InterPro

http://www.ebi.ac.uk/interpro

InterPro is a natural extension to Pfam and other databases for protein classification. InterPro also predicts functional characteristics of proteins by assigning them to families and by recognizing protein domains and relevant sites. However, InterPro is a meta-database: it combines protein signatures from multiple, independent databases into a single resource for an integrative sequence classification.

In InterPro, a 'protein signature' is just a computational model that reflect the site-specific amino acid conservation pattern of a multiple sequence alignment. It can take the form of a sequence pattern, a profile that describes a certain sequence motif or a sophisticated Hidden Markov Model that contemplates insertions and deletions in protein families. Initial models are used for iterative searches of databases such as UniProtKB to collect remote homologues and increase the number of distant sequences that the model represents. The final protein signature is a very descriptive prediction model that can be used for protein sequence analysis.

Among the many 'member databases' of InterPro are Pfam, SMART (a database of protein domain architectures), Superfamily (an HMM-based database of functional and structural annotation in proteins), CATH/Gene3D (a database of domain superfamilies with known structure and their prediction in complete genomes) and MobiDB (a central resource for annotation of protein intrinsic disorder in UniProt sequences)

Since the different databases would very likely have redundant information, a team of InterPro curators check and manually merge the signatures referring to the same family, domain or site into single InterPro entries. Therefore, each protein signature has as an InterPro accession code plus the corresponding code in the individual databases.

Notice that the strength of InterPro lies on the integration of multiple sources of information, each with its typical strengths, and not in the quantity of information extracted from them. By collecting multiples sources of annotation, InterPro serves as a powerful diagnostic tool. However, when signatures of interest are found, it is generally advisable to follow the links to the corresponding member databases for further information.

An InterPro entry can be of a few different types. 'Homologous Superfamily' entries collect proteins with common ancestry. They normally have very low sequence conservation but a noticeable structural similarity. Members of a 'Family' of proteins, while also sharing a common evolutionary origin, have similar sequences, structures and/or functions. 'Domains' in InterPro are discrete units with distinct sequence, structure or function that can be find in different biological contexts. 'Repeats' identify short sequences typically found many times in the same protein. A 'Site' is also a short sequence but with one or more conserved residues that may have a defined function (active site, binding site or post-translational modification sites.)

Some useful things to know about InterPro:

- At the homepage, a box on the top right corner allows searching InterPro.with distinct entry codes related with your protein of interest. You can use an InterPro accession code (consisting of 'IPR' plus a number), an identifier from UniProtKB or any of the member databases, etc. You can also search with keywords related with the function or activity of your protein (e.g. 'isomerase').
- Searches of the InterPro database with a sequence of interest can be made with InterProScan. It is automatically loaded when you click on the 'Search' tab but it is also available for download in the 'Download' tab.
- Also, you can download all InterPro entries and all mappings of InterPro entries to UniProtKB proteins, among other data sets, from the 'Download' tab.

Exercises

1. Do you have a protein of interest? If so, try to find which InterPro signatures it matches.

2. Search InterPro with the Pfam identifier **PF00571**. What is its InterPro accession code? Does it describe a protein, a domain, a family or something else? Does it describe the same entity that Pfam does? Do they look similar or at least have things in common?

3. Looking at the same InterPro entry, can you tell which are the signatures from other databases that were merged with Pfam to create this InterPro entry? (If so, follow the links and check if these member databases provide new information on the protein.)

4. How many proteins are matched by this entry? How many domain architectures? How many proteins have two repeated occurrences of this entry followed by a transporter-associated domain (IPR005170)?

5. Search for the InterPro entry describing the protein **Lysine-specific demethylase 5C** (UniProtKB entry P41229). The left-hand menu lets you filter out unwanted signatures, while the main section holds the prediction of signatures in the protein of interest.

   a. Which signatures are recognized in this protein? What changes if you de-select the box 'Domains', on the 'Entry type' section at the left-hand menu?

   b. Protein signatures are mapped to the query sequence in the main results. Look at the 'Homologous superfamilies' section. What is the architecture of this superfamily?

   c. The most C-terminal signature found in this superfamily is a certain zinc finger. Which InterPro signatures describe this region? What type of signatures are these? Where does InterPro takes the information for this assignment? (*Hint*: inspect the 'Detailed signature matches' section.)

   d. Among the matched signature is the IPR0001606 'domain' entry. Which databases were processed to create this entry? Do they all map to the exact same region of the protein of interest? If not, why?

   e. Can you identify any intrinsically disordered region in this protein? (Remember that 'mobidb' is a top resource for annotation of these regions.)

   f. Follow the link on the top-left to Download the protein sequences of all 221 similar proteins in FASTA format. Open the file on your system to check consistency.

6. Search for the InterPro entry describing the protein with the FASTA sequence copied at the bottom of the page.

   a. Does it belong to any protein family? Can you find the name of the PRINTS database entry used for building this InterPro family signature?

   b. What is the difference between InterPro signatures IPR001767 and IPR003587?

   c. Looking at the previous two signatures and IPR000320, can you describe the architecture of this protein? How does it differ from the architecture presented by the 'Homologous superfamilies' section at the top?

   d. What is the biological role of this protein? (*Hint*: click on the name of the protein family.)

   e. Does this protein have a signal peptide?

   f. How could you know, at a glance, that this protein has a peptidase activity?

```
>squirrel_seq | example from EBI Train Online
MALPARLVPLCCLALLALPAQSCGPGRGPVGRRRYVRKQLVPLLYKQFVPSVPERTLGASGPAEGRVARGSERFRDLVPN
YNPDIIFKDEENSGADRLMTERCKERVNALAIAVMNMWPGVRLRVTEGWDEDGHHAQDSLHYEGRALDITTSDRDRNKYG
LLARLAVEAGFDWVYYESRNHVHVSVKAGTVGGGCFRETEAAQLWGDARGLRELHRAWVLAADAAGRVVPTPVLLFLDRD
LQRRASFVAVETERPPRKLLLTPWHLVFAARGPAPAPGDFAPVFARRLRAGDSVLAPGGDALRPARVARVAREEAVGVFA
PLTAHGTLLVNDVLASCYAVLESHQWAHRAFAPLRLLHALGALLPGGAVQPTGMHWYSRFLYRLAEELLG
```

Protein Data Bank (PDB)

http://www.rcsb.org
http://www.ebi.ac.uk/pdbe

The Protein Data Bank (PDB) is a collective repository of the atomic coordinates of proteins and other biological molecules. These have been determined experimentally by X-ray crystallography, NMR spectroscopy and, more recently, cryo-electron microscopy. PDB files provide information on the tertiary structure of macromolecules, and at the same time, they allow users to inspect their secondary structures, domain arrangement and global folds. They can also be used as a high-confidence piece of evidence about interactions, from binding of small ligands to formation of oligomers and protein complexes.

PDB mainly stores coordinate files, which list each atom in a structure and their three-dimensional position in space. (Besides the main molecule, it can also hold small molecules, ions, water and other ligands). These files also have a 'header' section with information about the protein, the experimental procedure, reference bibliography, etc. All data should be stored as a plain text file and respecting a certain layout to allow an easier identification of the different sections.

A single PDB archive is maintained by three independent distribution sites, the RCSB PDB in USA, PDBe in Europe and PDBj in Japan. Although the main data and resources are shared, each site provides a set of exclusive services to users for the inspection of data. All PDB entries in these databases receive an identifier in the form of a four-character accession code. This should start with a number and be followed by any three alphanumeric characters, in upper or lowercase (e.g., 2c3v).

Although PDB files can be inspected with a text editor (to review the header data, for example), it is usually better to use a dedicated visualization program. It will display the structure on a virtual 3D coordinate system, allowing the user to zoom, rotate and translate the structure, change its representation, display bonds and calculate distances, find structural features of interest, etc. These tools can be accessed online and are available at the PDB sites, but more powerful and versatile programs (like UCSF Chimera or PyMol) should be downloaded.

Some useful things to know about PDB:

- Most PDBs have a 'chain ID' associated with its atoms. This is a unique identifier (usually a one-letter code starting on A) for all residues or nucleotides belonging to the same polypeptide or nucleotide chain, respectively. Thus, chain IDs help recognize the different molecules present in the file.

- PDB files normally have one model for each molecule they contain. However, a PDB entry may carry one or multiple models of the same molecule. For example, due to the characteristics imposed by the technique itself, structures solved by NMR usually have 20 alternative models in the same file.

- It is generally better to inspect first the 'Biological assembly' than the 'Asymmetric Unit' of an entry. The latter is useful for the crystallographer to refine the coordinates of the structure against experimental data. However, the biological assembly should reflect the functional form of the molecule.

- A central descriptor of the quality of the experimental data, and thus of the resulting PDB file, is its resolution. Expressed in Å (Ångstrom), low resolution values give more confidence on the location of atoms, while values exceeding 3 Å only allow a rough determination of the basic contours of the molecule.

- Many PDB entries have 'missing residues'. These are portions that were not observed during the experimental determination of the structure, possibly due to an increased flexibility. These could range from short loops within globular domains to long disordered segments.

- You should never assume blindly that a PDB file will follow the sequence of the corresponding UniProt entry. They may differ due to decision of the experimentalist or technical difficulties.

- Structure files must comply with the PDB format. Lately, PDB entries have been also made available in other interconvertible formats like mmCIF and XML, which are easier to understand by computers.

Exercises

1. Do you have a protein of interest? Using RCSB PDB, can you find its structure? (You may not know if someone solved it yet, or at least the structure of a homologous protein. You may BLAST your sequence against the PDB database and inspect the results for these hits.)

2. In this tutorial we'll work with the **oxy-myoglobin** of *Physeter catodon* (sperm whale). The PDB ID of this entry is 1A6M. Search for it using the top-right box.

   a. Use the buttons on the top right to download a PDB file of the protein.

   b. Inspect the 'Structure Summary' section. When did this structure first appeared in the PDB? Is it a good quality structure?

   c. On the 'Macromolecules' section of the same page, you'll find a mention to the Myoglobin entity. How many chains does it have? Can you find its UniProt accession?

   d. Still there, the 'Protein Feature View' provides a site-specific mapping between UniProt and PDB, with many additional features from other databases. What kind of secondary structures does this protein adopt? Besides, can you find the iron binding sites? (*Hint*: use the zoom buttons on top).

   e. Does this structure have a bound oxygen molecule?

   f. Go to the '3D View' tab, where you can play around with a three-dimensional representation of the protein structure. Hold the left button over the structure and move the mouse to rotate the structure. Use the right button to move it. Zoom in and out using the mouse wheel.

   g. Click over the structure to identify atoms. Can you find any of the iron binding sites you've seen before?

   h. Now, on the 'Display Options' section on the right, use the 'Interaction' drop-down menu to focus the display on the interaction with the oxygen atom carried by the heme group. Move around the structure until you can identify the residue numbers. Are they the same as above? If not, why?

   i. Set the 'Interaction' back to 'None'. Now play around with the 'Style' and 'Color' drop-down options to change the representation of the protein. You can mark the 'Spin' checkbox on the 'Viewer options' menu to see the structure from different sides.

   j. Check the 'Experiment' tab to see which pH was set on the crystallization experiments.

   k. Inspect the information on the 'Sequence Similarity' tab to find out how many protein chains in other PDB entries are 100% identical to this protein.

   l. Click on the number of chains in the 100% cluster and a table will load below showing the PDB entries of all these chains. Select 1A6M and any other structure solved at pH=4.0 (*Hint*: check the 'Details' column). Now, on the top of the list, select the 'jFATCAT-rigid' algorithm to perform a structural comparison of both structures. Wait until the superposition loads; Does the change in pH cause any significant difference between both structures?

   m. A similar superposition can be obtained on the 'Structure Similarity' section. Here, the server proposes 3QM9 as structurally similar to 1A6M, even though they only share 40% sequence similarity. Select 'jFATCAT-rigid' on the lower right to compare them. Do you see more differences than in the previous comparison?

   n. Without leaving these results, browse to the 'Alignment Block(s)' section below to see a structure-base sequence alignment of 1A6M and 3QM9. Judging from the sequence similarity, the structural superposition and the conservation of residues, would you say 3QM9 could is also a myglobin?

TMHMM

http://www.cbs.dtu.dk/services/TMHMM

TMHMM is a dedicated server for the prediction of transmembrane helices in proteins. Although it was first presented two decades ago, it has been updated and is still a reference for the task.

The program is based on the development of a Hidden Markov Model that served as a predictive tool. It was originally trained with a set of 160 membrane proteins. Its good performance was determined by a 10-fold cross validation and the comparison with a negative set of 645 non-membrane proteins from the PDB.

The first output of TMHMM is a set of statistics and a list of the predicted transmembrane helices. This list is the most relevant part, as it maps the start and end of each predicted helix and loop region. The location of these loops, either 'inside' or 'outside' the cell, is also given. Therefore, the server makes it possible to trace the path of the protein from one side to the other of the membrane.

A posterior probability plot allows identification of strong transmembrane segments that made it to the final model and weak segments that were not considered good enough. The HMM calculates the total probability that a residue is part of a helix, an inside loop or an outside loop; then combines these assessments on the final model.

Some useful things to know about TMHMM:

- The server does not accept sequences longer than 8,000 amino acids.
- If the whole sequence is labeled as inside or outside it means that no membrane helices were predicted.
- Raw data can be downloaded from the link in the bottom of the results and used to create custom plots.

Exercises

1. The interface to TMHMM Server 2.0 is quite simple. Use the search box on the homepage to run the prediction for **bacteriorhodopsin**, which sequence is pasted at the bottom of this page. Use the default output format ('Extensive, with graphics') to get more descriptive results.

2. Inspect the results of the search. How many transmembrane helices did you find?

3. Among the statistics you'll find the 'Exp number of AAs in TMHs'. This is the expected number of amino acids in transmembrane helices according to this HMM prediction method. When this number exceeds 18, the protein is likely a transmembrane protein. Is bacteriorhodopsin a transmembrane protein?

4. Another statistic is 'Exp number, first 60 AAs'. This is the same as above but limited to the first 60 amino acids. If this number is not small, then a transmembrane helix that was predicted in the N-terminal region could be a signal peptide instead. (You could try some of the dedicated tools for prediction of signal peptides, such as SignalP).

5. In the plot, red blocks correspond to transmembrane helices, blue lines indicate regions on the inside, and exterior segments are shown in purple. Could there be another transmembrane helix that the model is discarding? (Hint: look at the plot).

```
>AAA72184.1 bacteriorhodopsin
MQAQITGRPEWIWLALGTALMGLGTLYFLVKGMGVSDPDAKKFYAITTLVPAIAFTMYLSMLLGYGLTMVPFGGEQNPIY
WARYADWLFTTPLLLLDLALLVDADQGTILALVGADGIMIGTGLVGALTKVYSYRFVWWAISTAAMLYILYVLFFGFTSK
AESMRPEVASTFKVLRNVTVVLWSAYPVVWLIGSEGAGIVPLNIETLLFMVLDVSAKVGFGLILLRSRAIFGEAEAPEPS
AGDGAAATS
```

Additional Resources

InterPro: Quick tour

https://www.ebi.ac.uk/training/online/course/interpro-quick-tour

InterPro: Functional and structural analysis of protein sequences

https://www.ebi.ac.uk/training/online/course/interpro-functional-and-structural-analysis-protei

InterPro FAQS

https://www.ebi.ac.uk/interpro/faqs.html

Protein classification: An introduction to EMBL-EBI resources

https://www.ebi.ac.uk/training/online/course/protein-classification-introduction-embl-ebi-resou

PDBe: Quick tour

https://www.ebi.ac.uk/training/online/course/pdbe-quick-tour

PDBe: Exploring a Protein Data Bank (PDB) entry

https://www.ebi.ac.uk/training/online/course/pdbe-exploring-protein-data-bank-pdb-entry

PDBe: Searching the Protein Data Bank

https://www.ebi.ac.uk/training/online/course/pdbe-searching-protein-data-bank

PDBe: Searching for biological macromolecular structures

https://www.ebi.ac.uk/training/online/course/pdbe-searching-biological-macromolecular-structure