Chapter 10
Protein Structure Annotations

Mirko Torrisi and Gianluca Pollastri

Contents

10.1 Introduction to 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

M. Torrisi (✉) · G. Pollastri
School of Computer Science, University College Dublin, Dublin, Ireland
e-mail: gianluca.pollastri@ucd.ie

N. A. Shaik et al. (eds.), *Essentials of Bioinformatics, Volume I*,
https://doi.org/10.1007/978-3-030-02634-9_10

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 characterisation 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) (Berman et al. 2000) and the Universal Protein Resource (The UniProt Consortium 2016) – and their complexity make proteins an ideal domain where to apply the most recent and sophisticated ML techniques, such as deep learning (LeCun et al. 2015). Nonetheless, there are pitfalls to avoid and best practices to follow to correctly train and test any ML method on protein sequences (Walsh et al. 2016).

Deep learning is a collection of methods and techniques to efficiently train nuanced parametric models such as neural networks (NN) with multiple hidden layers (Schmidhuber 2015). 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 (Haas et al. n.d.). 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 (Pollastri et al. 2007).

Position-Specific Iterated Basic Local Alignment Search Tool (PSI-BLAST) (Altschul et al. 1997) 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, and 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, maximise 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 (Schäffer et al. 2001). 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 (Jones and Swindells 2002).

HHblits (Remmert et al. 2012) 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.

10.2 Secondary Structure

SS prediction is one of the great historical challenges in bioinformatics (Rost 2001; Yang et al. 2016). 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 (Pauling and Corey 1951). Notably, the very first high-resolution protein structure was determined only in 1958 (and led to a Nobel Prize to Kendrew and Perutz) (Kendrew et al. 1960; Perutz et al. 1960). 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 (Chou and Fasman 1974). 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 (Rost 2001). In the 1990s, 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 (Yang et al. 2016). Thus, the third generation of SS predictors was the first able to predict at better than 70% accuracy (Rost and Sander 1993), efficiently exploit PSI-BLAST (Jones 1999) and implement deep NN (Baldi et al. 1999). 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 (Aloy et al. 2003).

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 (Jones et al. 2015; Wang et al. 2017; Vullo et al. 2006), the recognition of protein folds (Yang et al. 2011) and protein tertiary structure (Baú et al. 2006)). 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 characterise the proteins in a common fold (Murzin et al. 1995).

The theoretical limit of SS prediction is usually set at 88–90% accuracy per AA (Yang et al. 2016). 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) (Kabsch and Sander 1983), 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 (Martin et al. 2005). While DSSP aims to provide an unambiguous and physically meaningful assignment, the PDB represents the ground truth in structural proteomics (Berman et al. 2000).

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 eight 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 three of the predictors presented here (Porter5, RaptorX-Property and SSpro) can predict in both three states and eight states. The DSSP classifications of SS in eight states are the following:

- G = three-turn helix (310-helix), minimum length three residues
- H = four-turn helix (α-helix), minimum length four residues
- I = five-turn helix (π-helix), minimum length five residues

- T = hydrogen bonded turn (three, four or five turn)
- E = extended strand (in β-sheet conformation), minimum length two 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-state prediction, respectively – or at the segment overlap score (SOV), i.e. the overlap between the predicted and the real segments of SS (Zemla et al. 1999), 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.

Table 10.1 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.

10.2.1 Jpred

Jpred is an SS predictor which was initially released in 1998 (Cuff et al. 1998). Jpred4 (Drozdetskiy et al. 2015), the last available version, has been released in 2015 to update HMMer (Finn et al. 2011) 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. FAQ and tutorials are available on its website (Fig. 10.1).

Table 10.1 Name, web server and notes on special features of every SS predictor presented in this chapter

Name	Web server	Notes
Jpred (Drozdetskiy et al. 2015)	http://www.compbio.dundee.ac.uk/jpred4/	HHMer, MSA as input, API
PSIPRED (Jones 1999)	http://bioinf.cs.ucl.ac.uk/psipred/	BLAST, cloud version, MSA as input
Porter (Pollastri and McLysaght 2005)	http://distilldeep.ucd.ie/porter/	three- or eight-states, HHblits or PSI-BLAST, light standalone
RaptorX-Property (Wang et al. 2016)	http://raptorx.uchicago.edu/StructurePropertyPred/predict/	three- or eight-states, no PSI-BLAST (only HHblits), option for no evolutionary information
SPIDER3 (Heffernan et al. 2017)	http://sparks-lab.org/server/SPIDER3/	Numpy or Tensorflow, HHblits and PSI-BLAST
SSpro (Magnan and Baldi 2014)	http://scratch.proteomics.ics.uci.edu/	three- or eight-states, BLAST, template-based

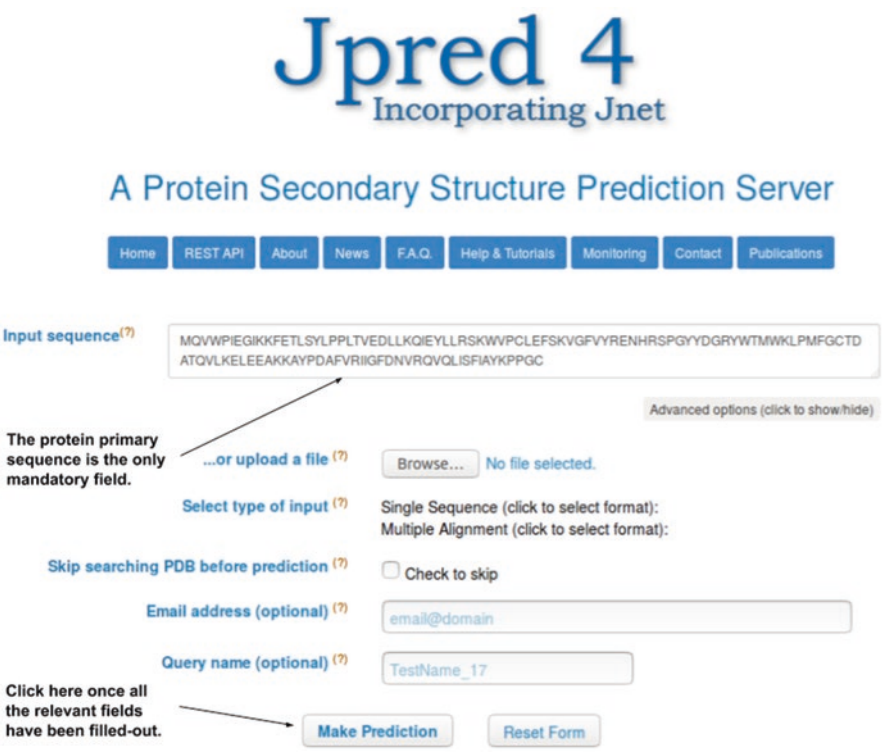

Fig. 10.1 The homepage of Jpred4. The input sequence is the only requirement while more options are made available

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 (Berman et al. 2000) 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 (The UniProt Consortium 2016) 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 (Waterhouse et al. 2009) (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.

10.2.2 PSIPRED

PSIPRED is a high-quality SS predictor freely available since 1999 (Jones 1999). 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, and thus provides less informative outputs for PSIPRED. BLAST+ is experimentally supported by PSIPRED (Fig. 10.2).

The web server of PSIPRED (Buchan et al. 2010), 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.

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 pro-

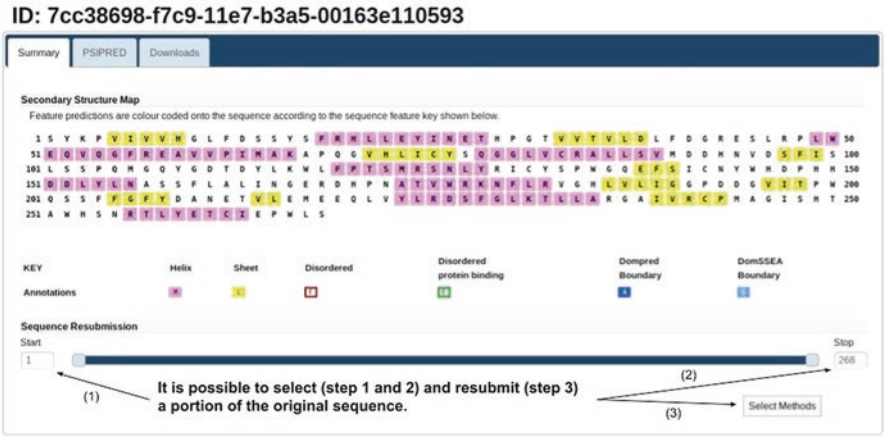

Fig. 10.2 A typical result page of PSIPRED web server. All the AA are listed and coloured according to the predicted SS class

cessing 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/.

10.2.3 Porter

Porter is a high-quality SS predictor which has been developed starting in 2005 (Pollastri and McLysaght 2005) and improved since then (Pollastri et al. 2007; Mirabello and Pollastri 2013). Porter is built on carefully tuned and trained ensembles of cascaded bidirectional recurrent neural networks (Baldi et al. 1999). It is typically built on very large datasets, which are released as well. Its last release (v5) is available as web server and standalone (Torrisi et al. 2018). 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-state and eight-state SS (Fig. 10.3).

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 64 KB 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

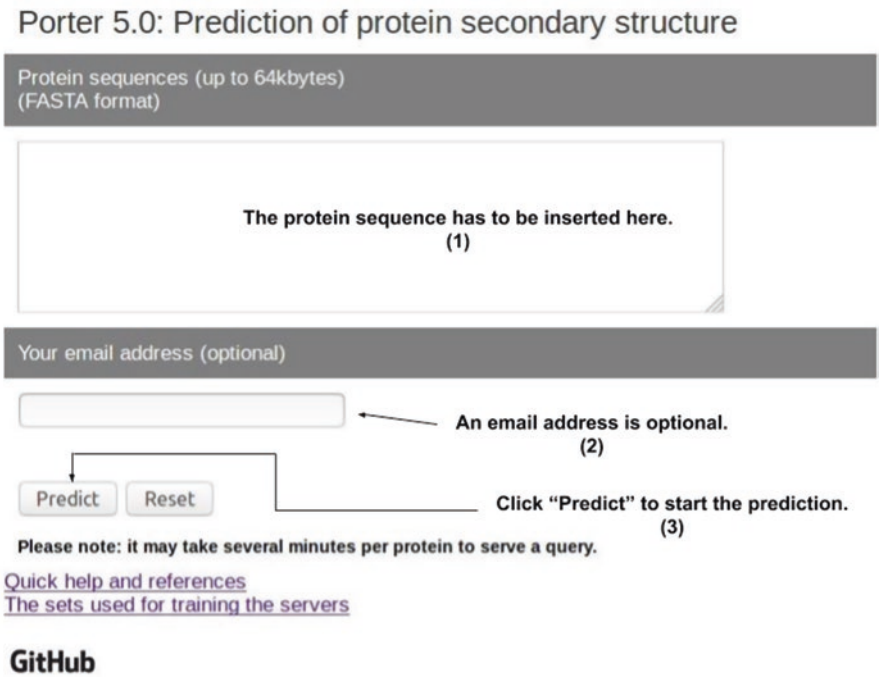

Fig. 10.3 The input form of Porter5. Around 200 proteins can be submitted at once in FASTA format

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.

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 two 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.

10.2.4 RaptorX-Property

RaptorX-Property, released in 2016, is a collection of methods to predict one-dimensional protein annotations (Wang et al. 2016). 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 (Fig. 10.4).

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 three- and eight-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.

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 50 times the storage required by Porter5.

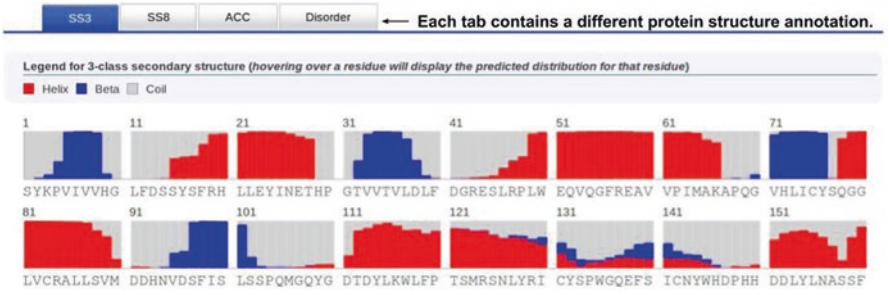

Fig. 10.4 A partial view of the result page of RaptorX-Property. Each bar in the charts represents the individual confidence

10.2.5 SPIDER3

SPIDER3 is the second version of a recent SS predictor first released in 2015 (Heffernan et al. 2015). Its last release is composed by 2 NN, the first of which predicts SS while the second predicts backbone angles, contact numbers and SA (Heffernan et al. 2017). It internally represents each AA using seven representative physiochemical properties (Fauchère et al. 1988). 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 (Fig. 10.5).

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 visualise the queue of jobs submitted from one's IP address. The result page presents the query sequence and the predicted SS and

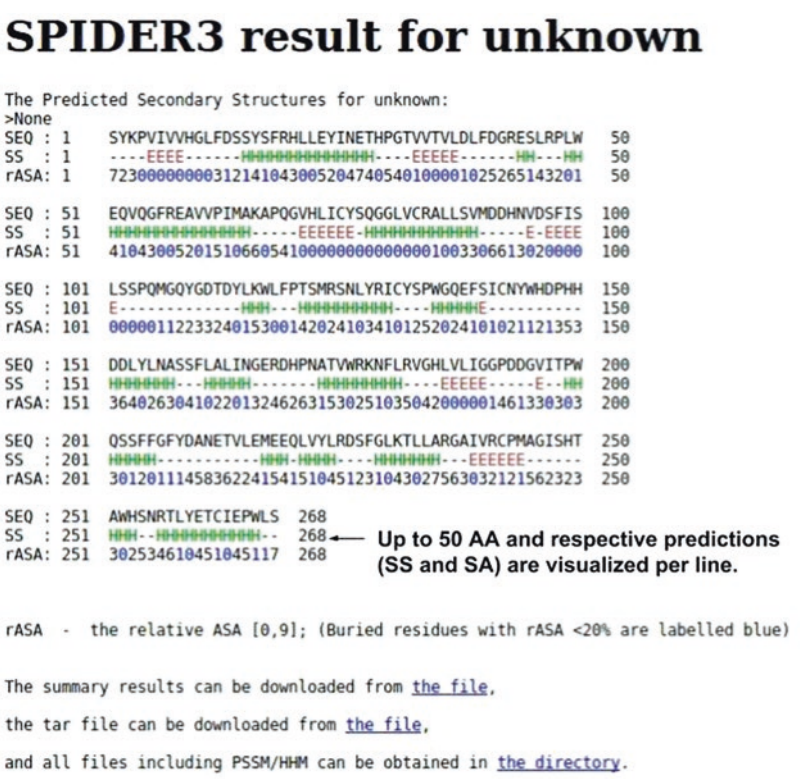

Fig. 10.5 An output example of SPIDER3

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 four 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.

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 101 MB and 117 MB, respectively, without considering the library of choice.

10.2.6 SSpro

SSpro is a historical SS predictor developed starting in 1999 (Baldi et al. 1999; Magnan and Baldi 2014). 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 (Magnan and Baldi 2014). More specifically, it exploits PSI-BLAST to look for homologues at both sequence and structure level (Pollastri et al. 2007). In other words, SSpro v5 has an additional final step in which it looks for similar proteins in the PDB (Fig. 10.6).

SSpro is available at http://scratch.proteomics.ics.uci.edu/ as part of the SCRATCH protein predictor (Cheng et al. 2005). 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-state (SSpro) and eight-state (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 five 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, 97 MB without databases).

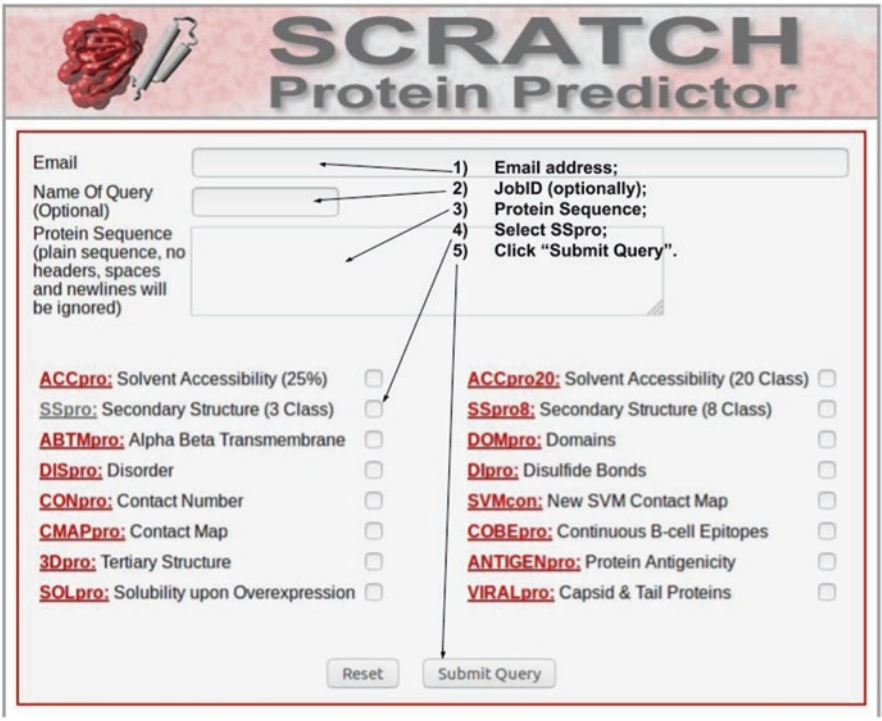

Fig. 10.6 A view of SCRATCH protein predictor

10.3 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 1990s and highly motivated from the successes obtained developing the third generation of SS predictors (Pascarella et al. 1998). In fact, similarly to SS prediction but sometimes with some time delay, mathematical and statistical methods (Cornette et al. 1987), NN (Rost and Sander 1994), evolutionary information (Holbrook et al. 1990) and deep NN (Pollastri et al. 2002) have been increasingly put to work to predict SA.

Although SA is less conserved than SS in homologous sequences (Rost and Sander 1994), 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 (Jones et al. 2015; Wang et al. 2017; Adhikari et al. 2017; Walsh et al. 2009) – protein fold recognition (Yang et al. 2011) and protein tertiary structure (Mooney and Pollastri 2009). Notably, a strong (negative) correlation of −0.734 between SA and contact numbers

Table 10.2 Solvent Accessibility prediction servers

Name	Web server	Notes
ACCpro (Baldi et al. 1999)	http://scratch.proteomics.ics.uci.edu/	two-state or twenty-states, BLAST, template-based
PaleAle (Pollastri et al. 2007)	http://distilldeep.ucd.ie/paleale/	four-states, HHblits or PSI-BLAST, light standalone
RaptorX-Property (Heffernan et al. 2017)	http://raptorx.uchicago.edu/StructurePropertyPred/predict/	three-states, no PSI-BLAST (only HHblits), option for no evolutionary information
SPIDER3 (Heffernan et al. 2017)	http://sparks-lab.org/server/SPIDER3/	HSE and ASA in R, Numpy or Tensorflow, HHblits and PSI-BLAST

has been observed by Yuan (Yuan 2005) and is motivating the development of predictors for contact number as a possible alternative to SA predictors (Heffernan et al. 2016).

Though there are promising examples of successful NN predictors considering adjacent AA to predict SA since the 1990s (Holbrook et al. 1990), different methods such as linear regression (Xia and Pan 2000) or substitution matrices (Pascarella et al. 1998) have been assessed, but the state of the art has been represented by deep NN since 2002 (Pollastri et al. 2002). Thus, all the SA predictors described below (and summarised in Table 10.2) implement deep NN (Wang et al. 2016; Heffernan et al. 2017; Magnan and Baldi 2014; Mirabello and Pollastri 2013) predicting SA as anything between a two-state problem – i.e. buried and exposed with an average two-state accuracy greater than 80% – and 20-state problem.

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 normalising the relative SA value observed by the maximum possible value of accessibility for the specific residue according to the DSSP (Kabsch and Sander 1983). The ASA of a protein can be visualised with ASAview, a tool developed in 2004 that requires real values extracted from the PDB or coming from predicted ASA (Ahmad et al. 2004). More recently, a different approach to measuring the SA, called half-sphere exposure (HSE), has been designed by Hamelryck (Thomas 2005). 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 (Thomas 2005). SPIDER3 can predict both HSE and ASA using real numbers (Heffernan et al. 2017).

10.3.1 ACCpro

ACCpro is a historical SA predictor initially released in 2002 (Pollastri et al. 2002). 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

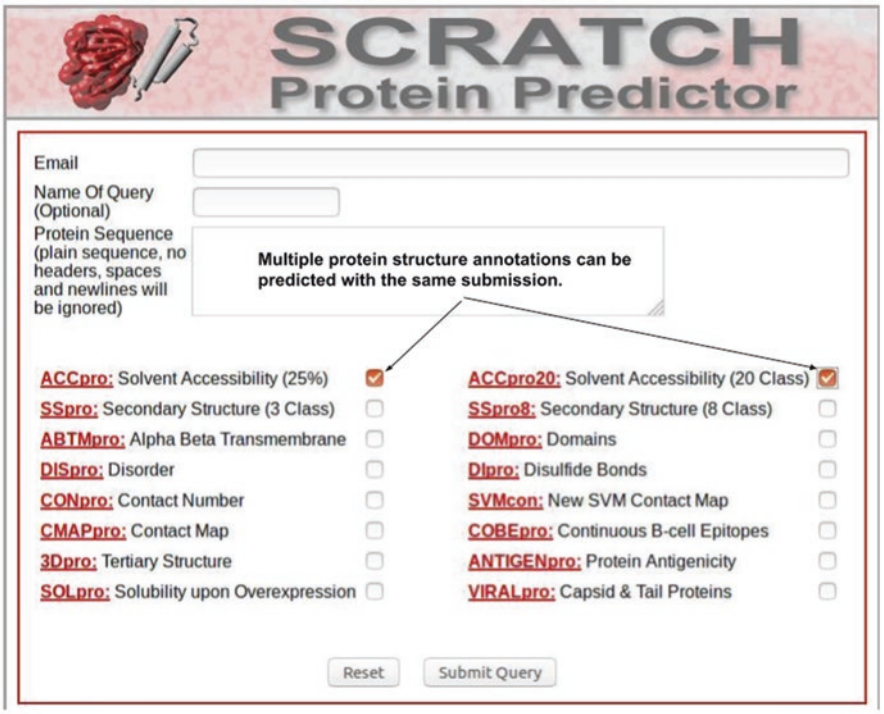

Fig. 10.7 A view of SCRATCH protein predictor where both ACCpro predictors have been selected

predictions (Magnan and Baldi 2014). 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 20 states from 0–95% with incremental steps of 5%, i.e. ACCpro classifies 20 classes, starting from 0–5% to 95–100% of SA (Fig. 10.7).

The web server of ACCpro and ACCpro20 is available at http://scratch.proteomics.ics.uci.edu/ as part of SCRATCH (Cheng et al. 2005). 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).

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.

10.3.2 PaleAle

PaleAle is a historical SA predictor developed in parallel with Porter (see Secondary Structure, Porter) since 2007 (Pollastri et al. 2007; Mirabello and Pollastri 2013) and is also based on ensembles of cascaded bidirectional recurrent neural networks (Baldi et al. 1999). PaleAle has been the first template-based SA predictor (Pollastri et al. 2007), while PaleAle (v5) is now able to predict four-state 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 (Fig. 10.8).

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.

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.

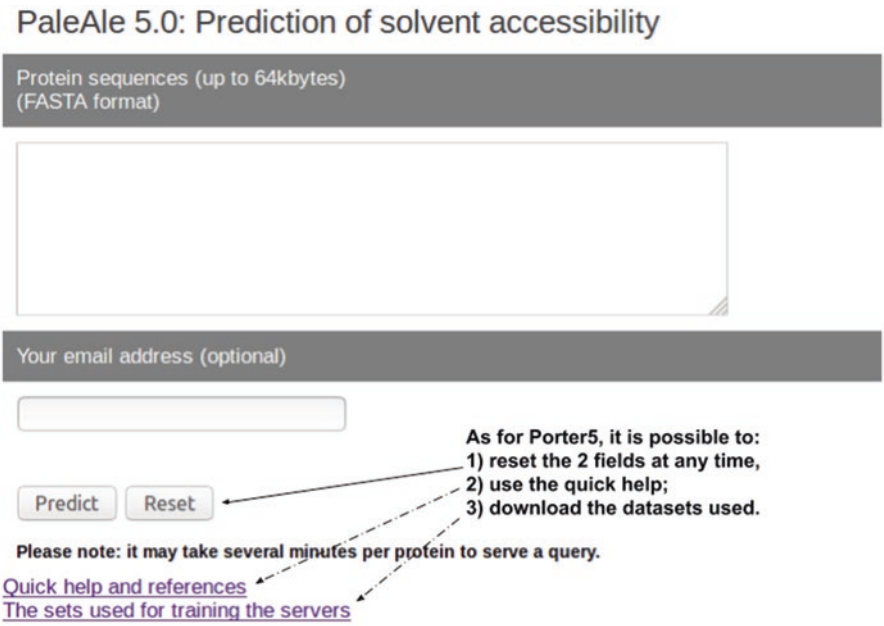

Fig. 10.8 A view of PaleAle5 where the reset button and the links are highlighted

10.3.3 RaptorX-Property

RaptorX-Property, described in section Secondary Structure, is 2016 suite of predictors able to predict SA, SS and disorder regions (Wang et al. 2016). 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 (Remmert et al. 2012) to gather evolutionary information (Fig. 10.9).

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 (Fig. 10.9 shows the three-state SA). The web server and the released standalone are described in section Secondary Structure, RaptorX-Property.

10.3.4 SPIDER3

SPIDER has been able to predict SA, SS and TA since 2015 (Heffernan et al. 2015) and was updated in 2017 (Heffernan et al. 2017). 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 (Heffernan et al. 2015). SPIDER2 has been the first HSE predictor (Heffernan et al. 2016), while SPIDER3 predicts HSEα-up and HSEα-down using real numbers, although Heffernan et al. reports result also in HSEβ-up and HSEβ-down (Heffernan et al. 2017).

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 Figs. 10.5 and 10.10).

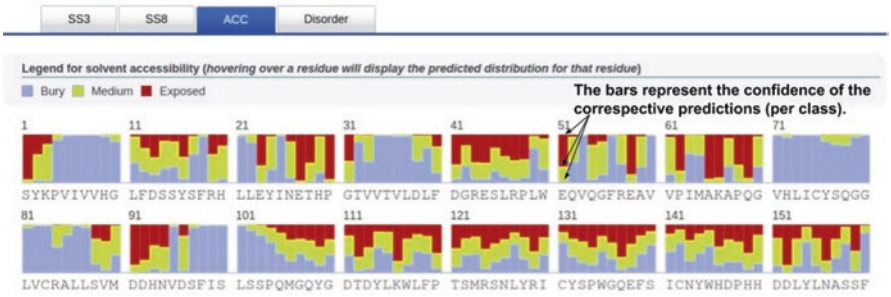

Fig. 10.9 The view on the predicted three-state SA performed by RaptorX-Property

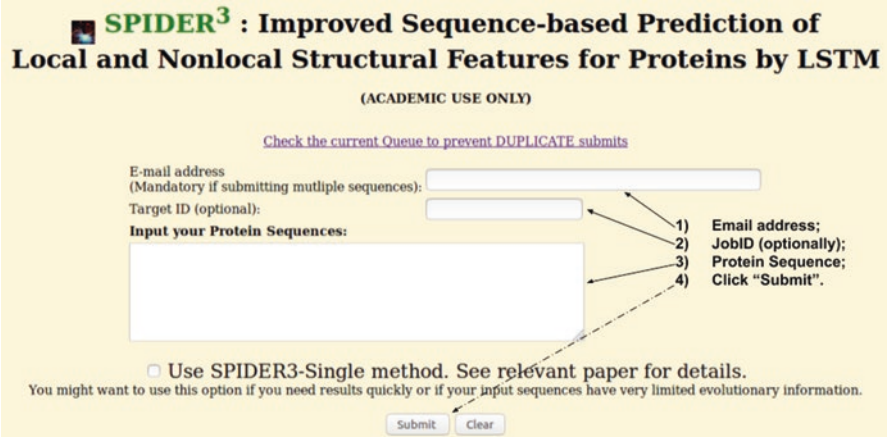

Fig. 10.10 A view of the input window of SPIDER3. The steps to follow to start a prediction are highlighted

10.4 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 (Kuang et al. 2004), 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) (Lyons et al. 2014). Thus, different annotations (i.e. SS, ϕ/ψ and θ/τ) can be adopted to describe the backbone of a protein (Fig. 10.11).

TA are essentially an alternative representation of local structure with respect to SS. Both TA and SS have been successfully used as restraints towards sequence alignment (Huang and Bystroff 2006), protein folding (Yang et al. 2011) and tertiary structure prediction (Faraggi et al. 2009). HMM (Bystroff et al. 2000), support vector machines (SVM) (Kuang et al. 2004) and several architectures of NN (e.g. iterative (Heffernan et al. 2017; Heffernan et al. 2015) and cascade-correlation (Wood and Hirst 2005)) have been analysed to predict TA since 2000. NN are currently the main tool to predict TA, in parallel with protein SS (Heffernan et al. 2017) or sequentially after it (Wood and Hirst 2005; Mooney et al. 2006).

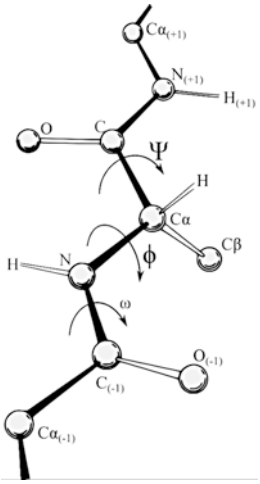

Fig. 10.11 Protein backbone dihedral angles phi, psi and omega; credits: https://commons.wikimedia.org/wiki/File:Protein_backbone_PhiPsiOmega_drawing.svg

Table 10.3 ϕ/ψ angles prediction web server

Name	Web server	Notes
Porter+ (Mooney et al. 2006)	http://distilldeep.ucd.ie/porter+	ϕ/ψ in 16 letters
SPIDER3 (Heffernan et al. 2017)	http://sparks-lab.org/server/SPIDER3/	ϕ/ψ and θ/τ , Numpy or Tensorflow

Table 10.4 Protein contact maps prediction servers

Name	Web server	Notes
DNCON (Adhikari et al. 2017)	http://sysbio.rnet.missouri.edu/dncon2/	Three coevolution algorithms, Computer Vision inspired
MetaPSICOV (Jones et al. 2015)	http://bioinf.cs.ucl.ac.uk/MetaPSICOV/	CCMpred, FreeContact and PSICOV, hydrogen bonds
RaptorX-Contact (Wang et al. 2017)	http://raptorx.uchicago.edu/ContactMap/	Inspired from Computer Vision, CCMpred only
XX-Stout (Walsh et al. 2009)	http://distilldeep.ucd.ie/xxstout/	Contact Density, template-based, multi-class CM

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 (De Brevern et al. 2000; Ting et al. 2010), multidimensional scaling (MDS) (Sims et al. 2005) and density plot (Kuang et al. 2004) approaches have been exploited to define different alphabets of various sizes (Tables 10.3 and 10.4).

10.4.1 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) (Mooney et al. 2006). Porter+ adopts an alphabet of 16 letters devised by Sims et al. using MDS on tetrapeptides (four contiguous residues) (Sims et al. 2005). 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 (Fig. 10.12).

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.

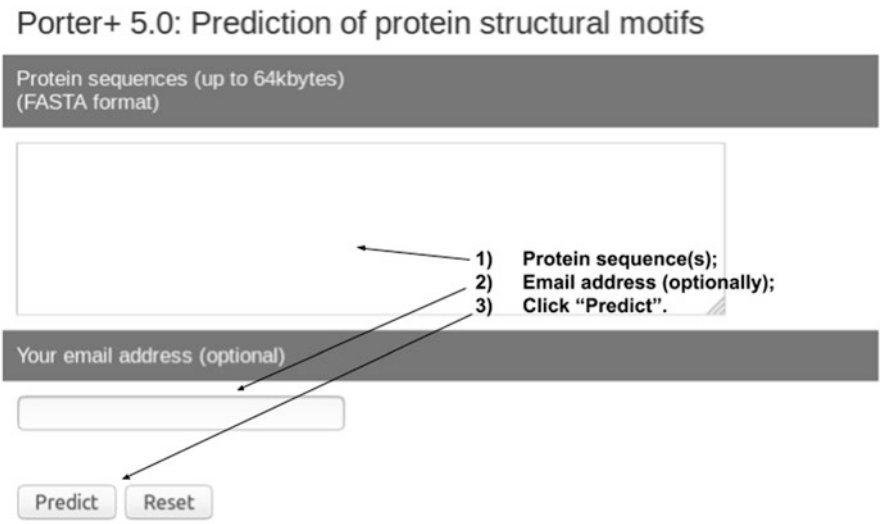

Fig. 10.12 A view of Porter+5 where the steps to start a prediction are highlighted

10.4.2 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 (Lyons et al. 2014). It has been further developed to also predict ϕ/ψ, in parallel with SS, SA and contact numbers (see the respective sections) (Heffernan et al. 2017; Heffernan et al. 2015). More details, regarding the pipeline implemented, the web server offered and the standalone available, are outlined in section Secondary Structure (Fig. 10.13).

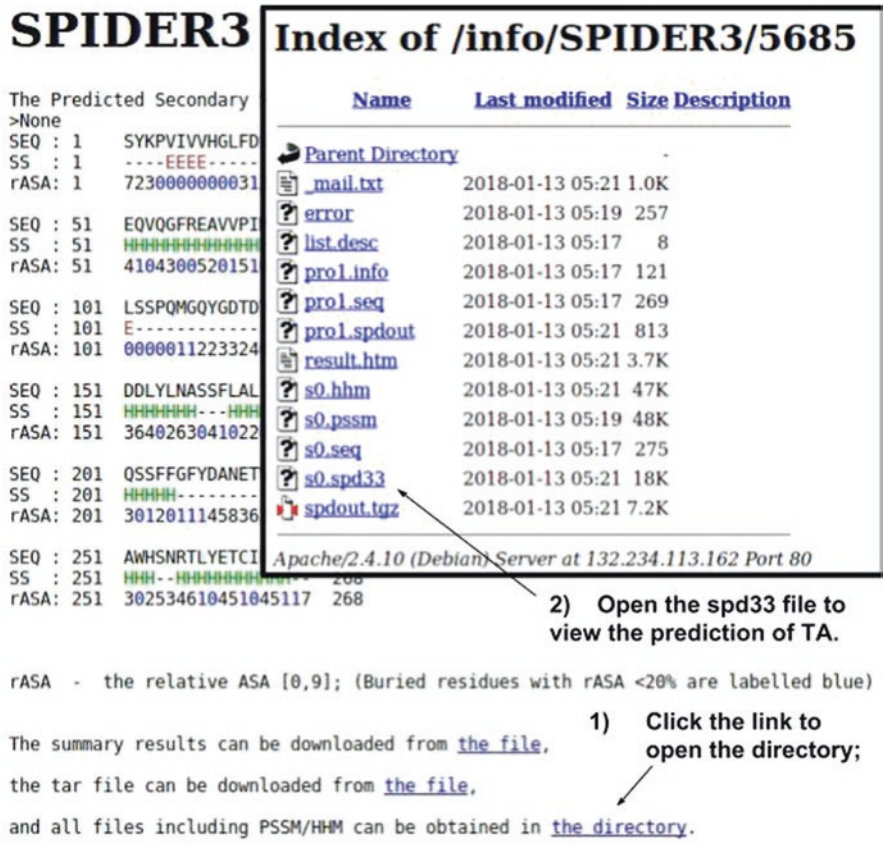

Fig. 10.13 A view of the results page of SPIDER3 where the steps to view the predicted TA are highlighted

10.5 Contact Maps

Contact Maps (CM) are the main two-dimensional protein structure annotation tools. 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 (Vassura et al. 2008; Vendruscolo et al. 1997). 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 (Vassura et al. 2011; Kukic et al. 2014). 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 (Pazos et al. 1997; Olmea and Valencia 1997; Göbel et al. 1994). These initial advancements led to the first generation of ML methods able to predict CM (Vullo et al. 2006; Fariselli et al. 2001; Cheng and Baldi 2007). 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 (Jones et al. 2012; Di Lena et al. 2011). 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 (Jones et al. 2015; Wang et al. 2016; Adhikari et al. 2017). Differently from the other protein annotations in this chapter, CM is currently assessed at CASP (Schaarschmidt et al. 2018) and CAMEO (Haas et al. n.d.).

The intrinsic properties of CM – namely, being compact and discrete two-state annotations, invariant to rotations and translations – make them a more appropriate target for ML techniques than protein tertiary structures or distance maps although still highly informative about the protein 3D structures (Bartoli et al. 2008). CM prediction is a typical intermediate step in many pipelines to predict protein tertiary structure (Mooney and Pollastri 2009; Roy et al. 2010; Kosciolek and Jones 2014). For example, it is a key component for contact-assisted structure prediction (Kinch et al. 2016), contact-assisted protein folding (Wang et al. 2017) and free and template-based modelling (Roy et al. 2010). CM have also been used to predict protein disorder (Schlessinger et al. 2007) and protein function (Pazos et al. 1997) and to detect challenging templates (Mooney and Pollastri 2009). In fact, even partial CM can greatly support robust and accurate protein structure modelling (Kim et al. 2014).

Being a 2D annotation, CM are typically gradually predicted starting from simpler but less informative 1D annotations – e.g. SA, SS and TA (Fariselli et al. 2001; Cheng and Baldi 2007; Pollastri and Baldi 2002). 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, Fig. 10.14 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 (Heffernan et al. 2017; Pollastri et al. 2001, 2002) to adjust and improve CM prediction (Olmea and Valencia 1997; Fariselli et al. 2001; Pollastri and Baldi 2002). Eigenvector decomposition has been used as a means for template search (Di Lena et al. 2010) and principal eigenvector (PE) prediction as an intermediate step towards CM prediction (Vullo et al. 2006). 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 (Pazos et al. 1997; Olmea and Valencia 1997; Göbel et al. 1994). 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 (Jones et al. 2012; Kaján et al. 2014; Seemayer et al. 2014).

As in Fig. 10.14, 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 (Bartoli et al. 2008). Notably, the number of AA in

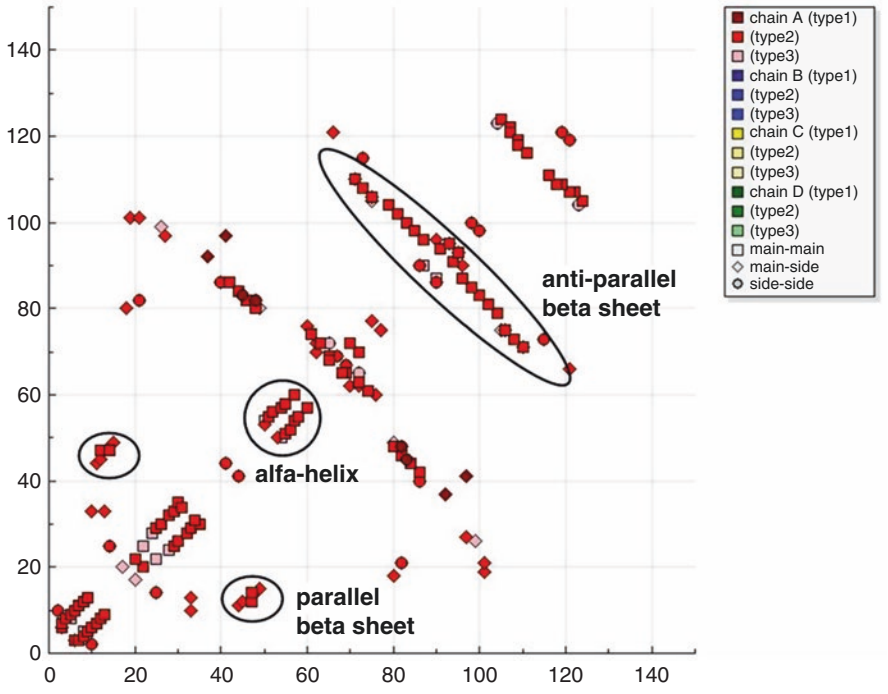

Fig. 10.14 CM with highlighted SS conformations; credits: https://commons.wikimedia.org/wiki/File:Elements_hb2.jpg

contact increases almost linearly with the protein length – i.e. shorter proteins are denser than longer ones (Bartoli et al. 2008). A pair of AA is in contact when the Euclidian distance between their C_β (or $C\alpha$, for glycine) atoms is closer than a given threshold. This threshold is usually set between 6 and 12 Å (8 Å at CASP (Schaarschmidt et al. 2018)), although values in the range of 10–18 Å may lead to better reconstructions (Vassura et al. 2008). 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 (Schaarschmidt et al. 2018). 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 (Schaarschmidt et al. 2018). Precision – i.e. the ratio between true contact and wrong contact (true contact + wrong contact) – is usually adopted to assess local (short range) contacts, i.e. involving AA within ten positions apart, and non-local (long range) contact, separately. Typically, CM predictors are evaluated at CASP through more complex measures (Schaarschmidt et al. 2018; Kinch et al. 2016; Monastyrskyy et al. 2014), 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; and root-mean-square deviation (RMSD) or TM-score, i.e. a measure more sensitive at the global (rather than local) structure than RMSD (Zemla 2003). Classic statistical and ML measures, such as the aforementioned precision, recall, F1 score and Matthews correlation coefficient (MCC), are also adopted in parallel with more unusual ones, such as alignment depth or entropy score (Schaarschmidt et al. 2018). 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 (Schaarschmidt et al. 2018).

Though correlated mutations and NN have been identified as promising instruments to also predict CM (Fariselli et al. 2001), pairwise contact potential (Schlessinger et al. 2007), self-organising maps (MacCallum 2004) and SVM (Cheng and Baldi 2007) have been used in the past. While 2D-BRNN (Pollastri and Baldi 2002; Tegge et al. 2009), multistage (Vullo et al. 2006; Di Lena et al. 2012) and template-based (Walsh et al. 2009) NN approaches have initially characterised the field (Martin et al. 2010), 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 (Adhikari et al. 2017; Buchan and Jones 2018) or multi-class maps (Kukic et al. 2014; Martin et al. 2010). The standard output format of any CM predictor is a text file organised in five columns as follows: the positions of the two AA in contact, a blank column, the set threshold (8 Å) and the confidence of each predicted contact.

10.5.1 DNCON2

DNCON has been initially released in 2012 (Eickholt and Cheng 2012), assessed at CASP10 (Monastyrskyy et al. 2014) and updated in 2017 (Adhikari et al. 2017). 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 8 Å. In the described two-stage approach, DNCON2 implements a total of six NN like RaptorX-Contact (Fig. 10.15). 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 24 h, 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 five 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 (Seemayer et al. 2014), FreeContact (Kaján et al. 2014), HHblits (Remmert et al. 2012), JackHMMER (Johnson et al. 2010) and PSICOV (Jones et al. 2012) 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

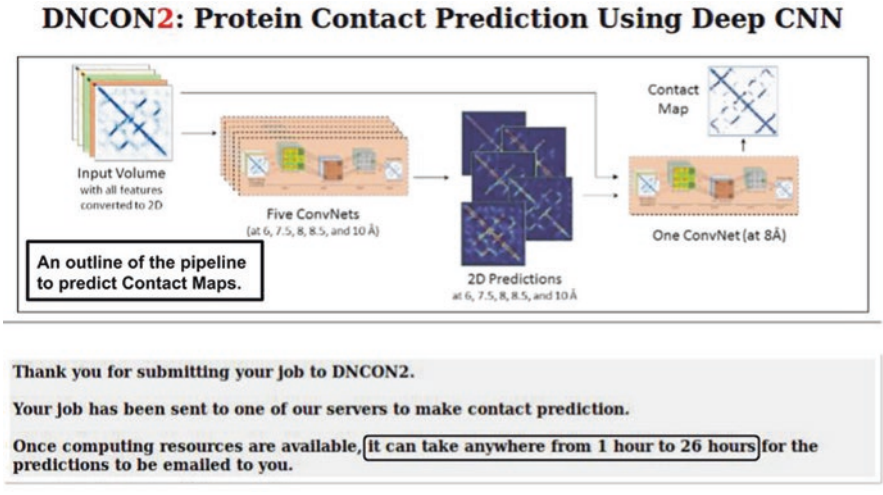

Fig. 10.15 The pipeline of DNCON2 is summarised in the confirmation page

DNCON2 is fully running dealing with the predictions of three proposed sequences. The results of each predictor and package involved are organised in directories.

10.5.2 MetaPSICOV

MetaPSICOV is a CM predictor which has been initially released in 2014 for CASP11 (Kosciolek and Jones 2016) and updated in 2016 for CASP12 (Buchan and Jones 2018). It is recognised as the first CM predictor successfully able to exploit the recent advancements in coevolutionary information extraction (Monastyrskyy et al. 2016). In particular, MetaPSICOV achieved this result implementing three different algorithms to extract coevolution signal from MSA generated with HHblits (Remmert et al. 2012) and HMMER (Finn et al. 2011) – i.e. CCMpred (Seemayer et al. 2014), FreeContact (Kaján et al. 2014) and PSICOV (Jones et al. 2012) – along with other local and global features used for SVMcon (Cheng and Baldi 2007). 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 (Jones et al. 2015). The web server and standalone of MetaPSICOV can be used to predict hydrogen-bonding patterns (Jones et al. 2015).

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 Fig. 10.16, the result page contains links to the output of MetaPSICOV stage 1 (also as image), of stage 2, of

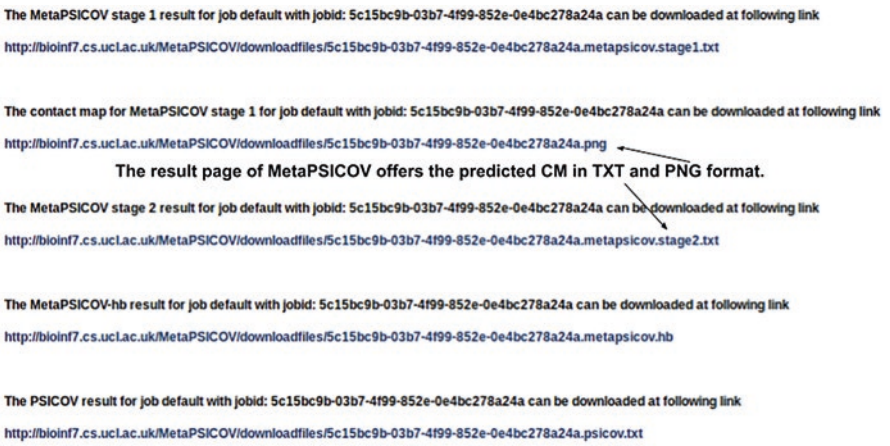

Fig. 10.16 A typical result page of MetaPSICOV. All the files, except the png, follow PSICOV's format

MetaPSICOV-hb (hydrogen bonds) and of PSICOV. A typical CM takes between 20 min and 6 h 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).

10.5.3 RaptorX-Contact

RaptorX-Contact is a 2016 CM predictor which performed well at the last CASP12 (Wang et al. 2017; Schaarschmidt et al. 2018; Wang et al. 2018). RaptorX-Contact aimed to exploit both computer vision (LeCun et al. 2015) and coevolution intuitions to further improve CM prediction. It employs RaptorX-Property (Wang et al. 2016) (see Secondary Structure and Solvent Accessibility) to predict SS and SA, CCMpred (Seemayer et al. 2014) to look for coevolutionary information and in-house algorithms for mutual information and pairwise potential extraction. RaptorX-Contact was trained using MSA generated with PSI-BLAST (Schäffer et al. 2001) while it uses HHblits (Remmert et al. 2012) at prediction time. Thus, the web server and standalone depend on HHblits only.

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 (Fig. 10.17). A JobID is recommended to distinguish among past submissions in the "My Jobs" page, while

Fig. 10.17 The confirmation page of RaptorX-Contact tells the pending jobs ahead and the result URL

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 (Seemayer et al. 2014) 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/.

10.5.4 XX-STOUT

XX STOUT is a CM predictor initially released in 2006 (Vullo et al. 2006) and further improved to be template-based (Walsh et al. 2009) and multi-class in 2009 (Martin et al. 2010). 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-state annotations. When either PSI-BLAST (Schäffer et al. 2001) or the in-house fold recognition software finds homology information, further inputs are provided to XX STOUT to perform template-based predictions – i.e. greatly improve the prediction quality exploiting proteins in the PDB (Berman et al. 2000; Mooney and Pollastri 2009).

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 localisation (Mooney et al. 2013). The predicted CM (threshold 8 Å), 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 are available on request (Fig. 10.18).

Fig. 10.18 XX STOUT sends the predicted protein structure annotations in the body email except the CM (which is attached)

10.6 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

Adhikari B, Hou J, Cheng J (2017) DNCON2: improved protein contact prediction using two-level deep convolutional neural networks. Bioinformatics 34(9):1466–1472

Ahmad S, Gromiha M, Fawareh H, Sarai A (2004) ASAView: database and tool for solvent accessibility representation in proteins. BMC Bioinformatics 5:51

Aloy P, Stark A, Hadley C, Russell RB (2003) Predictions without templates: new folds, secondary structure, and contacts in CASP5. Proteins Struct Funct Bioinforma 53(S6):436–456

Altschul SF et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25(17):3389–3402

Baldi P, Brunak S, Frasconi P, Soda G, Pollastri G (1999) Exploiting the past and the future in protein secondary structure prediction. Bioinforma Oxf Engl 15(11):937–946

Bartoli L, Capriotti E, Fariselli P, Martelli PL, Casadio R (2008) The pros and cons of predicting protein contact maps. Methods Mol Biol Clifton NJ 413:199–217

Baú D, Martin AJ, Mooney C, Vullo A, Walsh I, Pollastri G (2006) Distill: a suite of web servers for the prediction of one-, two- and three-dimensional structural features of proteins. BMC Bioinformatics 7:402

Berman HM et al (2000) The Protein Data Bank. Nucleic Acids Res 28(1):235–242

Buchan DWA, Jones DT (2018) Improved protein contact predictions with the MetaPSICOV2 server in CASP12. *Proteins* 86(Suppl 1):78–83

Buchan DWA, Ward SM, Lobley AE, Nugent TCO, Bryson K, Jones DT (2010) Protein annotation and modelling servers at University College London. *Nucleic Acids Res* 38(suppl_2):W563–W568

Bystroff C, Thorsson V, Baker D (2000) HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins. J Mol Biol 301(1):173–190

Cheng J, Baldi P (2007) Improved residue contact prediction using support vector machines and a large feature set. BMC Bioinformatics 8:113

Cheng J, Randall AZ, Sweredoski MJ, Baldi P (2005) SCRATCH: a protein structure and structural feature prediction server. *Nucleic Acids Res* 33(suppl_2):W72–W76

Chou PY, Fasman GD (1974) Prediction of protein conformation. Biochemistry (Mosc) 13(2):222–245

Cornette JL, Cease KB, Margalit H, Spouge JL, Berzofsky JA, DeLisi C (1987) Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. J Mol Biol 195(3):659–685

Cuff JA, Clamp ME, Siddiqui AS, Finlay M, Barton GJ (1998) JPred: a consensus secondary structure prediction server. Bioinformatics 14(10):892–893

De Brevern AG, Etchebest C, Hazout S (2000) Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. Proteins Struct Funct Bioinforma 41(3):271–287

Di Lena P, Fariselli P, Margara L, Vassura M, Casadio R (2010) Fast overlapping of protein contact maps by alignment of eigenvectors. Bioinformatics 26(18):2250–2258

Di Lena P, Fariselli P, Margara L, Vassura M, Casadio R (2011) Is there an optimal substitution matrix for contact prediction with correlated mutations? IEEEACM Trans Comput Biol Bioinforma 8(4):1017–1028

Di Lena P, Nagata K, Baldi P (2012) Deep architectures for protein contact map prediction. Bioinformatics 28(19):2449–2457

Drozdetskiy A, Cole C, Procter J, Barton GJ (2015) JPred4: a protein secondary structure prediction server. Nucleic Acids Res 43(W1):W389–W394

Eickholt J, Cheng J (2012) Predicting protein residue–residue contacts using deep networks and boosting. Bioinformatics 28(23):3066–3072

Faraggi E, Yang Y, Zhang S, Zhou Y (2009) Predicting continuous local structure and the effect of its substitution for secondary structure in fragment-free protein structure prediction. Structure 17(11):1515–1527

Fariselli P, Olmea O, Valencia A, Casadio R (2001) Prediction of contact maps with neural networks and correlated mutations. Protein Eng Des Sel 14(11):835–843

Fauchère JL, Charton M, Kier LB, Verloop A, Pliska V (1988) Amino acid side chain parameters for correlation studies in biology and pharmacology. Int J Pept Protein Res 32(4):269–278

Finn RD, Clements J, Eddy SR (2011) HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* 39(Web Server issue):W29–W37

Göbel U, Sander C, Schneider R, Valencia A (1994) Correlated mutations and residue contacts in proteins. Proteins 18(4):309–317

Haas J et al Continuous Automated Model Evaluation (CAMEO) complementing the critical assessment of structure prediction in CASP12. Proteins: Struct Funct Bioinf p. n/a-n/a

Heffernan R et al (2015) Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. *Sci Rep* 5:11476

Heffernan R et al (2016) Highly accurate sequence-based prediction of half-sphere exposures of amino acid residues in proteins. Bioinformatics 32(6):843–849

Heffernan R, Yang Y, Paliwal K, Zhou Y (2017) 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 33(18):2842–2849

Holbrook SR, Muskal SM, Kim SH (1990) Predicting surface exposure of amino acids from protein sequence. Protein Eng 3(8):659–665

Huang Y, Bystroff C (2006) Improved pairwise alignments of proteins in the Twilight Zone using local structure predictions. Bioinformatics 22(4):413–422

Johnson LS, Eddy SR, Portugaly E (2010) Hidden Markov model speed heuristic and iterative HMM search procedure. BMC Bioinformatics 11:431
Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 292(2):195–202
Jones DT, Swindells MB (2002) Getting the most from PSI–BLAST. Trends Biochem Sci 27(3):161–164
Jones DT, Buchan DWA, Cozzetto D, Pontil M (2012) PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics 28(2):184–190
Jones DT, Singh T, Kosciolek T, Tetchner S (2015) MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. Bioinformatics 31(7):999–1006
Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22(12):2577–2637
Kaján L, Hopf TA, Kalaš M, Marks DS, Rost B (2014) FreeContact: fast and free software for protein contact prediction from residue co-evolution. BMC Bioinformatics 15:85
Kendrew JC et al (1960) Structure of myoglobin: a three-dimensional Fourier synthesis at 2 A. resolution. Nature 185(4711):422–427
Kim DE, DiMaio F, Wang RY-R, Song Y, Baker D (2014) One contact for every twelve residues allows robust and accurate topology-level protein structure modeling. *Proteins* 82(2):208–218
Kinch LN, Li W, Monastyrskyy B, Kryshtafovych A, Grishin NV (2016) Assessment of CASP11 contact-assisted predictions. Proteins 84(Suppl 1):164–180
Kosciolek T, Jones DT (2014) De Novo structure prediction of globular proteins aided by sequence variation-derived contacts. PLoS One 9(3):e92197
Kosciolek T, Jones DT (2016) Accurate contact predictions using covariation techniques and machine learning. *Proteins* 84(Suppl 1):145–151
Kuang R, Leslie CS, Yang A-S (2004) Protein backbone angle prediction with machine learning approaches. Bioinformatics 20(10):1612–1621
Kukic P, Mirabello C, Tradigo G, Walsh I, Veltri P, Pollastri G (2014) Toward an accurate prediction of inter-residue distances in proteins using 2D recursive neural networks. BMC Bioinformatics 15:6
LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444
Lyons J et al (2014) Predicting backbone Cα angles and dihedrals from protein sequences by stacked sparse auto-encoder deep neural network. J Comput Chem 35(28):2040–2046
MacCallum RM (2004) Striped sheets and protein contact prediction. *Bioinformatics* 20(suppl_1):i224–i231
Magnan CN, Baldi P (2014) SSpro/ACCpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity. Bioinformatics 30(18):2592–2597
Martin J, Letellier G, Marin A, Taly J-F, de Brevern AG, Gibrat J-F (2005) Protein secondary structure assignment revisited: a detailed analysis of different assignment methods. BMC Struct Biol 5:17
Martin AJ, Mooney C, Walsh I, Pollastri G (2010) Contact map prediction by machine learning. In: Pan Y, Zomaya A, Rangwala H, Karypis G (eds) Introduction to protein structure prediction. Wiley. https://doi.org/10.1002/9780470882207.ch7
Mirabello C, Pollastri G (2013) Porter, PaleAle 4.0: high-accuracy prediction of protein secondary structure and relative solvent accessibility. Bioinformatics 29(16):2056–2058
Monastyrskyy B, D'Andrea D, Fidelis K, Tramontano A, Kryshtafovych A (2014) Evaluation of residue–residue contact prediction in CASP10. Proteins Struct Funct Bioinforma 82:138–153
Monastyrskyy B, D'Andrea D, Fidelis K, Tramontano A, Kryshtafovych A (2016) New encouraging developments in contact prediction: Assessment of the CASP11 results. *Proteins* 84(Suppl 1):131–144
Mooney C, Pollastri G (2009) Beyond the twilight zone: automated prediction of structural properties of proteins by recursive neural networks and remote homology information. Proteins Struct Funct Bioinforma 77(1):181–190

Mooney C, Vullo A, Pollastri G (2006) Protein structural motif prediction in multidimensional ø-ψ space leads to improved secondary structure prediction. J Comput Biol 13(8):1489–1502
Mooney C, Cessieux A, Shields DC, Pollastri G (2013) SCL-Epred: a generalised de novo eukaryotic protein subcellular localisation predictor. Amino Acids 45(2):291–299
Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247(4):536–540
Olmea O, Valencia A (1997) Improving contact predictions by the combination of correlated mutations and other sources of sequence information. Fold Des 2:S25–S32
Pascarella S, Persio RD, Bossa F, Argos P (1998) Easy method to predict solvent accessibility from multiple protein sequence alignments. Proteins Struct Funct Bioinforma 32(2):190–199
Pauling L, Corey RB (1951) Configurations of polypeptide chains with favored orientations around single bonds. Proc Natl Acad Sci U S A 37(11):729–740
Pazos F, Helmer-Citterich M, Ausiello G, Valencia A (1997) Correlated mutations contain information about protein-protein interaction. J Mol Biol 271(4):511–523
Perutz MF, Rossmann MG, Cullis AF, Muirhead H, Will G, North AC (1960) Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5-A. resolution, obtained by X-ray analysis. Nature 185(4711):416–422
Pollastri G, Baldi P (2002) Prediction of contact maps by GIOHMMs and recurrent neural networks using lateral propagation from all four cardinal corners. *Bioinformatics* 18(suppl_1):S62–S70
Pollastri G, McLysaght A (2005) Porter: a new, accurate server for protein secondary structure prediction. Bioinformatics 21(8):1719–1720
Pollastri G, Baldi P, Fariselli P, Casadio R (2001) Improved prediction of the number of residue contacts in proteins by recurrent neural networks. *Bioinformatics* 17(suppl_1):S234–S242
Pollastri G, Baldi P, Fariselli P, Casadio R (2002) Prediction of coordination number and relative solvent accessibility in proteins. Proteins 47(2):142–153
Pollastri G, Martin AJ, Mooney C, Vullo A (2007) Accurate prediction of protein secondary structure and solvent accessibility by consensus combiners of sequence and structure information. BMC Bioinformatics 8:201
Remmert M, Biegert A, Hauser A, Söding J (2012) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods 9(2):173–175
Rost B (2001) Review: protein secondary structure prediction continues to rise. J Struct Biol 134(2):204–218
Rost B, Sander C (1993) Prediction of protein secondary structure at better than 70% accuracy. J Mol Biol 232(2):584–599
Rost B, Sander C (1994) Conservation and prediction of solvent accessibility in protein families. Proteins 20(3):216–226
Roy A, Kucukural A, Zhang Y (2010) I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc 5(4):725–738
Schaarschmidt J, Monastyrskyy B, Kryshtafovych A, Bonvin AMJJ (2018) Assessment of contact predictions in CASP12: co-evolution and deep learning coming of age. Proteins Struct Funct Bioinforma 86:51–66
Schäffer AA et al (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res 29(14):2994–3005
Schlessinger A, Punta M, Rost B (2007) Natively unstructured regions in proteins identified from contact predictions. Bioinforma Oxf Engl 23(18):2376–2384
Schmidhuber J (2015) Deep learning in neural networks: an overview. Neural Netw 61:85–117
Seemayer S, Gruber M, Söding J (2014) CCMpred—fast and precise prediction of protein residue–residue contacts from correlated mutations. Bioinformatics 30(21):3128–3130
Sims GE, Choi I-G, Kim S-H (2005) Protein conformational space in higher order ϕ-Ψ maps. Proc Natl Acad Sci U S A 102(3):618–621
Tegge AN, Wang Z, Eickholt J, Cheng J (2009) NNcon: improved protein contact map prediction using 2D-recursive neural networks. *Nucleic Acids Res* 37(suppl_2):W515–W518
The UniProt Consortium (2016) UniProt: the universal protein knowledgebase. Nucleic Acids Res 45(D1):D158–D169

Thomas H (2005) An amino acid has two sides: a new 2D measure provides a different view of solvent exposure. Proteins Struct Funct Bioinforma 59(1):38–48

Ting D, Wang G, Shapovalov M, Mitra R, Jordan MI, Jr RLD (2010) Neighbor-dependent Ramachandran probability distributions of amino acids developed from a Hierarchical Dirichlet process model. *PLoS Comput Biol* 6(4):e1000763

Torrisi M, Kaleel M, Pollastri G (2018) Porter 5: state-of-the-art ab initio prediction of protein secondary structure in 3 and 8 classes. *bioRxiv*:289033

Vassura M, Margara L, Di Lena P, Medri F, Fariselli P, Casadio R (2008) Reconstruction of 3D structures from protein contact maps. IEEEACM Trans Comput Biol Bioinforma 5(3):357–367

Vassura M et al (2011) Blurring contact maps of thousands of proteins: what we can learn by reconstructing 3D structure. BioData Min 4:1

Vendruscolo M, Kussell E, Domany E (1997) Recovery of protein structure from contact maps. Fold Des 2(5):295–306

Vullo A, Walsh I, Pollastri G (2006) A two-stage approach for improved prediction of residue contact maps. BMC Bioinformatics 7:180

Walsh I, Baù D, Martin AJ, Mooney C, Vullo A, Pollastri G (2009) Ab initio and template-based prediction of multi-class distance maps by two-dimensional recursive neural networks. BMC Struct Biol 9:5

Walsh I, Pollastri G, Tosatto SCE (2016) Correct machine learning on protein sequences: a peer-reviewing perspective. Brief Bioinform 17(5):831–840

Wang S, Li W, Liu S, Xu J (2016) RaptorX-property: a web server for protein structure property prediction. Nucleic Acids Res 44(W1):W430–W435

Wang S, Sun S, Li Z, Zhang R, Xu J (2017) Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS Comput Biol* 13(1):e1005324

Wang S, Sun S, Xu J (2018) Analysis of deep learning methods for blind protein contact prediction in CASP12. *Proteins* 86(Suppl 1):67–77

Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ (2009) Jalview version 2—a multiple sequence alignment editor and analysis workbench. Bioinformatics 25(9):1189–1191

Wood MJ, Hirst JD (2005) Protein secondary structure prediction with dihedral angles. Proteins Struct Funct Bioinforma 59(3):476–481

Xia L, Pan X-M (2000) New method for accurate prediction of solvent accessibility from protein sequence. Proteins Struct Funct Bioinforma 42(1):1–5

Yang Y, Faraggi E, Zhao H, Zhou Y (2011) 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 27(15):2076–2082

Yang Y et al (2016) Sixty-five years of the long march in protein secondary structure prediction: the final stretch? Brief Bioinform 19(3):482–494

Yuan Z (2005) Better prediction of protein contact number using a support vector regression analysis of amino acid sequence. BMC Bioinformatics 6:248

Zemla A (2003) LGA: a method for finding 3D similarities in protein structures. Nucleic Acids Res 31(13):3370–3374

Zemla A, Venclovas Č, Fidelis K, Rost B (1999) A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. Proteins Struct Funct Bioinforma 34(2):220–223