Exploring Protein Structure Prediction using PSIPRED and Phyre2

Link to Video Tutorial: https://youtu.be/0a5Crq4SOmc

Background

Protein structure prediction is a very important field of protein research. Protein structure is often related to protein function. One common example of this is enzymes: each enzyme has an active site with a specific shape that allows only certain molecules to bind.

As sequencing technologies have improved, the amount of known protein sequences has increased dramatically. However, the amount of known protein structures remains relatively low, due to the time and cost of experimental techniques such as NMR and X-Ray crystallography. As a result, the UniProt/TrEMBL database contains over 85 million protein sequences, while the Protein Data Bank (PDB) contains only about 120,000 structures (as of April 2017[1]). As the vast majority of proteins do not have known structures, it is important to have accurate computational methods that can provide insight about protein structure. There has been a lot of work focusing on the prediction of a protein's secondary (local, two-dimensional) and tertiary (overall, three-dimensional) structure from its primary structure, or amino acid sequence. Quaternary structure, referring to the structure arising from multiple proteins/protein subunits interacting, is a more complicated challenge.

One common approach for the determination of tertiary structure is homology modeling. Homology modeling uses known structures and sequences to predict how different proteins with similar sequences may fold. This can provide relatively accurate predictions but falls short in a few major ways. Although part of the sequence will align to a known sequence, some of it will not, and for that region, other computational methods must be used to determine structure. In addition, homology modeling relies on fitting a protein to already discovered structures. Some three-dimensional structures have not yet been discovered and are therefore not represented by this modeling.

Secondary structure prediction is also a very important aspect of protein structure prediction as it can be helpful in answering some research questions on its own and can also be seen as a steppingstone to higher order structure prediction. Secondary structure plays an important role in determining how proteins will fold[2] and is therefore also important to consider when determining the overall structure. In addition, secondary structure has been proven to be useful in function determination[3]. Secondary structure prediction however, is not a foolproof method either, with its highest possible accuracy placed at 88-90%[4]. Overall, both techniques can provide valuable information about the structure and therefore function of different proteins, and how they fold, interact and interact with their environment.

This manual will explore the prediction of both of these types of structure using the prediction tools PSIPRED and Phyre2. How to use these tools will be described, and the advantages and disadvantages of each will be discussed. In addition, some sample sequences will be run, with the outputs discussed more in the accompanying video.

Bioinformatic Tools

The protein structure prediction tools that I will be using are PSIPRED and Phyre2[5]. PSIPRED is one of the highest performing secondary structure predictors[6], while Phyre2 focuses on homology modeling.

PSIPRED: http://bioinf.cs.ucl.ac.uk/psipred/

This manual will provide information on how to use PSIPRED workbench, which includes the PSIPRED secondary structure prediction and other associated tools including domain prediction and disorder prediction. PSIPRED uses PSI-BLAST, along with a two-layer neural network, to predict secondary structure. The details of this are beyond the scope of this manual, but more information can be found in PSIPRED publications[6-8].

Phyre2: http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index

Phyre2 is a homology modelling service that searches the millions of known sequences for homologues using protein-specific iterative BLAST (PSI-BLAST), which allows the user to find similar protein using position-specific scoring matrices. The homologous sequences are gathered and turned into a statistical model called the Hidden Markov model, which takes into account the patterns of mutations that occur in this sequence over time. The same process has already been performed for proteins of known structure (from the PDB for example), generating a database of Hidden Markov models of known structures. The Hidden Markov model of the target sequence is then aligned to the

Hidden Markov models in the database, and from that a 3D model of the target sequence is generated. This approach has shown to be accurate, even with sequence identity below 15%[5].

These two tools were chosen as they are integrated with each other (Phyre2 uses PSIPRED in its secondary structure determination) and allow for a demonstration of the different information that can be gained from such tools. In addition, both of these services are a collection of connected tools and services that allow you to gain a lot of information from one place, simplifying the process.

Dataset

In order to investigate these tools, the protein titin will be used. Titin is one of the three main filaments present in striated muscle, along with actin and myosin. This protein spans half of a sarcomere and connects the thick filament to the thin filament. It is thought to act as a molecular spring and play an integral role in passive tension and muscle contraction[9]. Titin is used as a dataset for this study as it contains many different regions and domains that can be modelled. In addition, the structure of some, but not all, of the domains in titin have been studied, allowing for the homology models to find these similar structures. In particular, some isoforms of titin contain an N2A region that is thought to be important in signaling/regulation[10]. This N2A region contains several immunoglobulin domains and a unique insertion sequence, which is partially disordered. These different domains/structural units of titin allow for an analysis of these tools. The sequence used in the video tutorial is the titin N2A isoform with NCBI Reference Sequence: NP_035782.3.

Manual

Phyre2

How to input a sequence

You can either input a sequence directly (in FASTA format or just the code) or use the sequence finder. As the sequence finder is in beta-testing, it would likely be easier to find your sequence through a database and manually input the sequence. If you input your sequence in FASTA format, the header will be ignored, so you should input your name/identifier in the “optional job description” field. You will also have to select a modeling mode (normal, intensive or test) for which you should select normal for most analyses, and a use (not for profit or commercial). As the jobs can take a while to run, you will need to provide an email address for the link with the results to be sent to.

Results

Once the job is complete, a link to the results page will be sent to your email, along with a .pdb file of the best model of the 3-dimensonal structure. Pdb files contain the structure of a protein, which can be visualized using a variety of methods. This manual will describe how to view the pdb file on the Phyre2 webpage, but Phyre2 also provides a help page which lists different software for visualizing pdb files: http://www.sbg.bio.ic.ac.uk/phyre2/html/help.cgi?id=help/faq.

The results page consists of several different sections which will be described below.

Top Model Information: This section provides information on the model that is the best match. It contains a rainbow 3-dimensional model of your protein with red as the N-terminus and blue as the C terminus. By clicking "Interactive 3D view in Jsmol", you can rotate your model in 3 dimensions. This section also contains the PDB entry of the template used for this model, the confidence value and coverage. The confidence value indicates how confident the system is that the template is homologous to your protein sequence, not how accurate the model is. Coverage indicates how much of the sequence you submitted was matched to the homology model. This section is where the website will give a warning if a lot of your protein appears to be disordered, because this technique is inaccurate for disordered proteins, as they adopt an ensemble of conformations. This section may also provide a link to Phyre alarm if the coverage is low, which is a service that will notify you if there is a new sequence uploaded that is a homolog to your sequence.

Sequence analysis: This section shows the output of the PSI-Blast search done to search for protein homologs.

Secondary structure/disorder: This section shows you your input sequence, what the likely secondary structure is at this point, disorder prediction and its confidence. They use PSIPPRED and DISOPRED for these predictions. See the section on PSIPRED and DISOPRED for further details.

Domain Analysis: This section provides a visual showing where the template aligned to the sequence of interest and with what confidence the match was, with red being high confidence, orange of slightly less confidence, yellow of even less confidence and so forth. The identity of the sequence is shown using an identifier where the first letter "c" indicates that the whole chain was taken from the PDB, while “d” indicates only a domain. Matches at the top of the list contain links to the detailed template information below. Matches further down the list were detected but not modelled.

Detailed template information: This section provides more detailed information about the different templates that could be used to model your protein sequence. It provides a confidence and percent identity value, as well as the title and link to the PDB entry showing this template (in the US, Europe and Japan PDB). To download a PDB file of the structure, simply click the picture of the structure. This section also has a link to run the Phyre investigator which gives more information about mutations and model quality. However, as this feature is still in beta-testing, it will not be discussed extensively. Another useful feature is the ability to superimpose several models over each other by selecting them on the left-hand side. This allows you to identify the differences between each.

Phyre2 Advanced Features

If you create a free account, you can use expert mode which gives you access to more options. Features includes one-to-one threading, batch processing, BackPhyre, and Genesearch. One-to-one threading is useful if you have a sequence and a specific known structure template (in PDB format) which you would like that sequence modeled to. The result will show the alignment of your sequence and template and the structure of the model generated. Batch processing allows you to run more than one sequence at a time. You can run up to 100 sequences (and even more if needed, provided you contact the Phyre administrators) by simply submitting a FASTA file containing the sequences. Backphyre performs a search in reverse. In other words, it produces a list of matches from different genomes (that can be selected) based on a 3-D structure (in the form of a PDB file). Genesearch performs a similar function as Backphyre, except the input is a protein sequence, and not a PDB file.

Exporting data

Most sections contain a PDF symbol that exports that section into a pdf file that you can save. The sequence analysis section contains a link to download the FASTA version of the results from that section. In addition, there is an option to download a zipped file of all results on top of the page.

Viewing results later

If you would like to refer to the webpage often for a project, this is also an option. You can find your job again through the link sent by email, or by saving the unique job ID (provided in the upper righthand corner). Your job will be saved for 30 days, and you can even extend this further by pressing the "renew for 30 days" button, also located in the top right of the webpage.

PSIPRED

How to input a sequence

There are two main input options: protein sequence or PDB structure file (in old format). For the sake of comparison, this manual will focus on the sequence input method.

If you chose protein sequence, then it is recommended that you enter the plain protein sequence (string of amino acids using their 1 letter codes). You may also enter the sequence in a FASTA format, but the header will be ignored. The maximum length sequence that you can submit is 1500 residues, with some of the functions having a maximum length of 500 residues. This means that the analysis is limited to small proteins/subsections of proteins, or requires multiple segments of a longer protein to be submitted separately.

PSIPRED offers several different kinds of servers/functions that you can select. Which you select will be dependent on what kind of protein you are analyzing and what information you hope to gain. Not all analyses will be useful for all proteins (in fact, it prevents you from selecting all the analyses). However, it is almost always useful to run the PSIPRED secondary structure prediction function no matter what other analyses you run from the overall PSIPRED workbench, as it provides an overview of the secondary structure which is likely useful. The following section will go over the different servers/functions and the results that can be obtained from each. From this information you can decide which are useful for your project.

Results

Every results page will begin with the name of the job entered and a link to the results page that can be copied. This is followed by a "Sequence plot" which displays the amino acid sequence. Depending on the analyses run different color-coded sequences are generated (psipred, memesat and aatypes) which you can toggle between. Psipred color codes based on secondary structure, memesat based on membrane association and aatypes based on the types of amino acids. Aatypes analysis will always be run, but in order for psipred to be displayed, PSIPRED 4.0 must be run, and for memsat to be displayed, MEMSAT, MESATSVM or MEMPACK must be run. You can save an image of the sequence plot as a PNG or SVG file by clicking the "Get PNG" or "Get SVG" buttons on the bottom right of the sequence plot. It is important to note that it will save the plot that is currently displayed.

Different functions

PSIPRED 4.0 (Predict Secondary Structure):

PSIPRED 4.0 provides an easy to visualize prediction of the secondary structure at each residue. It is a well-maintained, popular server that uses a neural network to make its predictions. It outputs a cartoon containing the input sequence (AA), the predicted structure (Pred), a cartoon of the predicted structure (Cart) and a confidence of that prediction (Conf). The predicted structure is depicted as a letter, with H, C and E referring to helix, coil, and strand respectively. The confidence of prediction is depicted visually with a higher, darker bar indicating greater confidence. The runtime of PSIPRED is short, from seconds to a few minutes.

DISOPRED3 (Disopred Prediction)

DISOPRED3 is a server designed to detect intrinsically disordered regions or IDRs. This neural network predictive tool was trained to identify intrinsically disordered regions using sequences from the PDB (Protein Data Bank) and DisProt (Disordered Protein Database). The output is a graph of the confidence level versus the residue number. The dashed lines indicate the minimum confidence value cutoff to determine that an area is disordered. The amino acids marked as disordered on the sequence plot at the top of the results page is determined by this cutoff value. More information on DISOPRED can be found here: http://bioinfadmin.cs.ucl.ac.uk/UCL-CS_Bioinformatics_DISOPRED_Overview.html.

pGenTHREADER (Profile Based Fold Recognition), GenTHREADER (Rapid Fold Recognition) and pDomTHREADER (Protein Domain Fold Recognition)

All of these are prediction tools that recognize specific folds (pGenTHREADER and GenTHREADER) or domains (pDomTHREADER). GenTHREADER is the original tool, while pGenTHREADER is a version that uses predicted secondary structure in addition to alignment to generate its predictions. This version (pGenTHREADER) is recommended but requires more time to run due to the additional PSI-BLAST required. DomTHREADER is similar except it recognizes domains. The result of both is a table containing information about the hit (fold or domain) and the confidence of this match. Some important columns include the p-value, links to search the fold/domain in different databases and a graphical alignment. An explanation of all the outputs of the table can be found here: http://bioinfadmin.cs.ucl.ac.uk/UCL-CS_Bioinformatics_Server_Tutorial.html#genthreader_method

DomPred (Domain Prediction):

This function determines the number of domains present in your sequence and the location of any domain boundaries. PSI-BLAST is used to align the sequence with a database of known domains (Pfam), and this along with a prediction of N and C terminal boundaries is used to generate the results. The output shows a graph of the alignment where a peak indicates a location where there is a predicted domain boundary. Below the plot, it will also indicate the number of domains predicted to be in the sequence, and the amino acid residue at which the boundaries occur.

More information on DomPred can be found here: http://bioinfadmin.cs.ucl.ac.uk/UCL-CS_Bioinformatics_Dompred_Overview.html.

FFPred (Function Prediction)

This function attempts to predict the gene ontology (GO) terms of proteins using a series of Support Vector Machines (SVMs). It is meant for eukaryotic proteins whose GO terms are otherwise difficult to predict. You can run the system optimized for fly or human genes (there is a choice on the sidebar upon submission of your sequence). The output displays the GO term predictions, with separate tables for each gene ontology domain. In other words, it contains tables for biological process, molecular function and cellular component predictions. Each table contains the predicted GO term, the name, the probability of the protein being annotated with the GO term (out of 1), and the reliability of the SVM used. Predictions shown with a red background near the bottom of the table are generated with less reliable SVMs (these are also indicated by an L for low in the SVM reliability column). More information on how FFPRED makes these predictions can be found here: http://bioinfadmin.cs.ucl.ac.uk/UCL-CS_Bioinformatics_FFPred_help.html

Structure Modelling: Bioserf, Domserf and DMPfold

Bioserf and Domserf are homology modelling services that are run using Modeller, and therefore require a Modeller key which you can request from the Sali lab at the following website:

https://salilab.org/modeller/registration.html. DMPFold can be run without a Modeller key but takes very long to run.

The PSIPRED workbench also has several tools meant for transmembrane proteins (MEMSAT3, MEMSATSVM, MEMPACK). As the dataset of interest is not membrane associated, the tools relating to membranes will not be discussed here.

Exporting data

On the side of the webpage, there is a downloads section where you can download a zipped file of all your results, or some specific results (listed by tool). However, it is important to note that results older than 5 days are deleted from the server, so data should be downloaded or resubmitted if required for a longer period of time.

Conclusion

The PSIPRED and Phyre2 servers provide a group of tools for the study of proteins and their structure that can yield different results. As an example, analyses of an intrinsically disordered titin exon and a section of the N2A region (I80-IS) were performed on both servers. The intrinsically disordered exon showed no significant results in Phyre2, along with a warning that this region was predicted to be disordered. This is expected as Phyre2 is not meant to be used with disordered sequences. Analysis with the PSIPRED workbench showed similarly few results, with no domain boundary or fold recognition results. However, only a small portion of the sequence was predicted to be disordered using the DISOPRED plot, compared to the 93% seen in Phyre2. The I80-IS section of the N2A region showed several high confidence high coverage results, presumably due to the large number of similar Ig domains that have been characterized. Sequence identity, however, was low for many of these models, with values ranging from about 15-40%. PSIPRED showed accurate predictions for the I80-IS in several of its tools. For example, DomPred correctly identified the number of domains and the relative location of the domain boundary. However, both tools failed to predict the helical core found in the insertion sequence (discovered in Gage Lab), which indicates that although these methods may be relatively reliable, validations are needed to confirm their accuracy.

It is also important to note that both PSIPRED and Phyre2 are primarily designed for ordered proteins; sequences containing sections that are intrinsically disordered should be used with caution. The DISOPRED function in PSIPRED and the secondary structure and disorder prediction section in Phyre2 are designed for the prediction of disordered proteins, with the PSIPRED appearing to have a much higher threshold for what it considers as disordered.

These tools have both been updated recently (Phyre2 is the second version of Phyre, PSIPRED is on its 4th version), which shows a commitment to improving, updating and maintaining these tools. Both are relatively easy to use and require no coding knowledge or experience, making them very accessible. Overall, both provide interesting information and predictions about your sequence of interest and could be very useful in protein research.

References:

1. Deng, H.; Jia, Y.; Zhang, Y., Protein structure prediction. *Int J Mod Phys B* **2018,** *32* (18), 1840009.
2. Ozkan, S. B.; Wu, G. A.; Chodera, J. D.; Dill, K. A., Protein folding by zipping and assembly. *Proc Natl Acad Sci U S A* **2007,** *104* (29), 11987-92.
3. Godzik, A.; Jambon, M.; Friedberg, I., Computational protein function prediction: are we making progress? *Cell Mol Life Sci* **2007,** *64* (19-20), 2505-11.
4. Rost, B., Review: protein secondary structure prediction continues to rise. *J Struct Biol* **2001,** *134* (2-3), 204-18.
5. Kelley, L. A.; Mezulis, S.; Yates, C. M.; Wass, M. N.; Sternberg, M. J. E., The Phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocols* **2015,** *10* (6), 845-858.
6. Buchan, D. W. A.; Jones, D. T., The PSIPRED Protein Analysis Workbench: 20 years on. *Nucleic Acids Res* **2019,** *47* (W1), W402-W407.
7. Jones, D. T., Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* **1999,** *292* (2), 195-202.
8. McGuffin, L. J.; Bryson, K.; Jones, D. T., The PSIPRED protein structure prediction server. *Bioinformatics* **2000,** *16* (4), 404-405.
9. Helmes, M.; Trombitás, K.; Centner, T.; Kellermayer, M.; Labeit, S.; Linke, W. A.; Granzier, H., Mechanically driven contour-length adjustment in rat cardiac titin's unique N2B sequence: titin is an adjustable spring. *Circ Res* **1999,** *84* (11), 1339-52.
10. Zhou, T.; Fleming, J. R.; Lange, S.; Hessel, A. L.; Bogomolovas, J.; Stronczek, C.; Grundei, D.; Ghassemian, M.; Biju, A.; Börgeson, E.; Bullard, B.; Linke, W. A.; Chen, J.; Kovermann, M.; Mayans, O., Molecular Characterisation of Titin N2A and Its Binding of CARP Reveals a Titin/Actin Cross-linking Mechanism. *Journal of Molecular Biology* **2021,** *433* (9), 166901.