the predictors described below offer a web server, are free for academic use and provide licenses for commercial users at the time of writing. The web servers described return a result(/prediction) in anything between a few minutes and a few hours.

Secondary Structure

SS prediction is one of the great historical challenges in Bioinformatics[12], [13]. Its history started in 1951, when Pauling and Corey predicted for the first time the existence of what were later discovered to be the two most common SS conformations: α-helix and β-sheet[14]. Notably, the very first high-resolution protein structure was determined only in 1958 (and led to a Nobel Prize to Kendrew and Perutz)[15], [16]. These early successes motivated the first generation of protein predictors, which were able to extrapolate statistical propensities of single AA (or residue) towards certain conformations[17]. The slow but steady growth of available data and more insights on protein structure led to the second generation of predictors, which expanded the input to segments of adjacent residues (3-51 AA) to gather more useful information, and assessed many available theoretical algorithms on SS[12]. In the 90s, more available computational power and data allowed the development and implementation of more advanced algorithms, able to look for and take advantage of evolutionary information[13]. Thus, the third generation of SS predictors was the first able to predict at better than 70% accuracy[18], efficiently exploit PSI-BLAST[19] and implement deep NN[20]. In 2002, SS was removed from CASP since the few and relatively short targets assessed at the venue were not considered statistically sufficient to evaluate the mature methods available[21].

The intrinsic nature of SS, being an intermediate structural representation between primary and tertiary structure, makes it a strategic and fundamental one-dimensional protein feature. It is often adopted as intermediate step towards more complex and informative features (i.e., contact maps[22]–[24], the recognition of protein folds[25] and protein tertiary structure[26]). In other words, a high-quality SS prediction can greatly help to understand the nature of a protein and lead to a better prediction of its structure. For example, SS regularities characterize the proteins in a common fold[27].

The theoretical limit of SS prediction is usually set at 88-90% accuracy per AA[13]. This limit is mainly derived from the disagreement on how to assign SS and from the intrinsic dynamic nature of protein structure — i.e. the protein structure changes according to the fluid in which the protein is immersed. In particular, Define Secondary Structure of Proteins (DSSP)[28], the gold standard algorithm to assign SS given the atomic-resolution coordinates of the protein, agrees with the PDB descriptions around 90.8% of the time[29]. While DSSP aims to provide an unambiguous and physically meaningful assignment, the PDB represents the ground truth in structural proteomics[1].

All the SS predictors described in this chapter exploit different architectures of NN to perform their predictions. The list of AA composing the protein of interest is the only input required. The SS is often classified in three-states – i.e., helices, sheets and coils –, although the DSSP identifies a total of 8 different classes. Because of the higher difficulty of the task, compounded also by the rare occurrence of certain classes – i.e., π-helix and β-bridge –, only 3 of the predictors presented here (Porter5, RaptorX-Property and SSpro) can predict in both three-states and eight-states. The DSSP classification of SS in eight-states is the following one:

- G = 3-turn helix (3_{10}-helix), minimum length 3 residues;
- H = 4-turn helix (α-helix), minimum length 4 residues;
- I = 5-turn helix (π-helix), minimum length 5 residues;
- T = hydrogen bonded turn (3, 4 or 5 turn);
- E = extended strand (in β-sheet conformation), minimum length 2 residues;
- B = residue in isolated β-bridge (single pair formation);
- S = bend (the only non-hydrogen-bond based assignment);
- C = coil (anything not in the above conformations).

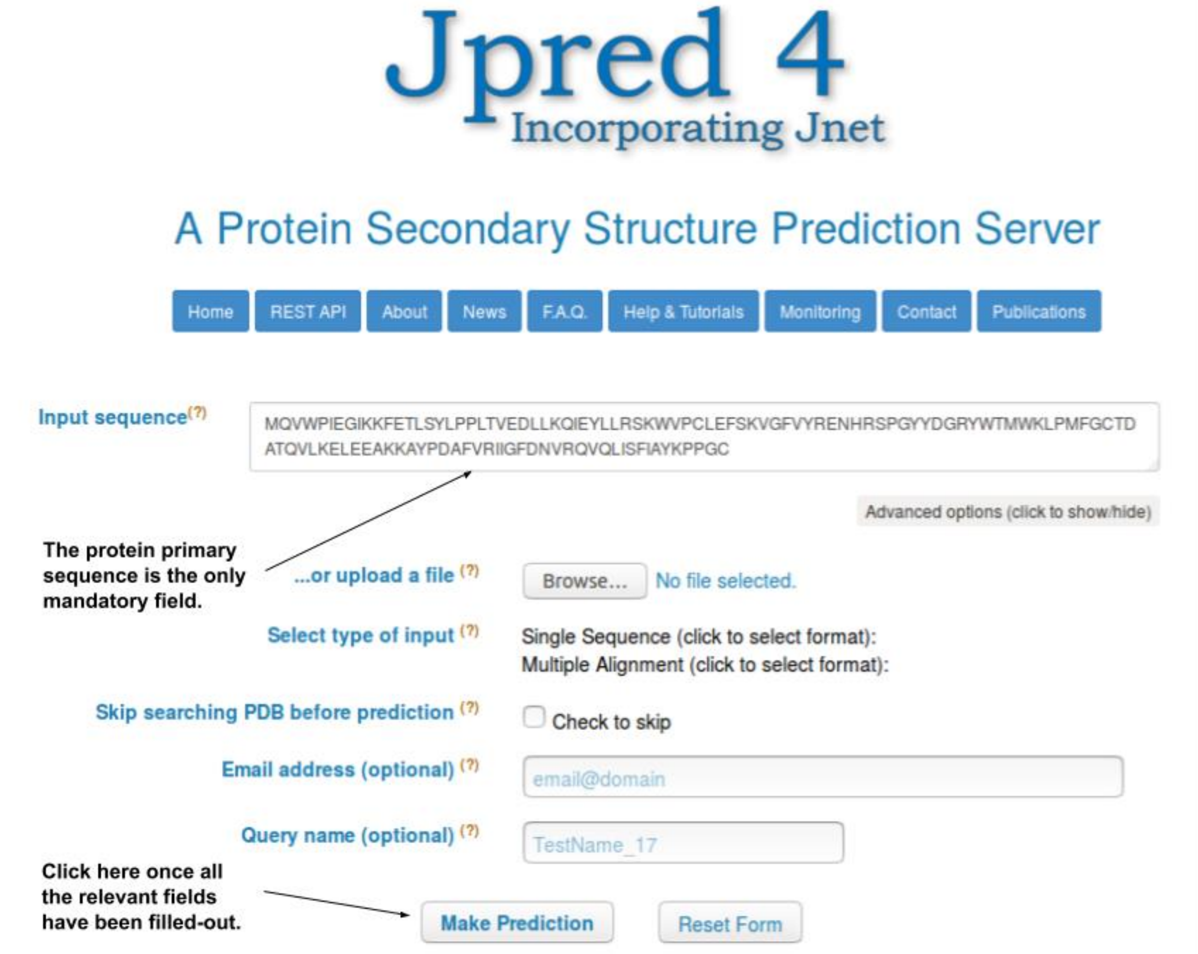
When SS is classified in three states, the first three (G, H, I) are generally considered helices, E and B are classified as strands and anything else as coils. SS prediction is evaluated looking at the rate of correctly classified residues (per class) – i.e., Q3 or Q8 for three- or eight-states prediction, respectively – or at the segment overlap score (SOV) – i.e., the overlap between the predicted and the real segments of SS[30] –, for a more biological viewpoint. The best performing ab initio SS predictors are able to predict three-state SS close to 85% Q3 accuracy and SOV score.

The table below gathers name, web server and notes on special features of every SS predictor presented in this chapter. A standalone – i.e., downloadable version that can run on a local machine – is currently available for all of them.

Name	**Web Server**	**Notes**
Jpred[31]	http://www.compbio.dundee.ac.uk/jpred4/	HHMer, MSA as input, API
PSIPRED[19]	http://bioinf.cs.ucl.ac.uk/psipred/	BLAST, cloud version, MSA as input
Porter[32]	http://distilldeep.ucd.ie/porter/	three- or eight-states, HHblits or PSI-BLAST, light standalone
RaptorX-Property[33]	http://raptorx.uchicago.edu/StructurePropertyPred/predict/	three- or eight-states, no PSI-BLAST (only HHblits), option for no evolutionary information
SPIDER3[34]	http://sparks-lab.org/server/SPIDER3/	Numpy or Tensorflow, HHblits and PSI-BLAST

SSpro[35]	http://scratch.proteomics.ics.uci.edu/	three- or eight-states, BLAST, template-based

Jpred



Input sequence(?) MQVWPIEGIKKFETLSYLPPLTVEDLLKQIEYLLRSKWVPCLEFSKVGFVYRENHRSPGYYDGRYWTMWKLPMFGCTD ATQVLKELEEAKKAYPDAFVRIIGFDNVRQVQLISFIAYKPPGC

The protein primary sequence is the only mandatory field.

...or upload a file (?) No file selected.

Select type of input (?) Single Sequence (click to select format): Multiple Alignment (click to select format):

Skip searching PDB before prediction (?) □ Check to skip

Email address (optional) (?) email@domain

Query name (optional) (?) TestName_17

Click here once all the relevant fields have been filled-out.

Figure 1: The homepage of Jpred4. The input sequence is the only requirement while more options are made available.

Jpred is an SS predictor which was initially released in 1998[36]. Jpred4[31], the last available version, has been released in 2015 to update HMMer[37] and the internal algorithm (a NN). Jpred4 relies on both PSI-BLAST and HMMer to gather evolutionary information, generating a PSSM and a HMM, respectively. It then predicts SS in three-states, along with SA and coiled-coil regions. Jpred4 aims to be easily usable also from smartphones and tablets. F.A.Q. and tutorials are available on its website.

The web server of Jpred4 is available at http://www.compbio.dundee.ac.uk/jpred4/. It requires a protein sequence in either FASTA or RAW format. Using the advanced options, it is also possible to submit multiple sequences (up to 200) or MSA as files. An email address and a JobID can be optionally provided. When a single sequence is given, Jpred4 looks for similar protein sequences in the PDB[1] and lists them when found. Checking a box it is possible to skip this step and force an ab-initio prediction. Jpred4 relies on a version of UniRef90[2] released in July 2014, while the PDB is regularly updated.

The result page is automatically shown and offers a graphical summary of the prediction along with links to possible views of the result in HTML (simple or full), PDF and Jalview[38] (in-browser or not). It is also possible to get an archive of all the files generated or navigate through them in the browser. If an email address is submitted, a link to the result page and a summary containing the query, predicted SS and confidence per AA will be sent. The full result, made available as HTML or PDF, lists the ID of similar sequences used at prediction time, the final and intermediate predictions for SS, the prediction of coiled-coil regions, the prediction of SA with three different thresholds (0, 5 and 25% exposure) and the reliability of such predictions.

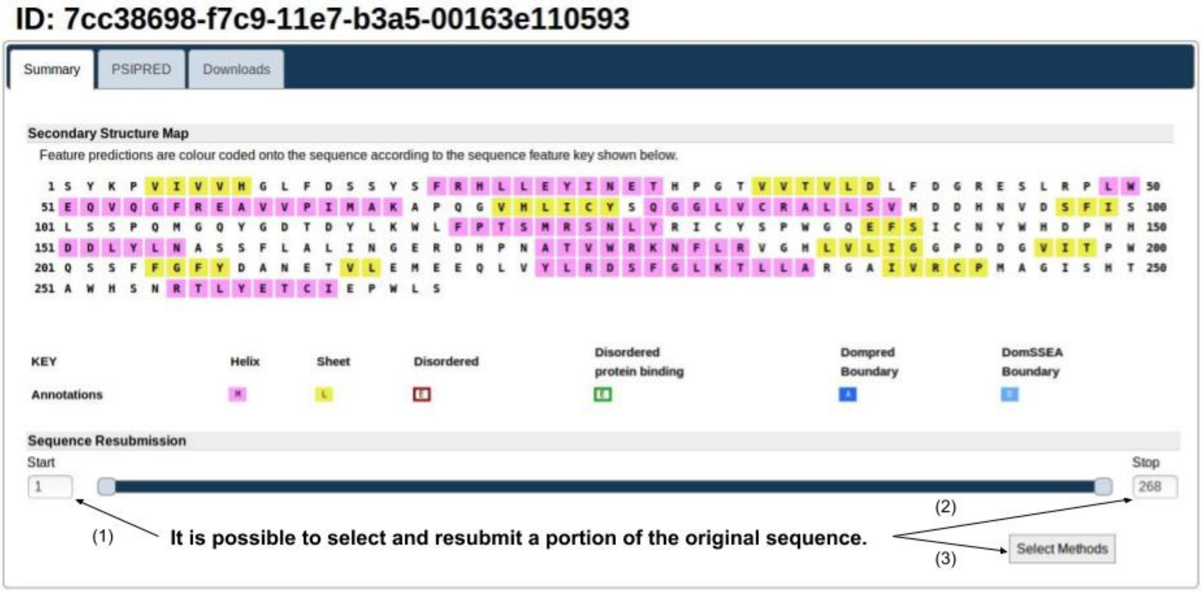
Jpred4 is not released as standalone but it is possible to submit, monitor and retrieve a prediction using the command line software available at http://www.compbio.dundee.ac.uk/jpred4/api.shtml. A second package of scripts is made available at the same address to facilitate the submission, monitoring and retrieving of multiple protein sequences. More instructions and examples on how to use the command line software are presented on the same page.

PSIPRED

PSIPRED is a high quality SS predictor freely available since 1999[19]. Its last version (v4.01) has been released in 2016. PSIPRED exploits the PSSM of the protein to generate its prediction by Neural Networks. Like SSpro (described below), it recommends the implementation of the legacy BLAST package (abandoned in 2011) to collect evolutionary information. The BLAST+ package (the active development of BLAST) fixes multiple bugs and provides improvements and new features, but scales by 10 and rounds the PSSM, thus provides less informative outputs for PSIPRED. BLAST+ is experimentally supported by PSIPRED.

The web server of PSIPRED[39], called *The PSIPRED Protein Sequence Analysis Workbench*, runs a 2012 release of PSIPRED (v3.3) and can be found at http://bioinf.cs.ucl.ac.uk/psipred/. A single sequence, (or its MSA) and a short identifier are expected as input. Optionally, an email address can be inserted to receive a confirmation email (with link to the result) when the prediction is ready. Several prediction methods (for other protein features) can be chosen. The default choice (picking only PSIPRED) is sufficient to predict the SS. If the submission proceeds successfully, a courtesy page will be shown until the result is ready.

The result page, organised in tabs, shows the list of AA composing the analysed protein (the query sequence) and the predicted SS class (using different colours). From the same tab, it is possible to select the full query sequence, or a subsequence, to pass it to one of the predictor methods available on the PSIPRED Workbench. The predicted SS is presented in the tab called *PSIPRED* using a diagram. In the same diagram, the confidence of each prediction and the query sequence are included. The *Downloads* tab, the last one, allows the download of the information in the diagram as text or PDF or postscript or of all three versions.



ID: 7cc38698-f7c9-11e7-b3a5-00163e110593

Secondary Structure Map

Feature predictions are colour coded onto the sequence according to the sequence feature key shown below.

```
  1 S Y K P V I V V H G L F D S S Y S F R H L L E Y I N E T H P G T V V T V L D L F D G R E S L R P L W  50
 51 E Q V Q G F R E A V V P I M A K A P Q G V H L I C Y S Q G G L V C R A L L S V M D D H N V D S F I S 100
101 L S S P Q M G Q Y G D T D Y L K W L F P T S M R S N L Y R I C Y S P W G Q E F S I C N Y W H D P H H 150
151 D D L Y L N A S S F L A L I N G E R D H P N A T V W R K N F L R V G H L V L I G G P D D G V I T P W 200
201 Q S S F F G F Y D A N E T V L E M E E Q L V Y L R D S F G L K T L L A R G A I V R C P M A G I S H T 250
251 A W H S N R T L Y E T C I E P W L S
```

KEY	Helix	Sheet	Disordered	Disordered protein binding	Dompred Boundary	DomSSEA Boundary
Annotations	H	L	E	E	A	B

Sequence Resubmission

Start: 1

Stop: 268

It is possible to select and resubmit a portion of the original sequence.

Figure 2: A typical result page of PSIPRED web server. All the AA are listed and coloured according to the predicted SS class.

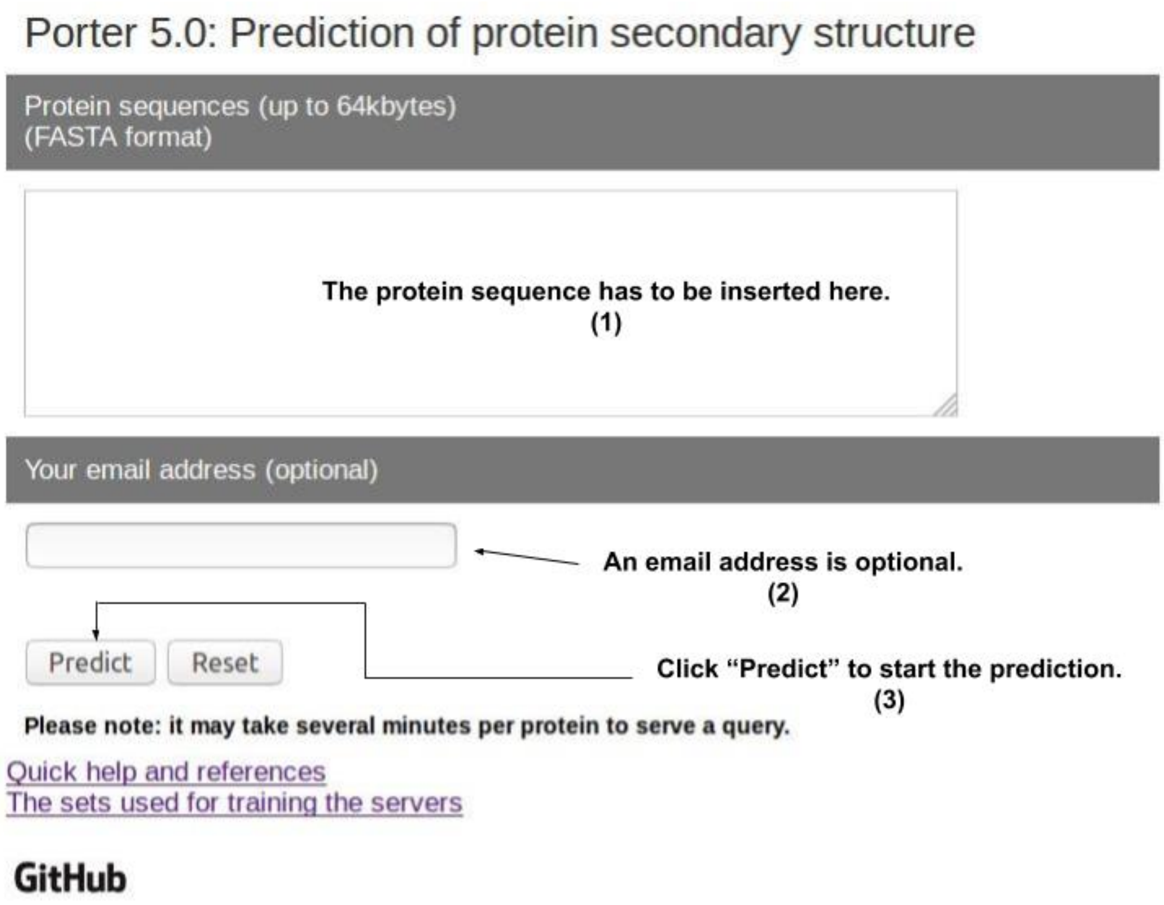
The last release of PSIPRED is typically available as a standalone at http://bioinfadmin.cs.ucl.ac.uk/downloads/psipred/. Once the standalone has been downloaded and extracted, it is sufficient to follow the instructions in the README to perform predictions on any machine. The output will be generated in text format only as horizontal or vertical format. The latter will contain also the individual confidence per helix, strand and coil. Notably, the results obtained from the standalone may very well differ from those obtained from the PSIPRED Workbench. The latter does not implement the last PSIPRED release, at the time of writing.

In 2013, a preliminary package (v0.4) has been released to run PSIPRED on Apache Hadoop. Hadoop is an open-source software to facilitate distributed processing in computer clusters. Although this PSIPRED package is intended as an alpha build, instructions to install it on Hadoop and on AWS (the cloud service of Amazon) are provided. This package does not contain any standalone of PSIPRED. Thus, it is an interface to run the selected PSIPRED release on Hadoop. It can be downloaded at http://bioinfadmin.cs.ucl.ac.uk/downloads/hadoop/.

Porter

Porter is a high quality SS predictor which has been developed starting in 2005[32] and improved since then[7], [40]. Porter is built on carefully tuned and trained ensembles of cascaded Bidirectional Recurrent Neural Networks[20]. It is typically built on very large datasets, which are released as well. Its last release (v5) is available as web server and standalone[41]. Differently from PSIPRED, it implements BLAST+ to gather evolutionary information. To maximise the gain obtained from evolutionary information, it also adopts HHblits alongside PSI-BLAST. Porter5 is one of the three SS predictors presented here that are able to predict both three-states and eight-states SS.

The web server can be found at http://distilldeep.ucd.ie/porter/. The basic interface asks for protein sequences (in FASTA format) and for an optional email address. Up to 64KB of protein sequences can be submitted at the same time, which approximately corresponds to 200 average proteins. Differently from other SS web servers, there is no limit of total submissions. The confirmation page will contain a summary of the job, the server load (how many jobs are to be processed) and the URL to the result page. It is automatically refreshed every minute.



Porter 5.0: Prediction of protein secondary structure

Protein sequences (up to 64kbytes)
(FASTA format)

The protein sequence has to be inserted here.
(1)

Your email address (optional)

An email address is optional.
(2)

Click “Predict” to start the prediction.
(3)

Please note: it may take several minutes per protein to serve a query.

Predict
Reset
Quick help and references
The sets used for training the servers
GitHub

Figure 3: The input form of Porter5. Around 200 proteins can be submitted at once in FASTA format.

The detailed result page will show the query, the SS prediction and the individual confidence. In other words, the same information shown by the PSIPRED Workbench is given in text format. The time to serve the job is shown as well. Optionally, if an email address has been inserted, all the information in the result page is sent by email. Thus, it can potentially be retrieved at any time. It is possible to predict SS and other protein structure annotations (one-dimensional or not) submitting one job at http://distillf.ucd.ie/distill/.

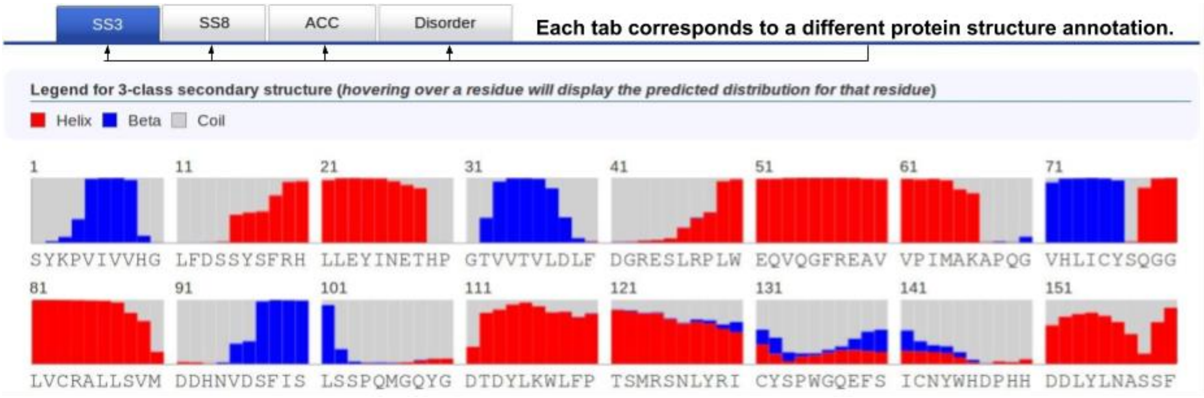
The very light standalone of Porter5 (7 MB) is available at http://distilldeep.ucd.ie/porter/. It is sufficient to extract the archive on any computer with Python3, HHblits and PSI-BLAST to start predicting any SS. Using the parameter --*fast,* it is possible to avoid PSI-BLAST and perform faster but generally slightly less accurate predictions. When the prediction in three-states and eight-states completes successfully, it is saved in 2 different files. Each file shows the query, the predicted SS and the individual confidence per class. The datasets adopted for training and testing purposes are available at the same address.

RaptorX-Property

RaptorX-Property, released in 2016, is a collection of methods to predict one-dimensional protein annotations[33]. Namely, SS, SA and disorder regions are predicted from the same suite. The SS is predicted in both three-states and eight-states, as with Porter5 and SSpro. At the cost of lower accuracy, evolutionary information can be avoided to perform faster predictions. Its last release substitutes PSI-BLAST with HHblits to get faster protein profiles.

The web server of RaptorX-Property is available at http://raptorx.uchicago.edu/StructurePropertyPred/predict/. Jobname and email address are recommended but not required. Query sequences can be uploaded directly from one's machine. Otherwise, up to 100 protein sequences (in FASTA format) can be passed at the same time through the input form. The system allows up to 500 pending (sequence) predictions at any time. The current server load, shown in the sidebar, tells the pending jobs to complete.

Once the job has been submitted, a courtesy page will provide the URL to the result page, how many pending jobs are ahead and the JobID. Less priority is given to intensive users. The jobs submitted in the previous 60 days are retrievable clicking on "My Jobs". Once the prediction is performed, the result page will show a summary of it using coloured text. At the bottom of the page, the same information is organised in tabs, one tab per feature predicted (SS in 3 and 8-state, SA and disorder). The individual confidence is provided in the tabs. All this information is sent by email (in txt and rtf format), if an email address has been provided. Otherwise, it can be downloaded clicking the specific button.



Each tab corresponds to a different protein structure annotation.

Legend for 3-class secondary structure (*hovering over a residue will display the predicted distribution for that residue*)

🟥 Helix 🟦 Beta ⬜ Coil

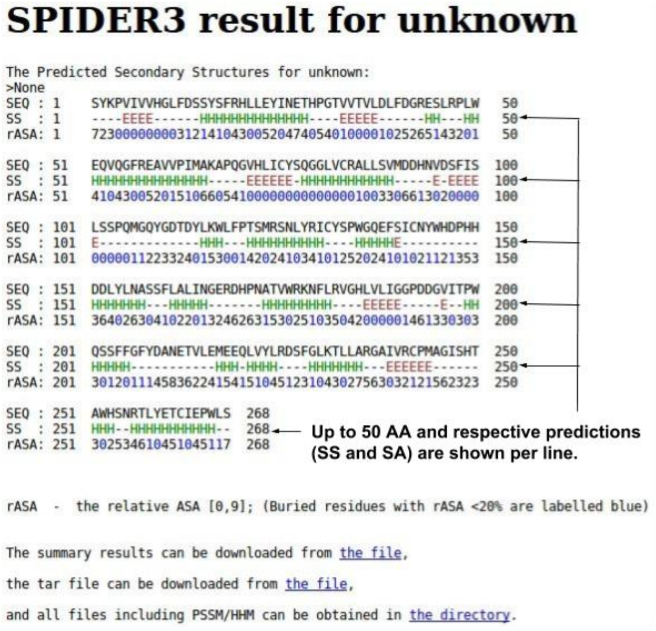
Figure 4: A partial view of the result page of RaptorX-Property. Each bar in the charts represents the individual confidence.

The last standalone of RaptorX-Property (v1.01) can be downloaded at http://raptorx.uchicago.edu/download/. Once it has been extracted, it is sufficient to read and follow the instruction in README to predict SS, SA and disorder regions on one's own machine. As in the web server, it is possible to use or not sequence profiles and the results are saved in txt and rtf format. The disk-space required is relatively considerable, 347 MB at the time of writing, almost fifty times the storage required by Porter5.

SPIDER3

SPIDER3 is the second version of a recent SS predictor first released in 2015[42]. Its last release is composed by 2 NN, the first of which predicts SS while the second predicts backbone angles, contact numbers and SA[34]. It internally represents each AA using 7 representative physio-chemical properties[43]. Like Porter5, it implements both HHblits and PSI-BLAST to look for more evolutionary information. SPIDER3 is also described in sections Solvent Accessibility and Torsion Angles, respectively.

The web server of SPIDER3 is available at http://sparks-lab.org/server/SPIDER3/. An email address is required when multiple sequences are submitted, or to receive a summary of the prediction. Otherwise, the query sequence is sufficient to submit the job and obtain an URL to the result page. The web server allows up to 100 protein sequences (in FASTA format) at a time and accepts optional JobID. To prevent duplicates, it is possible to visualize the queue of jobs submitted from one's IP address. The result page presents the query sequence and the predicted SS and SA, in a simple and colour-coded text format. In the same page, it is possible to download a summary (containing the same information) or an archive with the 4 features predicted and the individual confidence for SS. There is also a link to a temporary directory containing all the files created during the prediction, including the HMM and the PSSM.



SPIDER3 result for unknown

```
The Predicted Secondary Structures for unknown:
>None
SEQ : 1    SYKPVIVVHGLFDSSYSFRHLLEYINETHPGTVVTVLDLFDGRESLRPLW   50
SS  : 1    ----EEEE------HHHHHHHHHHHHHH----EEEEE------HH---HH   50
rASA: 1    723000000000312141043005204740540100001025265143201   50

SEQ : 51   EQVQGFREAVVPIMAKAPQGVHLICYSQGGLVCRALLSVMDDHNVDSFIS  100
SS  : 51   HHHHHHHHHHHHHHH-----EEEEEE-HHHHHHHHHHHH-----E-EEEE  100
rASA: 51   41043005201510660541000000000000000100330661302000 0  100

SEQ : 101  LSSPQMGQYGDTDYLKWLFPTSMRSNLYRICYSPWGQEFSICNYWHDPHH  150
SS  : 101  E-------------HHH---HHHHHHHHHH----HHHHHE----------  150
rASA: 101  00000112233240153001420241034101252024101021121353  150

SEQ : 151  DDLYLNASSFLALINGERDHPNATVWRKNFLRVGHLVLIGGPDDGVITPW  200
SS  : 151  HHHHHHH---HHHHH-------HHHHHHHHH----EEEEE-----E--HH  200
rASA: 151  36402630410220132462631530251035042000001461330303  200

SEQ : 201  QSSFFGFYDANETVLEMEEQLVYLRDSFGLKTLLARGAIVRCPMAGISHT  250
SS  : 201  HHHHH-----------HHH-HHHH----HHHHHHH---EEEEEE------  250
rASA: 201  30120111458362241541510451231043027563032121562323  250

SEQ : 251  AWHSNRTLYETCIEPWLS  268
SS  : 251  HHH--HHHHHHHHHHH--  268
rASA: 251  302534610451045117  268

rASA  -  the relative ASA [0,9]; (Buried residues with rASA <20% are labelled blue)

The summary results can be downloaded from the file,

the tar file can be downloaded from the file,

and all files including PSSM/HHM can be obtained in the directory.
```

Up to 50 AA and respective predictions (SS and SA) are shown per line.

Figure 5: An output example of SPIDER3.

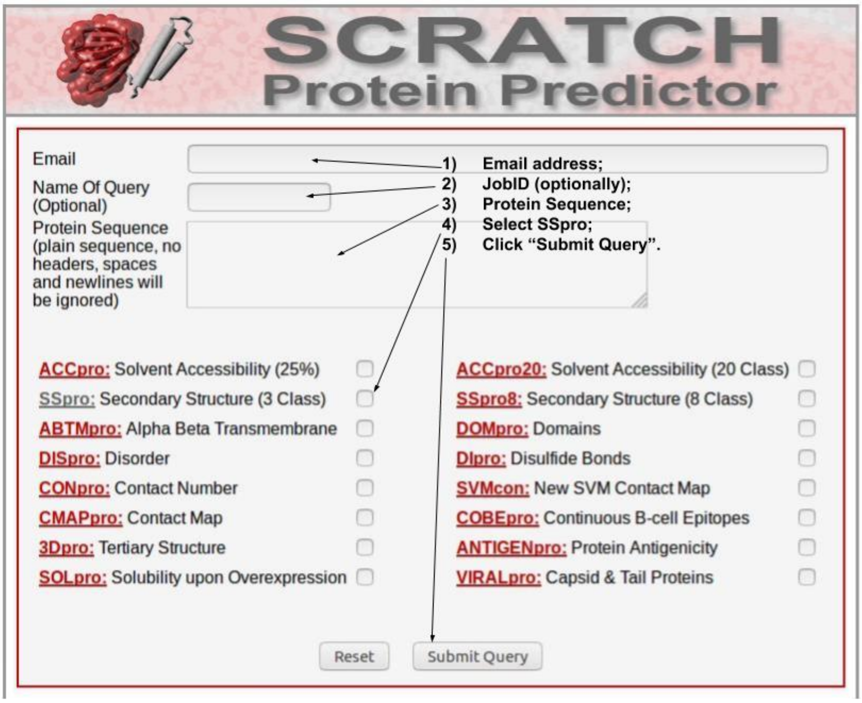
The standalone of SPIDER3, and the dataset used to train and test it, can be downloaded at http://sparks-lab.org/server/SPIDER3/. The main prerequisite is to install a python library of choice between Numpy and Tensorflow r0.11 (an older version). As for Porter5, it is then sufficient to install HHblits and PSI-BLAST to perform SS prediction on one's machine. The outcome of SS, SA, torque angles and contact number prediction will be saved in different columns of just one file. The storage required is 101MB and 117 MB, respectively, without considering the library of choice.

SSpro

SSpro is a historical SS predictor developed starting in 1999[20], [35]. Similarly to PSIPRED, it implements the BLAST package rather than the more recent BLAST+. The last version of SSpro (v5) has been released in 2014, together with ACCpro (see Solvent Accessibility, ACCpro), and performs template-based SS predictions[35]. More specifically, it exploits PSI-BLAST to look for homologues at both sequence and structure level[7]. In other words, SSpro v5 has an additional final step in which it looks for similar proteins in the PDB.

SSpro is available at http://scratch.proteomics.ics.uci.edu/ as part of the SCRATCH protein predictor[44]. SS is among the several (one-dimensional or not) protein features predictable on SCRATCH. Like Porter5 and RaptorX-Property, it is possible to predict both three-states (SSpro) and eight-states (SSpro8) predictions. Once SSpro or SSpro8 is selected, an email is required and optionally a JobID. Only one protein (of up to 1500 residues) can be submitted at a time. There are 5 total slots in the job queue per user. Once ready, the result of the prediction will be sent by email only. It will contain: the JobID, the query sequence, the predicted SS (in three or eight classes) and a link to the explanation of the output format.

The standalone of the last SSpro (v5.2) and ACCpro (described in section Solvent Accessibility) compose the SCRATCH suite of 1D predictors available at http://download.igb.uci.edu/. SCRATCH v1.1 is released with all the prerequisites to set-up and run SSpro. The BLAST package and the databases with both sequences and structural information are included. Thus, the amount of disk-space needed to download and extract SCRATCH v1.1 is considerable (5.7 GB, 97MB without databases).



SCRATCH
Protein Predictor

Email
Name Of Query (Optional)
Protein Sequence (plain sequence, no headers, spaces and newlines will be ignored)

1) Email address;
2) JobID (optionally);
3) Protein Sequence;
4) Select SSpro;
5) Click "Submit Query".

ACCpro: Solvent Accessibility (25%)	□	ACCpro20: Solvent Accessibility (20 Class)	□
SSpro: Secondary Structure (3 Class)	□	SSpro8: Secondary Structure (8 Class)	□
ABTMpro: Alpha Beta Transmembrane	□	DOMpro: Domains	□
DISpro: Disorder	□	DIpro: Disulfide Bonds	□
CONpro: Contact Number	□	SVMcon: New SVM Contact Map	□
CMAPpro: Contact Map	□	COBEpro: Continuous B-cell Epitopes	□
3Dpro: Tertiary Structure	□	ANTIGENpro: Protein Antigenicity	□
SOLpro: Solubility upon Overexpression	□	VIRALpro: Capsid & Tail Proteins	□

Figure 6: A view of SCRATCH Protein Predictor.

Solvent Accessibility

SA describes the degree of accessibility of a residue to the solvent surrounding the protein. SA is second only to SS among extensively studied and predicted one-dimensional protein structure annotations. The effort invested into SA predictors has been significant from the early 90s and highly motivated from the successes obtained developing the third generation of SS predictors[45]. In fact, similarly to SS prediction but sometimes with some time-delay, mathematical and statistical methods[46], NN[47], evolutionary information[48] and deep NN[49] have been increasingly put to work to predict SA.

Although SA is less conserved than SS in homologous sequences[47], it is typically adopted in parallel with SS in many pipelines towards more complex protein structure annotations such as CM – e.g., SA and SS are predicted for any CM predictor described in section Contact Maps [22], [23], [50], [51] –, protein fold recognition[25] and protein tertiary structure[52]. Notably, a strong (negative) correlation of -0.734 between SA and contact numbers has been observed by Yuan[53] and is motivating the development of predictors for contact number as a possible alternative to SA predictors[54].

Though there are promising examples of successful NN predictors considering adjacent AA to predict SA since the 90s[48], different methods such as linear regression[55] or substitution matrices[45] have been assessed but the state-of-the-art has been represented by deep NN since 2002[49]. Thus, all the SA predictors described below (and summarised in the table) implement deep NN[33]–[35], [40] predicting SA as anything between a two-state problem – i.e., buried and exposed with an average two-state accuracy greater than 80% – to twenty states.

SA has been typically measured as accessible surface area (ASA) – i.e., the protein's surface exposed to interactions with the external solvent. ASA is usually obtained normalizing the relative SA value observed by the maximum possible value of accessibility for the specific residue according to the DSSP[28]. The ASA of a protein can be visualized with ASAview, a tool developed in 2004 that requires real values extracted from the PDB or coming from predicted ASA[56]. More recently, a different approach to measuring the SA, called half-sphere exposure (HSE), has been designed by

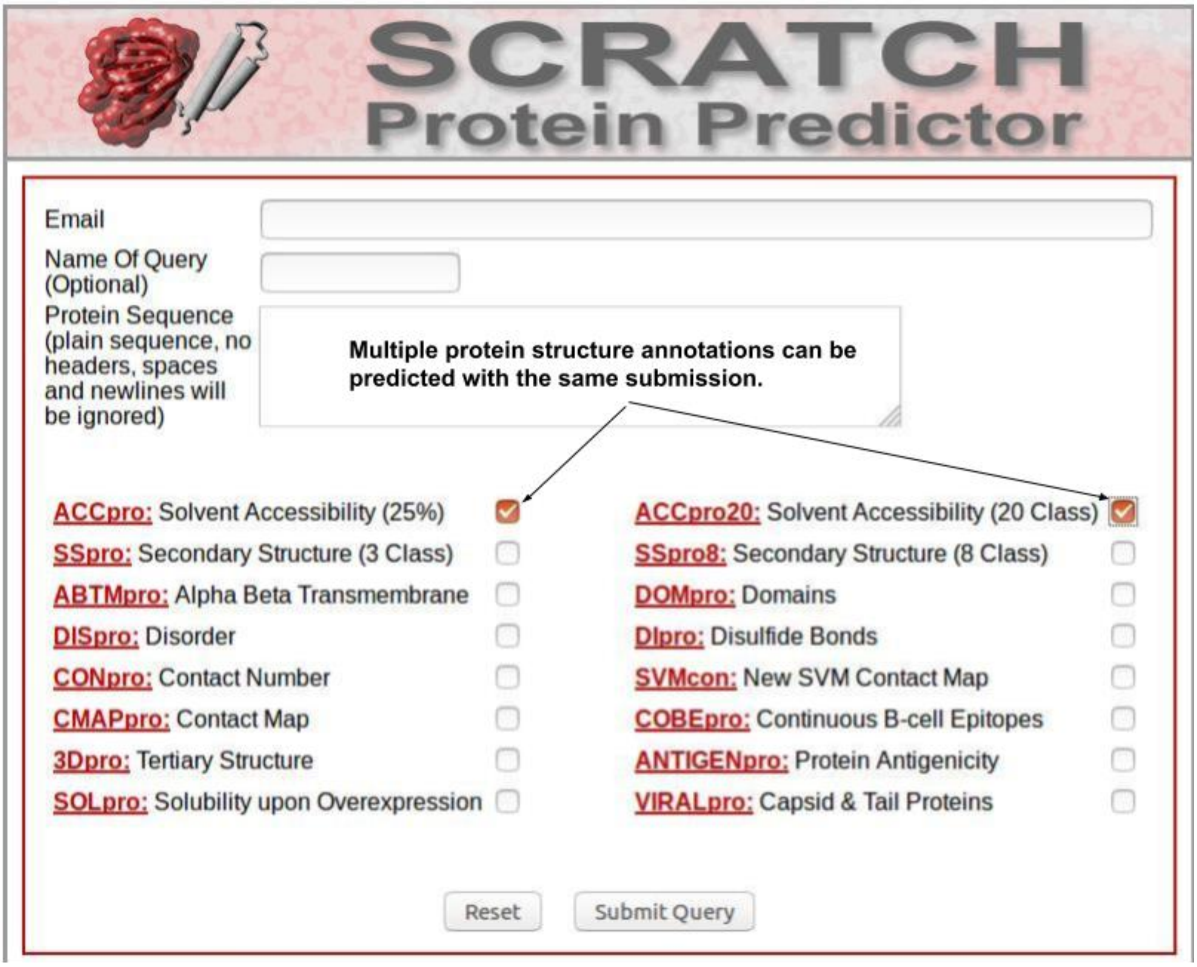
Hamelryck[57]. The idea is to split in half the sphere surrounding the Cα atom along the vector of Cα-C_β atoms aiming to provide a more informative and robust measure[57]. SPIDER3 can predict both HSE and ASA using real numbers[34].

Name	Web Server	Notes
ACCpro[20]	http://scratch.proteomics.ics.uci.edu/	two-state or twenty-states, BLAST, template-based
PaleAle[7]	http://distilldeep.ucd.ie/paleale/	four-states, HHblits or PSI-BLAST, light standalone
RaptorX-Property[33]	http://raptorx.uchicago.edu/StructurePropertyPred/predict/	three-states, no PSI-BLAST (only HHblits), option for no evolutionary information
SPIDER3[34]	http://sparks-lab.org/server/SPIDER3/	HSE and ASA in $\mathbb{R}$, Numpy or Tensorflow, HHblits and PSI-BLAST

ACCpro

ACCpro is a historical SA predictor initially released in 2002[49]. Since then, it has been developed in parallel with SSpro (see Secondary Structure, SSpro) and last updated to its v5 in 2014, adding support for template-base predictions[35]. Thus, like SSpro, ACCpro adopts the legacy BLAST to look for evolutionary information at both sequence and structure level. ACCpro predicts whether each residue is more exposed than 25% or not, while ACCpro20, an extension of ACCpro, distinguishes twenty-states from 0-95% with incremental steps of 5% – i.e., ACCpro classifies twenty classes, starting from 0-5% to 95-100% of SA.

The web server of ACCpro and ACCpro20 is available at http://scratch.proteomics.ics.uci.edu/ as part of SCRATCH[44]. Once an email and the sequence to predict have been inserted, it is possible to select ACCpro or ACCpro20 or any of the available protein predictors. More in Secondary Structure, SSpro.



Email

Name Of Query (Optional)

Protein Sequence (plain sequence, no headers, spaces and newlines will be ignored)

Multiple protein structure annotations can be predicted with the same submission.

ACCpro: Solvent Accessibility (25%) ☑
SSpro: Secondary Structure (3 Class) ☐
ABTMpro: Alpha Beta Transmembrane ☐
DISpro: Disorder ☐
CONpro: Contact Number ☐
CMAPpro: Contact Map ☐
3Dpro: Tertiary Structure ☐
SOLpro: Solubility upon Overexpression ☐

ACCpro20: Solvent Accessibility (20 Class) ☑
SSpro8: Secondary Structure (8 Class) ☐
DOMpro: Domains ☐
DIpro: Disulfide Bonds ☐
SVMcon: New SVM Contact Map ☐
COBEpro: Continuous B-cell Epitopes ☐
ANTIGENpro: Protein Antigenicity ☐
VIRALpro: Capsid & Tail Proteins ☐

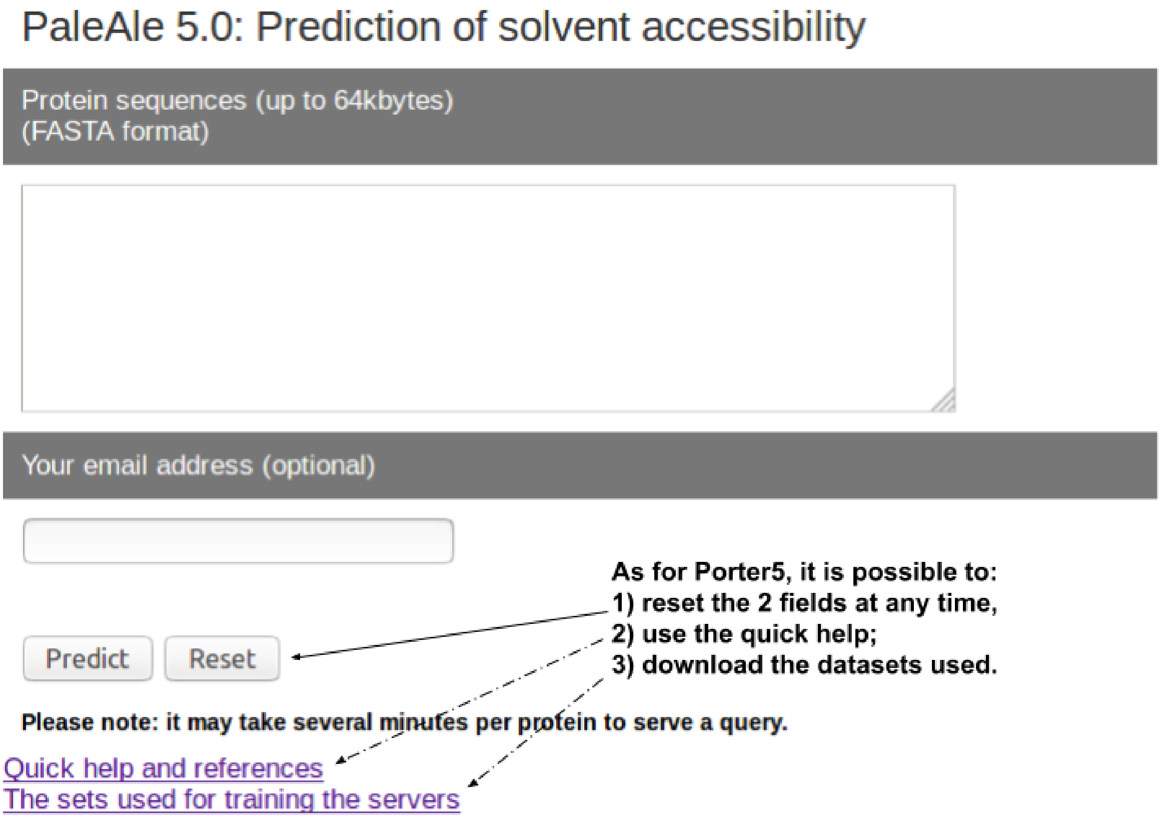
Figure 7: A view of Scratch Protein Predictor where both ACCpro predictors have been selected.

The standalone of ACCpro has been updated in 2015 and is available at http://download.igb.uci.edu/ as part of SCRATCH-1D v1.1. As described above (in Secondary Structure, SSpro) all the requirements are delivered together with the bundled predictors – i.e., ACCpro, ACCpro20, SSpro and SSpro8.

PaleAle

PaleAle is a historical SA predictor developed in parallel with Porter (see Secondary Structure, Porter) since 2007[7], [40], and is also based on ensembles of cascaded Bidirectional Recurrent Neural Networks[20]. PaleAle has been the first template-based SA predictor[7] while PaleAle (v5) is now able to predict four-states ASA – i.e., exposed at 0-4%, 4-25%, 25-50% or 50+%. Like Porter5 and Porter+5 (see Torsion Angles), PaleAle5 relies on both HHblits and PSI-BLAST to gather evolutionary information and, thus, improve its predictions.

The web server of PaleAle is available at http://distilldeep.ucd.ie/paleale/. As for Porter and Porter+ (see respective sections) the protein sequence is the only requirement while an email address is optional. More information about these servers is available in the Secondary Structure, Porter subsection.



PaleAle 5.0: Prediction of solvent accessibility

Protein sequences (up to 64kbytes)
(FASTA format)

Your email address (optional)

As for Porter5, it is possible to:
1) reset the 2 fields at any time,
2) use the quick help;
3) download the datasets used.

Please note: it may take several minutes per protein to serve a query.

Quick help and references
The sets used for training the servers

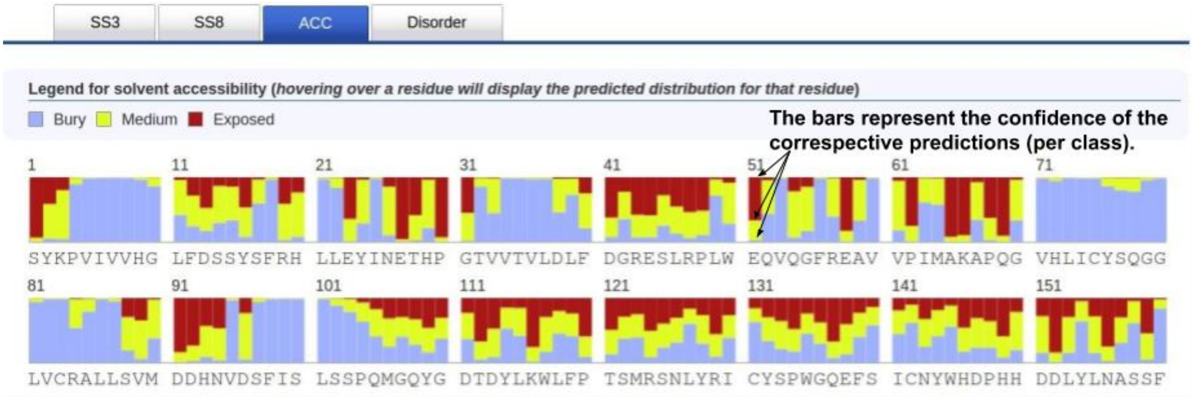
Figure 8: A view of PaleAle5 where the reset button and the links are highlighted.

The light standalone of PaleAle is available at the same address and requires only python3 and HHblits to perform SA predictions. As in Porter, PSI-BLAST can be optionally employed to gather further evolutionary information. The output file presents the confidence per each of the four-states predicted. The datasets are released at the same address.

RaptorX-Property

RaptorX-Property, described in section Secondary Structure, is 2016 suite of predictors able to predict SA, SS and disorder regions[33]. RaptorX-Property predicts SA in three-states with thresholds at 10% and 40%, respectively. As for SS predictions, RaptorX-Property can avoid to look for evolutionary information to speed up predictions at the cost of lower accuracy. It relies on HHblits[11] to gather evolutionary information.

The web server of RaptorX-Property is available at http://raptorx.uchicago.edu/StructurePropertyPred/predict/. The result page of RaptorX-Property provides the predicted 1D-annotations in different tabs (Figure 9 shows the three-states SA). The web server and the released standalone are described in section Secondary Structure, RaptorX-Property.


SS3
SS8
ACC
Disorder
1
11
21
31
41
51
61
71
SYKPVIVVHG LFDSSYSFRH LLEYINETHP GTVVTVLDLF DGRESLRPLW EQVQGFREAV VPIMAKAPQG VHLICYSQGG
81
91
101
111
121
131
141
151
LVCRALLSVM DDHNVDSFIS LSSPQMGQYG DTDYLKWLFP TSMRSNLYRI CYSPWGQEFS ICNYWHDPHH DDLYLNASSF

Legend for solvent accessibility (*hovering over a residue will display the predicted distribution for that residue*)

🟦 Bury 🟨 Medium 🟥 Exposed

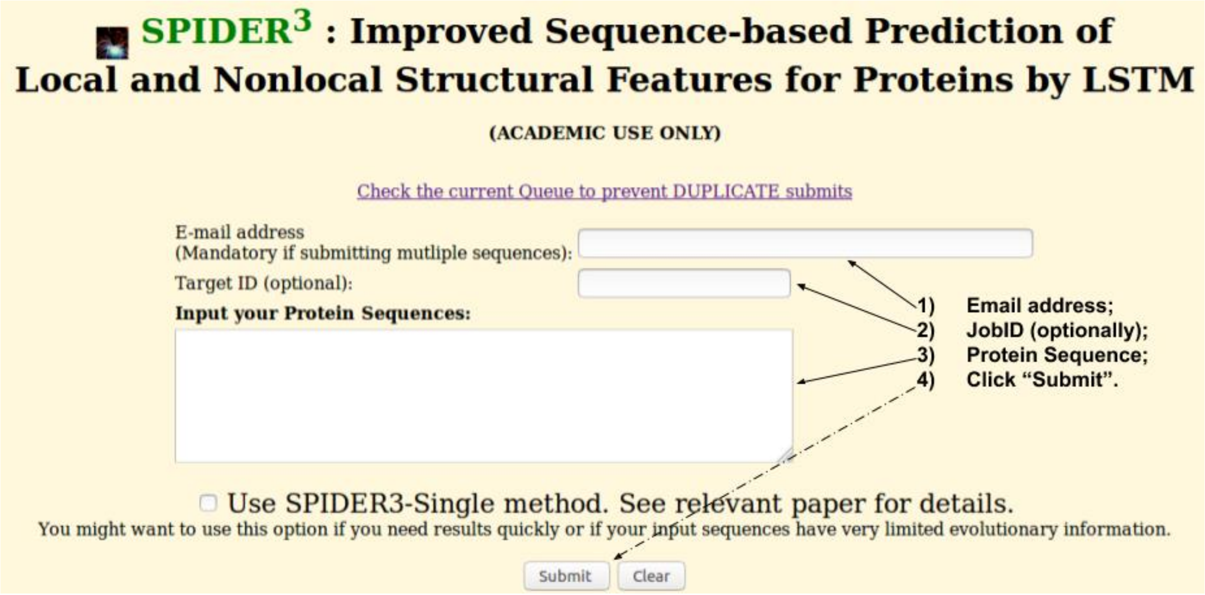
The bars represent the confidence of the correspective predictions (per class).

Figure 9: The view on the predicted three-states SA performed by RaptorX-Property.

SPIDER3

SPIDER has been able to predict SA, SS and TA since 2015[42] and was updated in 2017[34]. SPIDER3, described also in sections Secondary Structure and Torsion Angles, predicts the ASA using real numbers rather than classes, differently from the other predictors here presented[42]. SPIDER2 has been the first HSE predictor[54] while SPIDER3 predicts HSEα-up and HSEα-down using real numbers, although Heffernan et al. reports results also on HSEβ-up and HSEβ-down[34].

The web server and the standalone of SPIDER3 are described in Secondary Structure, SPIDER3. As a side note, the result page and the confirmation email of the web server show the predicted SA only as ASA in ten-classes – i.e., [0-9] – while the predicted ASA, HSEβ-up and HSEβ-down in real numbers are listed in the output file (“*.spd33”) in the temporary directory, along with PSSM/HMM files (see Figure 5).



SPIDER^3 : Improved Sequence-based Prediction of Local and Nonlocal Structural Features for Proteins by LSTM

(ACADEMIC USE ONLY)

Check the current Queue to prevent DUPLICATE submits

E-mail address
(Mandatory if submitting mutliple sequences):

Target ID (optional):

Input your Protein Sequences:

□ Use SPIDER3-Single method. See relevant paper for details.

You might want to use this option if you need results quickly or if your input sequences have very limited evolutionary information.

1) Email address;
2) JobID (optionally);
3) Protein Sequence;
4) Click “Submit”.

Figure 10: A view of the input window of SPIDER3. The steps to follow to start a prediction are highlighted.

Torsional Angles

Protein torsion (or dihedral or rotational) angles can accurately describe the local conformation of protein backbones. The main protein backbone dihedral angles are: phi (φ), psi (ψ) and omega (ω). The planarity of protein bonds restricts ω to be either 180° (typical case) or 0° (rarely). Therefore, it is generally sufficient to use φ and ψ to accurately describe the local shape of a protein.

TA are highly correlated to protein SS and particularly informative in highly variable loop regions. In fact, while TA of α-helices and β-sheets are mostly clustered and regularly distributed[58], φ and ψ can be more effective in describing the local conformation of residues when they are classified as coils (i.e., neither of the other SS classes). When four consecutive residues are considered, a different couple of angles can be observed: theta (θ) and tau (τ)[59]. Thus, different annotations (i.e., SS, φ/ψ and θ/τ) can be adopted to describe the backbone of a protein.

TA are essentially an alternative representation of local structure with respect to SS. Both TA and SS have been successfully used as restraints toward sequence alignment[60], protein folding[25] and tertiary structure prediction[61]. HMM[62], support vector machines (SVM)[58] and several architectures of NN (e.g., iterative[34], [42] and cascade-correlation[63]) have been analysed to predict TA since 2000. NN are currently the main tool to predict TA, in parallel with protein SS[34] or sequentially after it[63], [64].

φ and ψ can be predicted as real numbers or letters(/clusters). In fact, φ and ψ can range from 0° to 360° but are typically observed in certain ranges, given from chemical and physical characteristics of proteins. Bayesian probabilistic[65], [66], multidimensional scaling (MDS)[67] and density plot[58] approaches have been exploited to define different alphabets of various sizes.

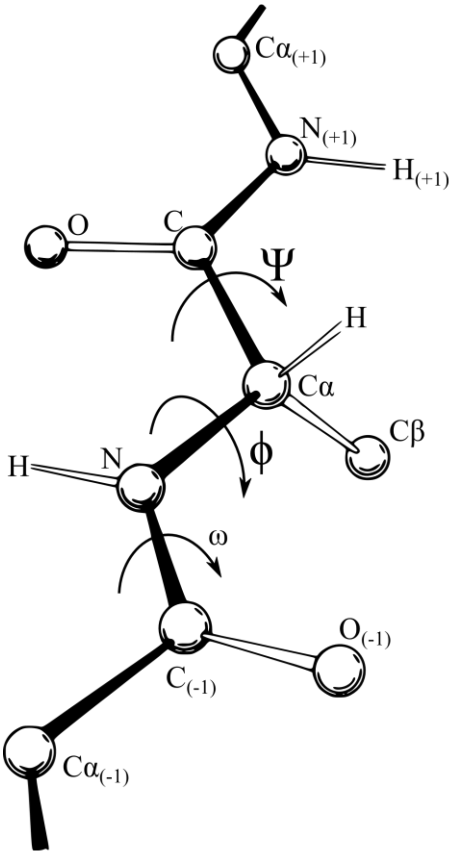


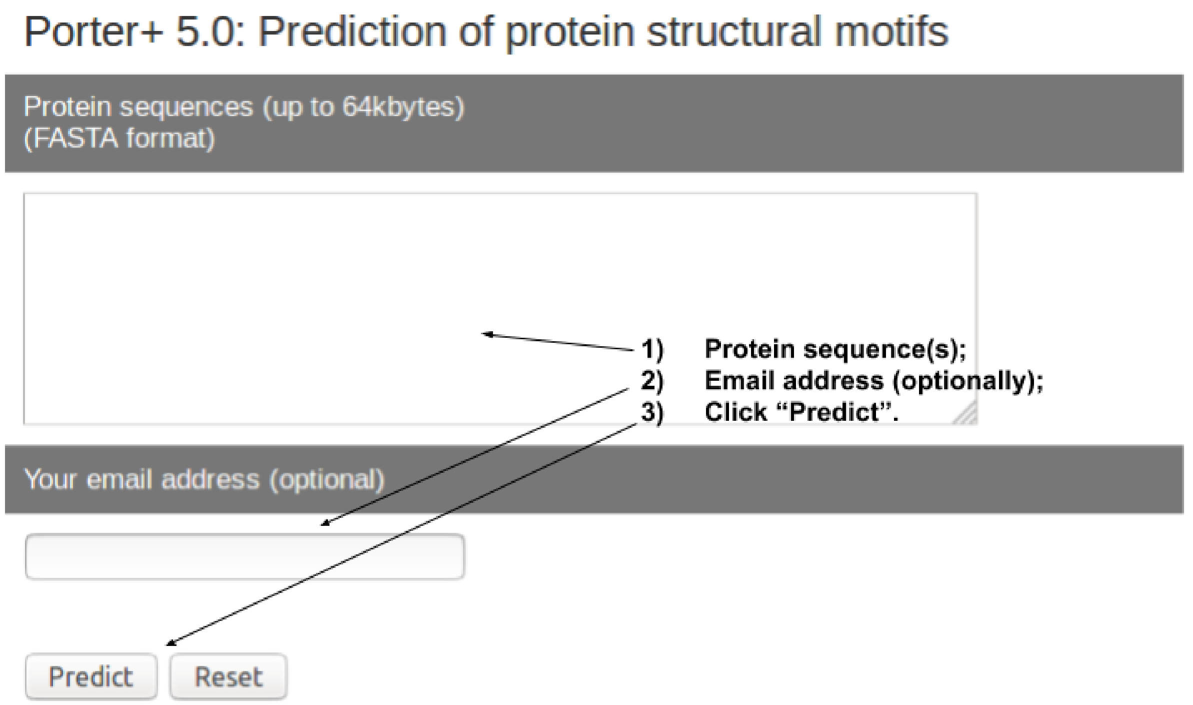
Figure 11: Protein backbone dihedral angles phi, psi and omega.

Name	Web Server	Notes
Porter+[64]	http://distilldeep.ucd.ie/porter+	φ/ψ in 16 letters
SPIDER3[34]	http://sparks-lab.org/server/SPIDER3/	φ/ψ and θ/τ , Numpy or Tensorflow

Porter+

Porter+ is a TA predictor able to classify the φ and ψ angles of a given protein. It was initially developed in 2006 as intermediate step to improve Porter (a SS predictor described in section Secondary Structure) [64]. Porter+ adopts an alphabet of 16 letters devised by Sims et al. using MDS on tetra-peptides (4 contiguous residues)[67]. Porter+, similarly to Porter and PaleAle (see Solvent Accessibility, PaleAle), implements BLAST+ to gather evolutionary information and improve the final prediction. As Porter and PaleAle, the most recent version of Porter+ (v5) adopts also HHblits to greatly improve its accuracy.

The web server of Porter+ is available at http://distilldeep.ucd.ie/porter+. The protein sequence is required, while an email address is optional. It will be then sufficient to confirm (clicking "Predict") to view a confirmation page with the overview of the job. Once ready, the prediction will be received by email. It will resemble the format adopted for Porter, see in section Secondary Structure. Porter+ can be executed in parallel with Porter or PaleAle, or several more protein predictors, at http://distillf.ucd.ie/distill/ to predict SS, SA, or other protein features, respectively.



Porter+ 5.0: Prediction of protein structural motifs

Protein sequences (up to 64kbytes)
(FASTA format)

1) Protein sequence(s);
2) Email address (optionally);
3) Click "Predict".

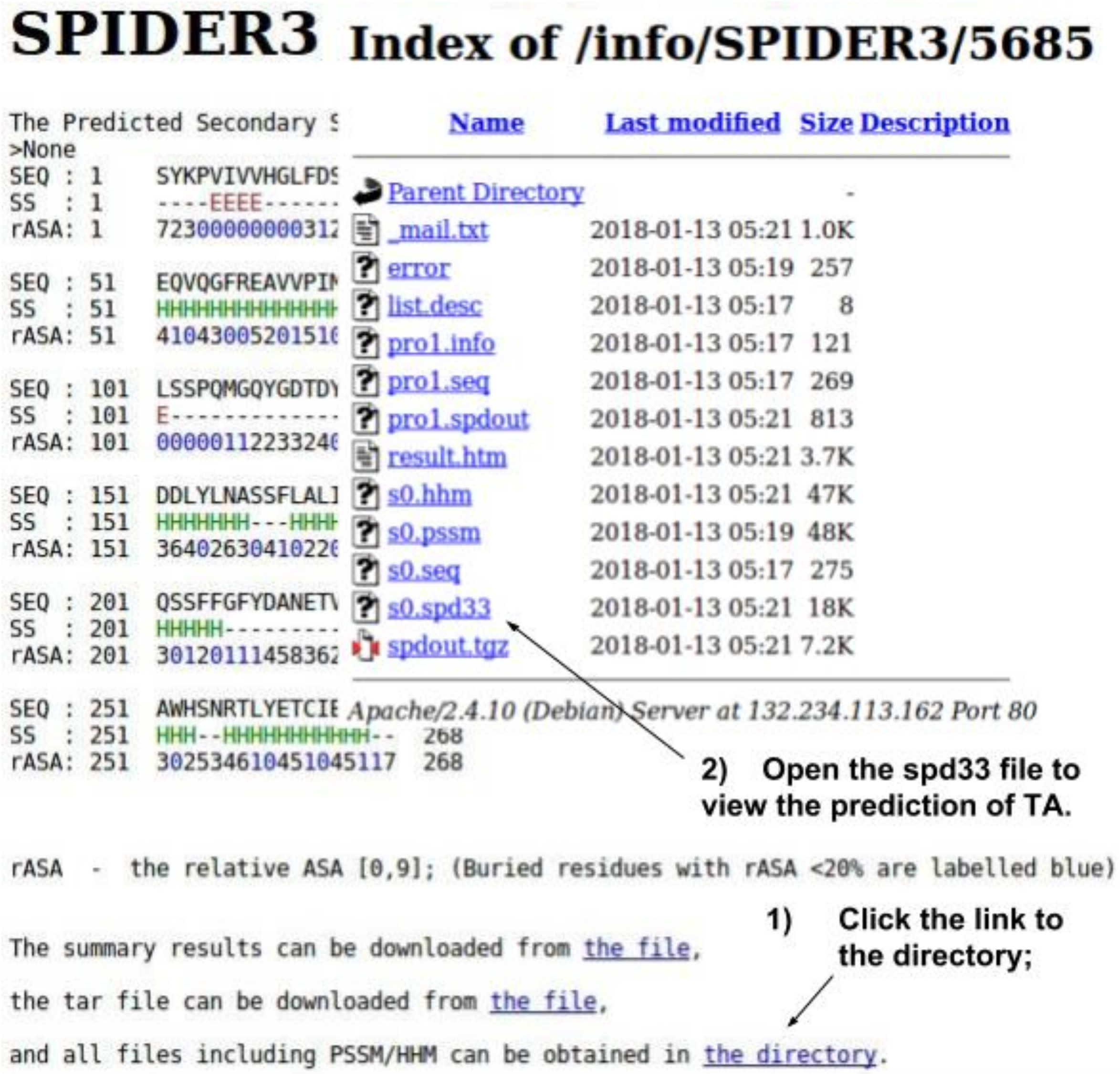
Your email address (optional)

Figure 12: A view of Porter+5 where the steps to start a prediction are highlighted.

The light standalone of Porter+ is available at http://distilldeep.ucd.ie/porter+ and closely resembles the one described in section Secondary Structure, Porter. The output of Porter+ overviews the confidence for all 14 classes predicted. The datasets adopted for training and testing purposes are also released.

SPIDER3

SPIDER3, also in section Secondary Structure and Solvent Accessibility, predicts TA using real numbers (ℝ). SPIDER was initially released in 2014 to predict only θ/τ [59]. It has been further developed to also predict φ/ψ, in parallel with SS, SA and contact numbers (see the respective sections) [34], [42]. More details, regarding the pipeline implemented, the web server offered and the standalone available, are outlined in section Secondary Structure.



SPIDER3 Index of /info/SPIDER3/5685

```
The Predicted Secondary S
>None
SEQ : 1    SYKPVIVVHGLFDS
SS  : 1    ----EEEE------
rASA: 1    72300000000312

SEQ : 51   EQVQGFREAVVPIN
SS  : 51   HHHHHHHHHHHHHH
rASA: 51   41043005201516

SEQ : 101  LSSPQMGQYGDTDY
SS  : 101  E-------------
rASA: 101  00000112233246

SEQ : 151  DDLYLNASSFLALI
SS  : 151  HHHHHHH---HHHH
rASA: 151  36402630410226

SEQ : 201  QSSFFGFYDANETV
SS  : 201  HHHHH---------
rASA: 201  30120111458362

SEQ : 251  AWHSNRTLYETCIE
SS  : 251  HHH--HHHHHHHHHH--   268
rASA: 251  302534610451045117  268
```

```
Name              Last modified     Size Description

Parent Directory                       -
_mail.txt         2018-01-13 05:21 1.0K
error             2018-01-13 05:19  257
list.desc         2018-01-13 05:17    8
pro1.info         2018-01-13 05:17  121
pro1.seq          2018-01-13 05:17  269
pro1.spdout       2018-01-13 05:21  813
result.htm        2018-01-13 05:21 3.7K
s0.hhm            2018-01-13 05:21  47K
s0.pssm           2018-01-13 05:19  48K
s0.seq            2018-01-13 05:17  275
s0.spd33          2018-01-13 05:21  18K
spdout.tgz        2018-01-13 05:21 7.2K

Apache/2.4.10 (Debian) Server at 132.234.113.162 Port 80
```

2) Open the spd33 file to view the prediction of TA.

```
rASA  -  the relative ASA [0,9]; (Buried residues with rASA <20% are labelled blue)

The summary results can be downloaded from the file,

the tar file can be downloaded from the file,

and all files including PSSM/HHM can be obtained in the directory.
```

1) Click the link to the directory;

Figure 13: A view of the results page of SPIDER3 where the steps to view the predicted TA are highlighted.

Contact Maps

Contact Maps (CM) are the main two-dimensional protein structure annotation. A plain 2D representation of protein tertiary structure would describe the distance between all possible pairs of AA using a matrix containing real values. Such dense representation, referred as distance map, is reduced to a more compact abstraction – i.e., CM – by quantising a distance map through a fixed threshold, i.e. describing distances not as real numbers but as contacts (distance smaller than the threshold) or no. This latter abstraction is routinely exploited to reconstruct protein tertiary structures implementing heuristic methods[68], [69]. Thus, 3D structure prediction being a computationally expensive problem motivates the development of the aforementioned heuristic methods that aim to be both robust against noise in the CM – i.e., to ideally fix CM prediction errors – and computationally applicable on a large scale[70], [71]. Following closely the development of the third generation of SS predictors, motivated by the same abundance of available data and computational resources, MSA have been thoroughly tested and successfully exploited to extract promising features for CM prediction – e.g., correlated mutations, sequence conservation, alignment stability and family size[72]–[74]. These initial advancements led to the first generation of ML methods able to predict CM[24], [75], [76]. Though, given that MSA are replete with useful but noisy information, statistical insights have been necessary to further exploit the growing amount of evolutionary information – e.g., distinguishing between indirect- and direct-coupling[77], [78]. The most recent CM predictors gather recent intuitions in both statistics and advanced ML, aiming to collect, clean and employ as much useful data as possible[22], [33], [50]. Differently from the other protein annotations in this chapter, CM is currently assessed at CASP[79] and CAMEO[6].

The intrinsic properties of CM – namely, being compact and discrete two-state annotations, invariant to rotations and translations – makes them a more appropriate target for ML techniques than protein tertiary structures or distance maps although still highly informative about the protein 3D structures[80]. CM prediction is a typical intermediate step in many pipelines to predict protein tertiary structure[52], [81], [82]. For example, it is a key component for contact-assisted structure prediction[83], contact assisted protein folding[23], free and template-based modelling[81]. CM have also been used to predict protein disorder[84], protein function[72] and to detect challenging templates[52]. In fact, even partial CM can greatly support robust and accurate protein structure modelling[85].

Being a 2D annotation, CM are typically gradually predicted starting from simpler but less informative 1D annotations – e.g., SA, SS and TA[75], [76], [86]. The advantages of this incremental approach lie in the intrinsic nature of protein abstractions – i.e., 1D annotations are easier to predict while providing useful insights. For example, Figure 14 highlights the strong relations between SS conformations and CM. The contact occupancy – i.e., contact number, or number of contacts per AA – is another 1D protein annotation which has been successfully predicted [34], [49], [87] to adjust and improve CM prediction[73], [75], [86]. Eigenvector decomposition has been used as a means for template-search[88] and principal eigenvector (PE) prediction as an intermediate step towards CM prediction[24]. Finally, correlated mutations appear to be the most informative protein feature for CM prediction – i.e., residues in contact tend to coevolve to maintain the physiochemical equilibrium[72]–[74]. Thus, statistical methods have been extensively assessed to look for coevolving residues, gathering mutual information from MSA while aiming to discriminate direct- from indirect-coupling mutations – e.g., implementing sparse inverse covariance estimation to remove indirect-coupling[77], [89], [90].

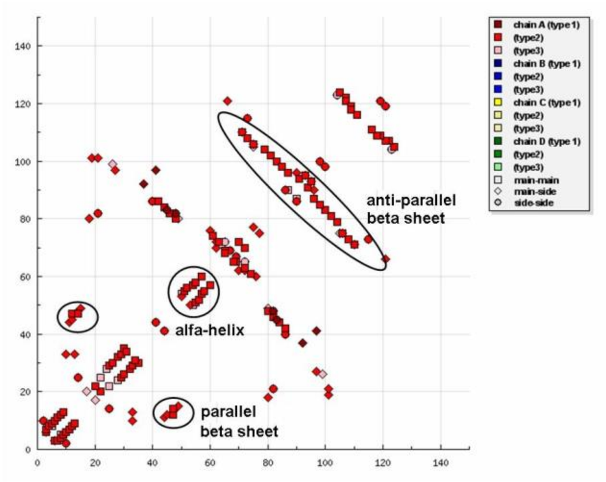


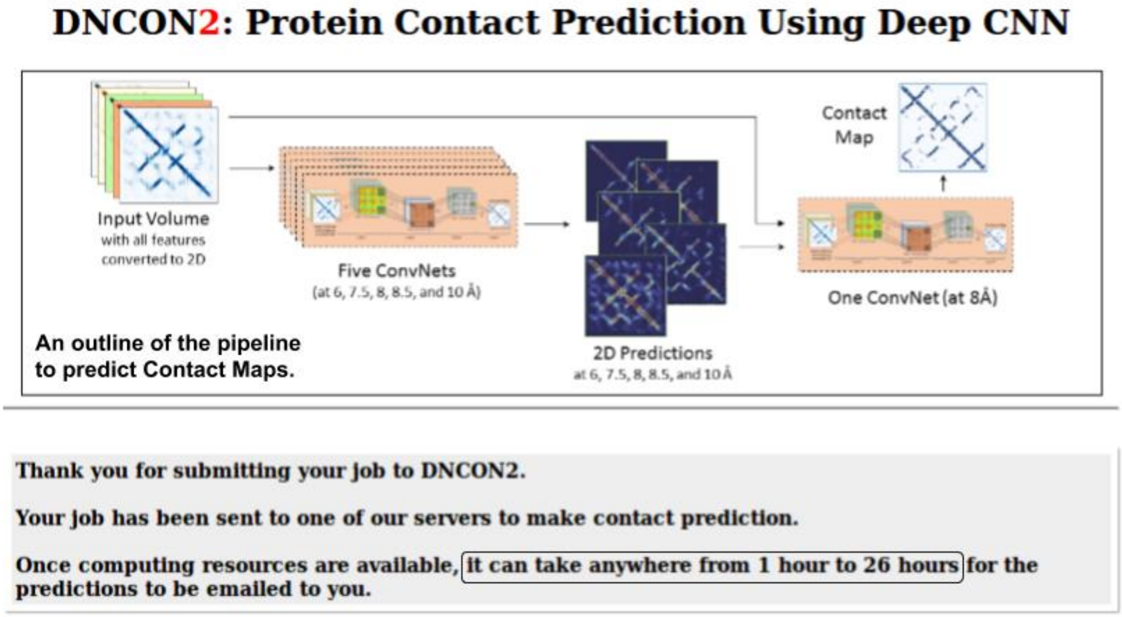
Figure 14: CM with highlighted SS conformations.

As in Figure 14, CM are represented as (symmetric) matrices or graphs – rather than vectors – where around 2-5% of all possible pairs of AA are “in contact” – i.e., an unbalanced problem in ML[80]. Notably, the number of AA in contact increases almost linearly with the protein length – i.e., shorter proteins are denser than longer ones[80]. A pair of AA is in contact when the Euclidian distance between their C_β (or Cα, for glycine) atoms is closer than a given threshold. This threshold is usually set between 6 and 12 Å (8Å at CASP[79]), although values in the range of 10-18 Å may lead to better reconstructions[68]. In fact, it is arguable whether all predicted "contacts" should be taken in consideration or certain criteria should be applied, such as focusing on those predicted with the highest confidence – i.e., the top 10, L/5, L/2 or L contacts, with L = protein length – or with a minimum probability threshold[79]. For example, tertiary structure modelling benefits more from well distributed contacts, thus the entropy score is one of the measures of interest to evaluate CM predictors[79]. Precision – i.e., the ratio between true contact and (true contact + wrong contact) – is usually adopted to assess local (short range) contacts – i.e., involving AA within 10 positions apart – and non-local (long range) contact, separately. Typically, CM predictors are evaluated at CASP through more complex measures[79], [83], [91], such as z-scores – i.e., weighted sum of energy separation with the true structure for each domain –, GDT_TS – i.e., score of optimal superposition between the predicted and the true structure –, root mean squared deviation (RMSD) or TM-score – i.e., a measure more sensitive at the global (rather than local) structure than RMSD [92]. Classic statistical and ML measures, such as the aforementioned precision, recall, F1 score, Matthews Correlation Coefficient (MCC) are also adopted in parallel with more unusual ones, such as alignment depth or entropy score[79]. The average precision of the top predictors at CASP12 was 47% on L/5 long range contacts for the difficult category, while the highest GDT_TS for each of the 14 domains assessed went from 12 to 70[79].

Though correlated mutations and NN have been identified as promising instruments to also predict CM [75], pairwise contact potential[84], self-organising maps[93] and SVM[76] have been used in the past. While 2D-BRNN[86], [94], multi-stage[24], [95] and template-based[51] NN approaches have initially characterized the field[96], the most recent CM predictors rely on multiple 1D protein annotation predictors – e.g., predicting SA and SS along with other protein features –, two-stage approaches and coevolution information[50], [97] or multi-class maps[71], [96]. The standard output format of any CM predictor is a text file organised in 5 columns as follow: the positions of the two AA in contact, a blank column, the set threshold (8Å) and the confidence of each predicted contact.

Name	**Web Server**	**Notes**
DNCON[50]	http://sysbio.rnet.missouri.edu/dncon2/	Three coevolution algorithms, Computer Vision inspired
MetaPSICOV[22]	http://bioinf.cs.ucl.ac.uk/MetaPSICOV/	CCMpred, FreeContact and PSICOV, hydrogen bonds
RaptorX-Contact[23]	http://raptorx.uchicago.edu/ContactMap/	Inspired from Computer Vision, CCMpred only
XX-Stout[51]	http://distilldeep.ucd.ie/xxstout/	Contact Density, template-based, multi-class CM

DNCON2


DNCON2: Protein Contact Prediction Using Deep CNN
Input Volume
with all features converted to 2D
Five ConvNets
(at 6, 7.5, 8, 8.5, and 10 Å)
2D Predictions
at 6, 7.5, 8, 8.5, and 10 Å
Contact Map
One ConvNet (at 8Å)
An outline of the pipeline to predict Contact Maps.

Thank you for submitting your job to DNCON2.

Your job has been sent to one of our servers to make contact prediction.

Once computing resources are available, it can take anywhere from 1 hour to 26 hours for the predictions to be emailed to you.

Figure 15: The pipeline of DNCON2 is summarised in the confirmation page.

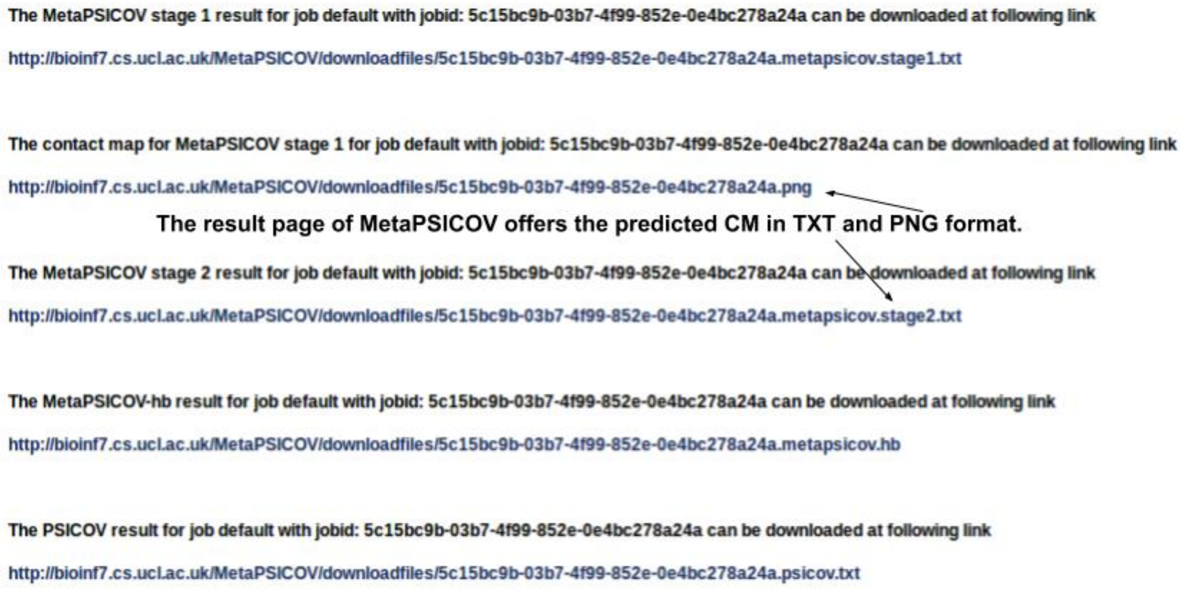
DNCON has been initially released in 2012[98], assessed at CASP10[91] and updated in 2017[50]. DNCON2 gathers coevolution signal along with 1D protein features – e.g., PSIPRED and SSpro (see section Secondary Structure) – with a similar approach to MetaPSICOV2 (see below). It then predicts CM with different thresholds – namely, 6, 7.5, 8, 8.5 and 10 Å – resembling the multi-class maps of XX-Stout (see below) and finally refines them generating only one CM at 8Å. In the described two-stage approach, DNCON2 implements a total of 6 NN like RaptorX-Contact (Figure 15). Thus, DNCON2 further exploits the most recent intuitions in CM prediction, including recent ML algorithms.

The web server and dataset of DNCON2 are available at http://sysbio.rnet.missouri.edu/dncon2/. JobID and email are required, along with the sequence to predict (up to two sequences at time). Once the prediction is ready, typically in less than 24h, the predicted CM is sent by email in both text and image format as email content and attachment, respectively. The email content specifies the number of alignments found and the predicted CM (in the standard 5 columns text format).

The standalone of DNCON2 is available at https://github.com/multicom-toolbox/DNCON2/. The same page lists all the instructions to install every requirement – i.e., CCMpred[90], FreeContact[89], HHblits[11], JackHMMER[99] and PSICOV[77] for coevolution information, python libraries (such as Tensorflow), MetaPSICOV and PSIPRED (see Secondary Structure, PSIPRED) for SS and SA prediction. Once all the requirements are met, it is possible to verify whether DNCON2 is fully running dealing with the predictions of 3 proposed sequences. The results of each predictor and package involved is organised in directories.

MetaPSICOV

MetaPSICOV is a CM predictor which has been initially released in 2014 for CASP11[100] and updated in 2016 for CASP12[97]. It is recognised as the first CM predictor successfully able to exploit the recent advancements in co-evolutionary information extraction[101]. In particular, MetaPSICOV achieved this result implementing three different algorithms to extract coevolution signal from MSA generated with HHblits[11] and HMMER[37] – i.e., CCMpred[90], FreeContact[89] and PSICOV[77] – along with other local and global features used for SVMcon[76]. It relies on PSIPRED (see Secondary Structure, PSIPRED) to predict SS and a similar ML method to predict SA. As a final step, MetaPSICOV adopts a two-stage NN to infer CM from the features described[22]. The web server and standalone of MetaPSICOV can be used to predict hydrogen bonding patterns[22].



The MetaPSICOV stage 1 result for job default with jobid: 5c15bc9b-03b7-4f99-852e-0e4bc278a24a can be downloaded at following link

http://bioinf7.cs.ucl.ac.uk/MetaPSICOV/downloadfiles/5c15bc9b-03b7-4f99-852e-0e4bc278a24a.metapsicov.stage1.txt

The contact map for MetaPSICOV stage 1 for job default with jobid: 5c15bc9b-03b7-4f99-852e-0e4bc278a24a can be downloaded at following link

http://bioinf7.cs.ucl.ac.uk/MetaPSICOV/downloadfiles/5c15bc9b-03b7-4f99-852e-0e4bc278a24a.png

The result page of MetaPSICOV offers the predicted CM in TXT and PNG format.

The MetaPSICOV stage 2 result for job default with jobid: 5c15bc9b-03b7-4f99-852e-0e4bc278a24a can be downloaded at following link

http://bioinf7.cs.ucl.ac.uk/MetaPSICOV/downloadfiles/5c15bc9b-03b7-4f99-852e-0e4bc278a24a.metapsicov.stage2.txt

The MetaPSICOV-hb result for job default with jobid: 5c15bc9b-03b7-4f99-852e-0e4bc278a24a can be downloaded at following link

http://bioinf7.cs.ucl.ac.uk/MetaPSICOV/downloadfiles/5c15bc9b-03b7-4f99-852e-0e4bc278a24a.metapsicov.hb

The PSICOV result for job default with jobid: 5c15bc9b-03b7-4f99-852e-0e4bc278a24a can be downloaded at following link

http://bioinf7.cs.ucl.ac.uk/MetaPSICOV/downloadfiles/5c15bc9b-03b7-4f99-852e-0e4bc278a24a.psicov.txt

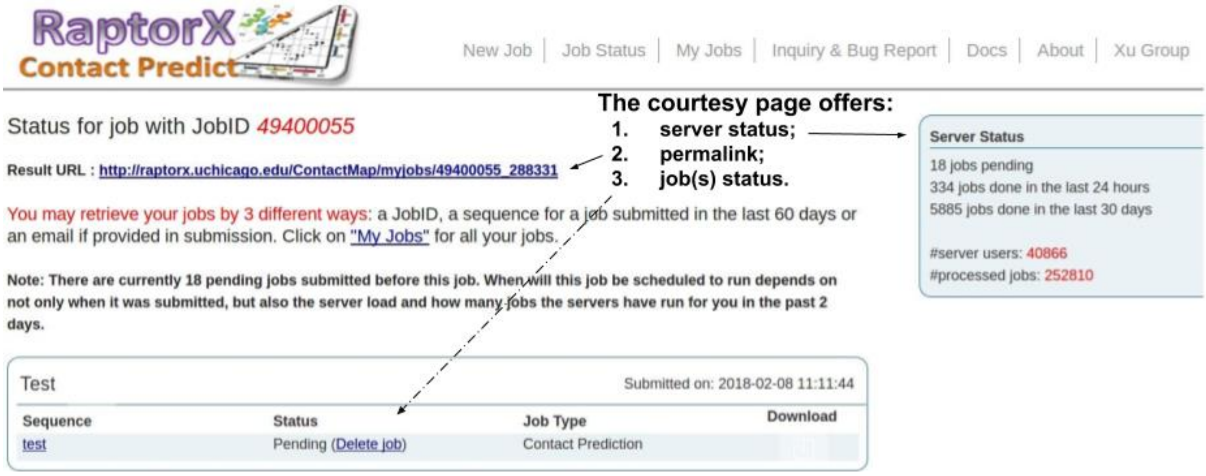
Figure 16: A typical result page of MetaPSICOV. All the files, except the png, follow PSICOV's format.

The web server of the 2014 version of MetaPSICOV is available at http://bioinf.cs.ucl.ac.uk/MetaPSICOV. A simple interface, which resembles the web server of PSIPRED (see Secondary Structure, PSIPRED), asks for a single sequence in FASTA format and a short identifier. A confirmation page is automatically shown when the job is completed. If an email address is inserted, an email containing only the permalink to the result page will be sent. As in Figure 16, the result page contains links to the output of MetaPSICOV stage 1 (also as image), of stage 2, of MetaPSICOV-hb (hydrogen bonds) and of PSICOV. A typical CM takes between 20 minutes and 6 hours to be predicted.

The very last version of MetaPSICOV is usually available as standalone at http://bioinfadmin.cs.ucl.ac.uk/downloads/MetaPSICOV/. To run MetaPSICOV2, it is required to install (legacy) BLAST, PSIPRED, PSICOV, FreeContact, CCMpred, HHblits and HMMER, separately. Once the required packages are installed, it is sufficient to follow the README to complete the setup and run MetaPSICOV2. Each run of MetaPSICOV2 will generate the needed features – i.e., the output of the required packages, such as PSIPRED and PSICOV – along with the predicted CM (in standard text format).

RaptorX-Contact

RaptorX-Contact is a 2016 CM predictor which performed well at the last CASP12[23], [79], [102]. RaptorX-Contact aimed to exploit both Computer Vision[3] and coevolution intuitions to further improve CM prediction. It employs RaptorX-Property[33] (see Secondary Structure and Solvent Accessibility) to predict SS and SA, CCMpred[90] to look for co-evolutionary information and in-house algorithms for mutual information and pairwise potential extraction. RaptorX-Contact was trained using MSA generated with PSI-BLAST[9] while it uses HHblits[11] at prediction time. Thus, the web server and standalone depends on HHblits only.



The courtesy page offers:

1. server status;
2. permalink;
3. job(s) status.

Status for job with JobID 49400055

Result URL : http://raptorx.uchicago.edu/ContactMap/myjobs/49400055_288331

You may retrieve your jobs by 3 different ways: a JobID, a sequence for a job submitted in the last 60 days or an email if provided in submission. Click on "My Jobs" for all your jobs.

Note: There are currently 18 pending jobs submitted before this job. When will this job be scheduled to run depends on not only when it was submitted, but also the server load and how many jobs the servers have run for you in the past 2 days.

Server Status

18 jobs pending
334 jobs done in the last 24 hours
5885 jobs done in the last 30 days

#server users: 40866
#processed jobs: 252810

Test — Submitted on: 2018-02-08 11:11:44

Sequence	Status	Job Type	Download
test	Pending (Delete job)	Contact Prediction	

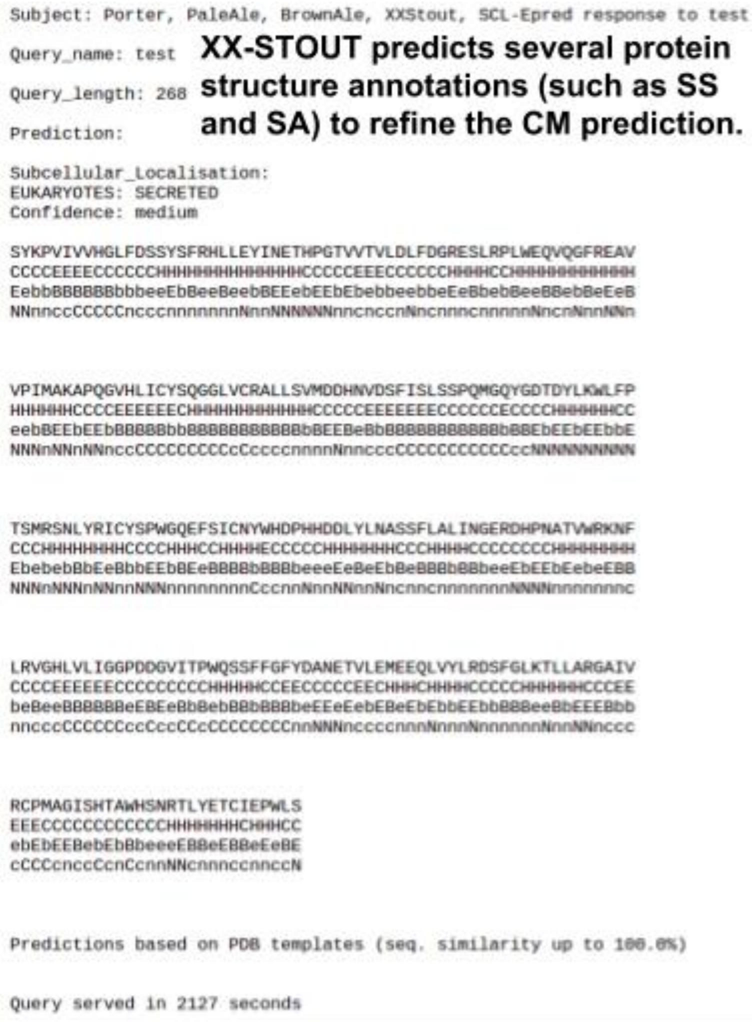
Figure 17: The confirmation page of RaptorX-Contact tells the pending jobs ahead and the result URL.

The web server of RaptorX-Contact is available at http://raptorx.uchicago.edu/ContactMap/. Once a protein sequence (in FASTA format) has been inserted, it is possible to submit it and a result URL will be provided (Figure 17). A JobID is recommended to distinguish among past submissions in the "My Jobs" page, while an email address can be specified to receive the outcome of RaptorX-Contact by email – i.e., the result URL and, as attachments, the predicted CM in text and image format. The tertiary structure is also predicted by default but it is possible to uncheck the respective box to speed up the CM prediction. Up to 50 protein primary structures can be submitted at the same time through the input form or uploaded from one's computer. Optionally, a MSA (of up to 20,000 sequences) can be sent instead of a protein sequence. The result URL links to an interactive page where it is possible to navigate the predicted CM besides downloading it in text or image format. The MSA generated (in A2M format), the CCMpred[90] output and the 3D models (if requested) are also made available. Finally, it is also possible to query the web server from command-line (using curl) as explained at http://raptorx.uchicago.edu/ContactMap/documentation/.

XX-STOUT

XX-STOUT is a CM predictor initially released in 2006[24] and further improved to be template-based[51] and multi-class in 2009[96]. XX-STOUT employs the predictions by BrownAle, PaleAle and Porter (see Secondary Structure and Solvent Accessibility) – i.e., contact density, SS and SA predictions, respectively – to generate multi-class CM – i.e., CM with four-states annotations. When either PSI-BLAST[9] or the in-house fold recognition software finds homology information, further inputs are provided to XX-STOUT to perform template-base predictions – i.e., greatly improve the prediction quality exploiting proteins in the PDB[1], [52].

The web server of XX-STOUT is available at http://distilldeep.ucd.ie/xxstout/. An email address and the plain protein sequence are required to start the prediction, a JobID is optional. The confirmation page summarises the information provided and the predictors which are going to be used – i.e., the aforementioned 1D predictors and SCL-Epred, a predictor of subcellular localization[103]. The predicted CM (threshold 8Å), the prediction per residue of SS, SA and contact density, and the predicted protein's location are sent by email. The same email describes the confidence of SCL-Epred's prediction and whether the whole prediction has been based on PDB templates and, if found, of which similarity with the query sequence. The standalone of XX-STOUT and required 1D predictors is available on request.



```
Subject: Porter, PaleAle, BrownAle, XXStout, SCL-Epred response to test

Query_name: test
Query_length: 268
Prediction:

Subcellular_Localisation:
EUKARYOTES: SECRETED
Confidence: medium

SYKPVIVVHGLFDSSYSFRHLLEYINETHPGTVVTVLDLFDGRESLRPLWEQVQGFREAV
CCCCEEEECCCCCCHHHHHHHHHHHHHHCCCCCEEECCCCCCHHHHCCHHHHHHHHHHHH
EebbBBBBBBbbbeeEbBeeBeebBEEebEEbEbebbeebbeEeBbebBeeBBebBeEeB
NNnnccCCCCCncccnnnnnnnnNnnNNNNNNnncnccnNncnnncnnnnnNncnNnnNNn


VPIMAKAPQGVHLICYSQGGLVCRALLSVMDDHNVDSFISLSSPQMGQYGDTDYLKWLFP
HHHHHHCCCCEEEEEEECHHHHHHHHHHHHHCCCCCEEEEEEEECCCCCCECCCCHHHHHHCC
eebBEEbEEbBBBBBbbBBBBBBBBBBBbBEEBeBbBBBBBBBBBBBbBBEbEEbEEbbE
NNNnNNnNNnccCCCCCCCCCcCccccnnnnNnncccCCCCCCCCCCCccNNNNNNNNN


TSMRSNLYRICYSPWGQEFSICNYWHDPHHDDLYLNASSFLALINGERDHPNATVWRKNF
CCCHHHHHHHHCCCCHHHCCHHHHECCCCCHHHHHHHCCCHHHHCCCCCCCCHHHHHHHH
EbebebBbEeBbbEEbBEeBBBBbBBBbeeeEeBeEbBeBBBbBBbeeEbEEbEebeEBB
NNNnNNNnNNnnNNNnnnnnnnnCccnnNnnNNnnNncnncnnnnnnnNNNNnnnnnnnc


LRVGHLVLIGGPDDGVITPWQSSFFGFYDANETVLEMEEQLVYLRDSFGLKTLLARGAIV
CCCCEEEEEEECCCCCCCCHHHHHCCEECCCCCEECHHHCHHHHCCCCCHHHHHHCCCEE
beBeeBBBBBBeEBEeBbBebBBbBBBbeEEeEebEBeEbEbbEEbbBBBeeBbEEEBbb
nncccCCCCCCcccCcccCCcCCCCCCCCnnNNNncccccnnnNnnnNnnnnnnNnnNNnccc


RCPMAGISHTAWHSNRTLYETCIEPWLS
EEECCCCCCCCCCCCHHHHHHHCHHHCC
ebEbEEBebEbBbeeeEBBeEBBeEeBE
cCCCcnccCcnCcnnNNcnnnccnnccN


Predictions based on PDB templates (seq. similarity up to 100.0%)

Query served in 2127 seconds
```

XX-STOUT predicts several protein structure annotations (such as SS and SA) to refine the CM prediction.

Figure 18: XX-STOUT sends the predicted protein structure annotations in the body email except the CM (which is attached).

Conclusions

In this chapter we have discussed the importance of protein structure to understand protein functions and the need for abstractions – i.e., protein structural annotations – to overcome the difficulties of determining such structures *in vitro*. We have then presented an overview of the role Bioinformatics – i.e., *in silico* Biology – has played in advancing such understanding, thanks to one- and two-dimensional abstractions and efficient techniques to predict them that are applicable on a large scale, such as Machine Learning and Deep Learning in particular. The typical pipeline to predict protein structure annotations was also presented, highlighting the key tools adopted and their characteristics.

The chapter then described the main one- and two-dimensional protein structure annotations, from their definition to samples of state-of-the-art methods to predict them. We have given a concise introduction to each protein structure annotation trying to highlight what, why and how is predicted. We also tried to give a sense of how different abstractions are linked to one another and how this is reflected in the systems that predict them.

A considerable part of this chapter is dedicated to presenting, describing and comparing state-of-the-art predictors of protein structure annotations. The methods presented are typically available as both web servers and standalone programs and, thus, can be used for small or large scale experiments and studies. The general aim of this chapter is to introduce and facilitate the adoption of *in silico* methods to study proteins by the broader research community.

References

[1] H. M. Berman *et al.*, "The Protein Data Bank," *Nucleic Acids Res.*, vol. 28, no. 1, pp. 235–242, Jan. 2000.

[2] The UniProt Consortium, "UniProt: the universal protein knowledgebase," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D158–D169, Nov. 2016.

[3] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[4] I. Walsh, G. Pollastri, and S. C. E. Tosatto, "Correct machine learning on protein sequences: a peer-reviewing perspective," *Brief. Bioinform.*, vol. 17, no. 5, pp. 831–840, Sep. 2016.

[5] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, Jan. 2015.

[6] J. Haas *et al.*, "Continuous Automated Model Evaluation (CAMEO) Complementing the Critical Assessment of Structure Prediction in CASP12.," *Proteins Struct. Funct. Bioinforma.*, p. n/a-n/a.

[7] G. Pollastri, A. J. Martin, C. Mooney, and A. Vullo, "Accurate prediction of protein secondary structure and solvent accessibility by consensus combiners of sequence and structure information," *BMC Bioinformatics*, vol. 8, p. 201, Jun. 2007.

[8] S. F. Altschul *et al.*, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.," *Nucleic Acids Res.*, vol. 25, no. 17, pp. 3389–3402, Sep. 1997.

[9] A. A. Schäffer *et al.*, "Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements," *Nucleic Acids Res.*, vol. 29, no. 14, pp. 2994–3005, Jul. 2001.

[10] D. T. Jones and M. B. Swindells, "Getting the most from PSI–BLAST," *Trends Biochem. Sci.*, vol. 27, no. 3, pp. 161–164, Mar. 2002.

[11] M. Remmert, A. Biegert, A. Hauser, and J. Söding, "HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment," *Nat. Methods*, vol. 9, no. 2, pp. 173–175, Feb. 2012.

[12] B. Rost, "Review: Protein Secondary Structure Prediction Continues to Rise," *J. Struct. Biol.*, vol. 134, no. 2, pp. 204–218, May 2001.

[13] Y. Yang *et al.*, "Sixty-five years of the long march in protein secondary structure prediction: the final stretch?," *Brief. Bioinform.*, Dec. 2016.

[14] L. Pauling and R. B. Corey, "Configurations of Polypeptide Chains With Favored Orientations Around Single Bonds," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 37, no. 11, pp. 729–740, Nov. 1951.

[15] J. C. Kendrew *et al.*, "Structure of myoglobin: A three-dimensional Fourier synthesis at 2 A. resolution," *Nature*, vol. 185, no. 4711, pp. 422–427, Feb. 1960.

[16] M. F. Perutz, M. G. Rossmann, A. F. Cullis, H. Muirhead, G. Will, and A. C. North, "Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5-A. resolution, obtained by X-ray analysis," *Nature*, vol. 185, no. 4711, pp. 416–422, Feb. 1960.

[17] P. Y. Chou and G. D. Fasman, "Prediction of protein conformation," *Biochemistry (Mosc.)*, vol. 13, no. 2, pp. 222–245, Jan. 1974.

[18] B. Rost and C. Sander, "Prediction of Protein Secondary Structure at Better than 70% Accuracy," *J. Mol. Biol.*, vol. 232, no. 2, pp. 584–599, Jul. 1993.

[19] D. T. Jones, "Protein secondary structure prediction based on position-specific scoring matrices," *J. Mol. Biol.*, vol. 292, no. 2, pp. 195–202, Sep. 1999.

[20] P. Baldi, S. Brunak, P. Frasconi, G. Soda, and G. Pollastri, "Exploiting the past and the future in protein secondary structure prediction," *Bioinforma. Oxf. Engl.*, vol. 15, no. 11, pp. 937–946, Nov. 1999.

[21] P. Aloy, A. Stark, C. Hadley, and R. B. Russell, "Predictions without templates: New folds, secondary structure, and contacts in CASP5," *Proteins Struct. Funct. Bioinforma.*, vol. 53, no. S6, pp. 436–456, Jan. 2003.

[22] D. T. Jones, T. Singh, T. Kosciolek, and S. Tetchner, “MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins,” *Bioinformatics*, vol. 31, no. 7, pp. 999–1006, Apr. 2015.

[23] S. Wang, S. Sun, Z. Li, R. Zhang, and J. Xu, “Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model,” *PLOS Comput. Biol.*, vol. 13, no. 1, p. e1005324, Jan. 2017.

[24] A. Vullo, I. Walsh, and G. Pollastri, “A two-stage approach for improved prediction of residue contact maps,” *BMC Bioinformatics*, vol. 7, p. 180, Mar. 2006.

[25] Y. Yang, E. Faraggi, H. Zhao, and Y. Zhou, “Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates,” *Bioinformatics*, vol. 27, no. 15, pp. 2076–2082, Aug. 2011.

[26] D. Baú, A. J. Martin, C. Mooney, A. Vullo, I. Walsh, and G. Pollastri, “Distill: a suite of web servers for the prediction of one-, two- and three-dimensional structural features of proteins,” *BMC Bioinformatics*, vol. 7, p. 402, Sep. 2006.

[27] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, “SCOP: A structural classification of proteins database for the investigation of sequences and structures,” *J. Mol. Biol.*, vol. 247, no. 4, pp. 536–540, Apr. 1995.

[28] W. Kabsch and C. Sander, “Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features,” *Biopolymers*, vol. 22, no. 12, pp. 2577–2637, Dec. 1983.

[29] J. Martin, G. Letellier, A. Marin, J.-F. Taly, A. G. de Brevern, and J.-F. Gibrat, “Protein secondary structure assignment revisited: a detailed analysis of different assignment methods,” *BMC Struct. Biol.*, vol. 5, p. 17, Sep. 2005.

[30] A. Zemla, Č. Venclovas, K. Fidelis, and B. Rost, “A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment,” *Proteins Struct. Funct. Bioinforma.*, vol. 34, no. 2, pp. 220–223, Feb. 1999.

[31] A. Drozdetskiy, C. Cole, J. Procter, and G. J. Barton, “JPred4: a protein secondary structure prediction server,” *Nucleic Acids Res.*, vol. 43, no. W1, pp. W389–W394, Jul. 2015.

[32] G. Pollastri and A. McLysaght, “Porter: a new, accurate server for protein secondary structure prediction,” *Bioinformatics*, vol. 21, no. 8, pp. 1719–1720, Dec. 2005.

[33] S. Wang, W. Li, S. Liu, and J. Xu, “RaptorX-Property: a web server for protein structure property prediction,” *Nucleic Acids Res.*, vol. 44, no. W1, pp. W430–W435, Apr. 2016.

[34] R. Heffernan, Y. Yang, K. Paliwal, and Y. Zhou, “Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers and solvent accessibility,” *Bioinformatics*, vol. 33, no. 18, pp. 2842–2849, Sep. 2017.

[35] C. N. Magnan and P. Baldi, “SSpro/ACCpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity,” *Bioinformatics*, vol. 30, no. 18, pp. 2592–2597, May 2014.

[36] J. A. Cuff, M. E. Clamp, A. S. Siddiqui, M. Finlay, and G. J. Barton, “JPred: a consensus secondary structure prediction server.,” *Bioinformatics*, vol. 14, no. 10, pp. 892–893, Jan. 1998.

[37] R. D. Finn, J. Clements, and S. R. Eddy, “HMMER web server: interactive sequence similarity searching,” *Nucleic Acids Res.*, vol. 39, no. Web Server issue, pp. W29–W37, Jul. 2011.

[38] A. M. Waterhouse, J. B. Procter, D. M. A. Martin, M. Clamp, and G. J. Barton, “Jalview Version 2—a multiple sequence alignment editor and analysis workbench,” *Bioinformatics*, vol. 25, no. 9, pp. 1189–1191, May 2009.

[39] D. W. A. Buchan, S. M. Ward, A. E. Lobley, T. C. O. Nugent, K. Bryson, and D. T. Jones, “Protein annotation and modelling servers at University College London,” *Nucleic Acids Res.*, vol. 38, no. suppl_2, pp. W563–W568, Jul. 2010.

[40] C. Mirabello and G. Pollastri, “Porter, PaleAle 4.0: high-accuracy prediction of protein secondary structure and relative solvent accessibility,” *Bioinformatics*, vol. 29, no. 16, pp. 2056–2058, Jun. 2013.

[41] M. Torrisi, M. Kaleel, and G. Pollastri, “Porter 5: state-of-the-art ab initio prediction of protein secondary structure in 3 and 8 classes,” *bioRxiv*, p. 289033, Mar. 2018.

[42] R. Heffernan *et al.*, “Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning,” *Sci. Rep.*, vol. 5, Jun. 2015.

[43] J. L. Fauchère, M. Charton, L. B. Kier, A. Verloop, and V. Pliska, “Amino acid side chain parameters for correlation studies in biology and pharmacology,” *Int. J. Pept. Protein Res.*, vol. 32, no. 4, pp. 269–278, Oct. 1988.

[44] J. Cheng, A. Z. Randall, M. J. Sweredoski, and P. Baldi, “SCRATCH: a protein structure and structural feature prediction server,” *Nucleic Acids Res.*, vol. 33, no. suppl_2, pp. W72–W76, Jul. 2005.

[45] S. Pascarella, R. D. Persio, F. Bossa, and P. Argos, “Easy method to predict solvent accessibility from multiple protein sequence alignments,” *Proteins Struct. Funct. Bioinforma.*, vol. 32, no. 2, pp. 190–199, Aug. 1998.

[46] J. L. Cornette, K. B. Cease, H. Margalit, J. L. Spouge, J. A. Berzofsky, and C. DeLisi, “Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins,” *J. Mol. Biol.*, vol. 195, no. 3, pp. 659–685, Jun. 1987.

[47] B. Rost and C. Sander, “Conservation and prediction of solvent accessibility in protein families,” *Proteins*, vol. 20, no. 3, pp. 216–226, Nov. 1994.

[48] S. R. Holbrook, S. M. Muskal, and S. H. Kim, “Predicting surface exposure of amino acids from protein sequence,” *Protein Eng.*, vol. 3, no. 8, pp. 659–665, Aug. 1990.

[49] G. Pollastri, P. Baldi, P. Fariselli, and R. Casadio, “Prediction of coordination number and relative solvent accessibility in proteins,” *Proteins*, vol. 47, no. 2, pp. 142–153, May 2002.

[50] B. Adhikari, J. Hou, and J. Cheng, “DNCON2: improved protein contact prediction using two-level deep convolutional neural networks,” *Bioinformatics*.

[51] I. Walsh, D. Baù, A. J. Martin, C. Mooney, A. Vullo, and G. Pollastri, “Ab initio and template-based prediction of multi-class distance maps by two-dimensional recursive neural networks,” *BMC Struct. Biol.*, vol. 9, p. 5, Jan. 2009.

[52] C. Mooney and G. Pollastri, “Beyond the Twilight Zone: Automated prediction of structural properties of proteins by recursive neural networks and remote homology information,” *Proteins Struct. Funct. Bioinforma.*, vol. 77, no. 1, pp. 181–190, Oct. 2009.

[53] Z. Yuan, “Better prediction of protein contact number using a support vector regression analysis of amino acid sequence,” *BMC Bioinformatics*, vol. 6, p. 248, Oct. 2005.

[54] R. Heffernan *et al.*, “Highly accurate sequence-based prediction of half-sphere exposures of amino acid residues in proteins,” *Bioinformatics*, vol. 32, no. 6, pp. 843–849, Mar. 2016.

[55] Li Xia and Pan Xian-Ming, “New method for accurate prediction of solvent accessibility from protein sequence,” *Proteins Struct. Funct. Bioinforma.*, vol. 42, no. 1, pp. 1–5, Nov. 2000.

[56] S. Ahmad, M. Gromiha, H. Fawareh, and A. Sarai, “ASAView: Database and tool for solvent accessibility representation in proteins,” *BMC Bioinformatics*, vol. 5, p. 51, May 2004.

[57] Hamelryck Thomas, “An amino acid has two sides: A new 2D measure provides a different view of solvent exposure,” *Proteins Struct. Funct. Bioinforma.*, vol. 59, no. 1, pp. 38–48, Feb. 2005.

[58] R. Kuang, C. S. Leslie, and A.-S. Yang, “Protein backbone angle prediction with machine learning approaches,” *Bioinformatics*, vol. 20, no. 10, pp. 1612–1621, Jul. 2004.

[59] J. Lyons *et al.*, “Predicting backbone Cα angles and dihedrals from protein sequences by stacked sparse auto-encoder deep neural network,” *J. Comput. Chem.*, vol. 35, no. 28, pp. 2040–2046, Oct. 2014.

[60] Y. Huang and C. Bystroff, “Improved pairwise alignments of proteins in the Twilight Zone using local structure predictions,” *Bioinformatics*, vol. 22, no. 4, pp. 413–422, Feb. 2006.

[61] E. Faraggi, Y. Yang, S. Zhang, and Y. Zhou, “Predicting Continuous Local Structure and the Effect of Its Substitution for Secondary Structure in Fragment-Free Protein Structure Prediction,” *Structure*, vol. 17, no. 11, pp. 1515–1527, Nov. 2009.

[62] C. Bystroff, V. Thorsson, and D. Baker, “HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins,” *J. Mol. Biol.*, vol. 301, no. 1, pp. 173–190, Aug. 2000.

[63] M. J. Wood and J. D. Hirst, “Protein secondary structure prediction with dihedral angles,” *Proteins Struct. Funct. Bioinforma.*, vol. 59, no. 3, pp. 476–481, May 2005.

[64] C. Mooney, A. Vullo, and G. Pollastri, “Protein Structural Motif Prediction in Multidimensional ø-ψ Space Leads to Improved Secondary Structure Prediction,” *J. Comput. Biol.*, vol. 13, no. 8, pp. 1489–1502, Oct. 2006.

[65] A. g. de Brevern, C. Etchebest, and S. Hazout, “Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks,” *Proteins Struct. Funct. Bioinforma.*, vol. 41, no. 3, pp. 271–287, Nov. 2000.

[66] D. Ting, G. Wang, M. Shapovalov, R. Mitra, M. I. Jordan, and R. L. D. Jr, “Neighbor-Dependent Ramachandran Probability Distributions of Amino Acids Developed from a Hierarchical Dirichlet Process Model,” *PLOS Comput. Biol.*, vol. 6, no. 4, p. e1000763, Apr. 2010.

[67] G. E. Sims, I.-G. Choi, and S.-H. Kim, “Protein conformational space in higher order φ-Ψ maps,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 102, no. 3, pp. 618–621, Jan. 2005.

[68] M. Vassura, L. Margara, P. Di Lena, F. Medri, P. Fariselli, and R. Casadio, “Reconstruction of 3D Structures From Protein Contact Maps,” *IEEEACM Trans Comput Biol Bioinforma.*, vol. 5, no. 3, pp. 357–367, Jul. 2008.

[69] M. Vendruscolo, E. Kussell, and E. Domany, “Recovery of protein structure from contact maps,” *Fold. Des.*, vol. 2, no. 5, pp. 295–306, Oct. 1997.

[70] M. Vassura *et al.*, “Blurring contact maps of thousands of proteins: what we can learn by reconstructing 3D structure,” *BioData Min.*, vol. 4, p. 1, Jan. 2011.

[71] P. Kukic, C. Mirabello, G. Tradigo, I. Walsh, P. Veltri, and G. Pollastri, “Toward an accurate prediction of inter-residue distances in proteins using 2D recursive neural networks,” *BMC Bioinformatics*, vol. 15, p. 6, Jan. 2014.

[72] F. Pazos, M. Helmer-Citterich, G. Ausiello, and A. Valencia, “Correlated mutations contain information about protein-protein interaction,” *J. Mol. Biol.*, vol. 271, no. 4, pp. 511–523, Aug. 1997.

[73] O. Olmea and A. Valencia, “Improving contact predictions by the combination of correlated mutations and other sources of sequence information,” *Fold. Des.*, vol. 2, pp. S25–S32, Jun. 1997.

[74] U. Göbel, C. Sander, R. Schneider, and A. Valencia, “Correlated mutations and residue contacts in proteins,” *Proteins*, vol. 18, no. 4, pp. 309–317, Apr. 1994.

[75] P. Fariselli, O. Olmea, A. Valencia, and R. Casadio, “Prediction of contact maps with neural networks and correlated mutations,” *Protein Eng. Des. Sel.*, vol. 14, no. 11, pp. 835–843, Nov. 2001.

[76] J. Cheng and P. Baldi, “Improved residue contact prediction using support vector machines and a large feature set,” *BMC Bioinformatics*, vol. 8, p. 113, Apr. 2007.

[77] D. T. Jones, D. W. A. Buchan, D. Cozzetto, and M. Pontil, “PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments,” *Bioinformatics*, vol. 28, no. 2, pp. 184–190, Jan. 2012.

[78] P. Di Lena, P. Fariselli, L. Margara, M. Vassura, and R. Casadio, “Is There an Optimal Substitution Matrix for Contact Prediction with Correlated Mutations?,” *IEEEACM Trans Comput Biol Bioinforma.*, vol. 8, no. 4, pp. 1017–1028, Jul. 2011.

[79] J. Schaarschmidt, B. Monastyrskyy, A. Kryshtafovych, and A. M. J. J. Bonvin, “Assessment of contact predictions in CASP12: co-evolution and deep learning coming of age,” *Proteins Struct. Funct. Bioinforma.*, Nov. 2017.

[80] L. Bartoli, E. Capriotti, P. Fariselli, P. L. Martelli, and R. Casadio, “The pros and cons of predicting protein contact maps,” *Methods Mol. Biol. Clifton NJ*, vol. 413, pp. 199–217, 2008.

[81] A. Roy, A. Kucukural, and Y. Zhang, “I-TASSER: a unified platform for automated protein structure and function prediction,” *Nat. Protoc.*, vol. 5, no. 4, pp. 725–738, Apr. 2010.

[82] T. Kosciolek and D. T. Jones, “De Novo Structure Prediction of Globular Proteins Aided by Sequence Variation-Derived Contacts,” *PLOS ONE*, vol. 9, no. 3, p. e92197, Mar. 2014.

[83] L. N. Kinch, W. Li, B. Monastyrskyy, A. Kryshtafovych, and N. V. Grishin, “Assessment of CASP11 Contact-Assisted Predictions,” *Proteins*, vol. 84, no. Suppl 1, pp. 164–180, Sep. 2016.

[84] A. Schlessinger, M. Punta, and B. Rost, “Natively unstructured regions in proteins identified from contact predictions,” *Bioinforma. Oxf. Engl.*, vol. 23, no. 18, pp. 2376–2384, Sep. 2007.

[85] D. E. Kim, F. DiMaio, R. Y.-R. Wang, Y. Song, and D. Baker, “One contact for every twelve residues allows robust and accurate topology-level protein structure modeling,” *Proteins*, vol. 82, no. 0 2, pp. 208–218, Feb. 2014.

[86] G. Pollastri and P. Baldi, “Prediction of contact maps by GIOHMMs and recurrent neural networks using lateral propagation from all four cardinal corners,” *Bioinformatics*, vol. 18, no. suppl_1, pp. S62–S70, Jul. 2002.

[87] G. Pollastri, P. Baldi, P. Fariselli, and R. Casadio, “Improved prediction of the number of residue contacts in proteins by recurrent neural networks,” *Bioinformatics*, vol. 17, no. suppl_1, pp. S234–S242, Jun. 2001.

[88] P. Di Lena, P. Fariselli, L. Margara, M. Vassura, and R. Casadio, “Fast overlapping of protein contact maps by alignment of eigenvectors,” *Bioinformatics*, vol. 26, no. 18, pp. 2250–2258, Sep. 2010.

[89] L. Kaján, T. A. Hopf, M. Kalaš, D. S. Marks, and B. Rost, “FreeContact: fast and free software for protein contact prediction from residue co-evolution,” *BMC Bioinformatics*, vol. 15, p. 85, Mar. 2014.

[90] S. Seemayer, M. Gruber, and J. Söding, “CCMpred—fast and precise prediction of protein residue–residue contacts from correlated mutations,” *Bioinformatics*, vol. 30, no. 21, pp. 3128–3130, Nov. 2014.

[91] B. Monastyrskyy, D. D’Andrea, K. Fidelis, A. Tramontano, and A. Kryshtafovych, “Evaluation of residue–residue contact prediction in CASP10,” *Proteins Struct. Funct. Bioinforma.*, vol. 82, pp. 138–153, Feb. 2014.

[92] A. Zemla, “LGA: a method for finding 3D similarities in protein structures,” *Nucleic Acids Res.*, vol. 31, no. 13, pp. 3370–3374, Jul. 2003.

[93] R. M. MacCallum, “Striped sheets and protein contact prediction,” *Bioinformatics*, vol. 20, no. suppl_1, pp. i224–i231, Aug. 2004.

[94] A. N. Tegge, Z. Wang, J. Eickholt, and J. Cheng, “NNcon: improved protein contact map prediction using 2D-recursive neural networks,” *Nucleic Acids Res.*, vol. 37, no. suppl_2, pp. W515–W518, Jul. 2009.

[95] P. Di Lena, K. Nagata, and P. Baldi, “Deep architectures for protein contact map prediction,” *Bioinformatics*, vol. 28, no. 19, pp. 2449–2457, Oct. 2012.

[96] AJM Martin, C Mooney, I Walsh, and G Pollastri, “Contact Map Prediction by Machine Learning,” in *Introduction to Protein Structure Prediction: Methods and Algorithms*, pp. 137–163.

[97] D. W. A. Buchan and D. T. Jones, “Improved protein contact predictions with the MetaPSICOV2 server in CASP12,” *Proteins*, vol. 86 Suppl 1, pp. 78–83, Mar. 2018.

[98] J. Eickholt and J. Cheng, “Predicting protein residue–residue contacts using deep networks and boosting,” *Bioinformatics*, vol. 28, no. 23, pp. 3066–3072, Dec. 2012.

[99] L. S. Johnson, S. R. Eddy, and E. Portugaly, “Hidden Markov model speed heuristic and iterative HMM search procedure,” *BMC Bioinformatics*, vol. 11, p. 431, Aug. 2010.

[100] T. Kosciolek and D. T. Jones, “Accurate contact predictions using covariation techniques and machine learning,” *Proteins*, vol. 84 Suppl 1, pp. 145–151, Sep. 2016.

[101] B. Monastyrskyy, D. D’Andrea, K. Fidelis, A. Tramontano, and A. Kryshtafovych, “New encouraging developments in contact prediction: Assessment of the CASP11 results,” *Proteins*, vol. 84 Suppl 1, pp. 131–144, Sep. 2016.

[102] S. Wang, S. Sun, and J. Xu, “Analysis of deep learning methods for blind protein contact prediction in CASP12,” *Proteins*, vol. 86 Suppl 1, pp. 67–77, Mar. 2018.

[103] C. Mooney, A. Cessieux, D. C. Shields, and G. Pollastri, “SCL-Epred: a generalised de novo eukaryotic protein subcellular localisation predictor,” *Amino Acids*, vol. 45, no. 2, pp. 291–299, Aug. 2013.

View publication stats