Manual

Welcome to Promod's manual. Here you will find all the information you need to run this software. This program is a project for the Structural Bioinformatics and Introduction to Python subjects of the MSc in Bioinformatics for Health Sciences at Universitat Pompeu Fabra (Barcelona).

What is Promod?

Promod is a Python tool whose goal is to model macrocomplexes of molecules starting from the interacting pairs of chains that will form the complex. It is compatible with protein chains, single and double DNA strands and RNA strands. It has several parameters available to adjust according to the users needs.

In essence, the builder approach is similar to a genomic assembler: it seeks for similar chains in the different PDB files given as input and overlaps them. The builder places pair by pair all the possible chains into a single model.

Background and scientific explanation

Introduction

Proteins are the executive molecules in all organisms. They perform a wide variety of functions, from small compounds transport to signals transduction or immune responses. However, proteins do not usually work individually in cells. Most proteins interact with other proteins (protein-protein interactions, PPIs) or molecules (DNA, hormones, drugs, among others), which are essential for a proper development of their activities. That is the reason why collecting PPIs can lead to a better understanding of protein functions, biological pathways and mechanisms of disease. The identification of such PPIs has been a challenging task for years. Experimental methods provide irrefutable evidence to test interactions between proteins, but they are expensive and time-consuming. Nowadays, computational approaches are being developed to reduce as much as possible the necessity of experimental data. Some advances have been made in this field, but the range of successful organisms is short and general frameworks are lacking at the moment.

Different experimental approaches for the identification of individual PPIs are available. The most accurate methods that allow the determination of the exact atom coordinates in the complexes are X-ray crystallography, Nuclear Magnetic Resonance and Electron Microscopy. These techniques are difficult to perform, they need very specialised equipments and a strong data analysis procedure. There are also high-throughput methods based on biochemical properties, such as Tandem Affinity Purification, which can detect multiple PPIs at once, but a high rate of false positive results can be expected and the information is not very precise. Protein microarrays are also being used as a screening method that can be easily automated and parallelised. Other traditional techniques, namely co-immunoprecipitation, yeast two-hybrid or pull-down, are still being used as well. All these approaches provide complementary views of PPIs and have their advantages and problems. The main one is the cost-effectiveness of the experiments, making in silico approximations more feasible and adequate for understanding PPIs at the atomic level.

Regarding the structural properties of the proteins, there are three main categories of methods for computational modelling of PPIs: homology or template-based modelling, *ab initio* or template-free modelling, and hybrid or integrative modelling. Homology modelling is based on the fact that the evolutionary information in both the sequences and the structures is important for PPI prediction. The latter are preferred, as most existing predictors use surface patch data, and only the residues in the interface have to be analysed, instead of the whole sequence of aminoacids. A pitfall of this method is that not all the 3D structures are available, although public databases such as PDB are unstoppably increasing in size. Template-free modelling needs more computational resources, as it explores all the possible orientation between the interacting molecules. The combination of prior knowledge about the individual structures of the components may help reduce the searching space, but still this approach is more challenging. This kind of techniques also require a careful

evaluation of the results by various means and refinement of the candidate models using biological information. Finally, integrative modelling combines experimental data and bioinformatics developments to narrow the possible complexes and save computational resources and time.

In this work, our aim is to develop a software that builds protein macrocomplexes using as input pairs of protein-protein interactions. Interaction with other molecules, such as nucleic acids, is also considered. To do this, our program is based on homology modelling. Therefore, the homology of the different chains is analysed, tridimensional structures are superimposed and energy levels of the final models are considered to propose the best possible solution. Another option would have been template-free modelling, but the required computational resources and the steps of evaluation and refinement excluded this possibility, because of the limited means and knowledge that our team has got.

Here we propose Promod as a first step approach to protein-protein and protein-nucleic acid complex modelling. It is distributed as both a Python package and a standalone application to be run in both UNIX-based and Windows operating systems. Several parameters can be tuned to adjust to the user's needs. The final result is a single model with the best option according to the algorithm that is explained below.

The Promod method

The algorithm used by our program can be divided in three main steps. The first one is the analysis of the homology of the subunits. Then, the superimposition of the homologous structures is performed. Finally, energy levels in the models are taken into account to discard unlikely complexes.

Homology of the subunits In order to build the model, the program begins with several files, each of them containing information about two interacting chains, whether peptidic, DNA or RNA. We can overlap similar proteins from two different PDB files to check if two chains in those files are alike. If they effectively are, they can be joined in the same model. This is known as protein superimposition, and the whole process can be compared to a DNA sequence assembly. The main objective is to recognise that two sequences are the same and put them together in the correct spot.

Two chains in different PDB files must be homologous to be considered as part of the same protein and overlap them. To check if two proteins are homologous or not, a pairwise alignment is performed. This alignment consists of calculating the similarity between two sequences, taking into account both mutations and gaps, and then returning a score between 0 and 1. A score close to 1 means that both sequences are identical, therefore high score values are evidence of a large proportion of sequence identity. Only the homologous proteins will be overlapped and used to build the model later.

Superimposition of the 3D structure Two homologous sequences will probably have similar structures. However, it is also possible that two proteins with lower identity score may have similar structures in the tridimensional space, as a result of convergent evolution. Therefore, when two proteins are significantly different in sequence but they have similar structures, the may have the same role in our model and it is desirable to consider them. For example, two homologous proteins from distant species whose alignment do not pass our homology threshold, but they still preserve the structure.

In these cases, to test if two proteins really have similar structures, we need a measurement, such as the root median square deviation (RMSD). First, to calculate this value, we have to superimpose these two proteins. That means placing the atoms of both proteins in the same coordinates and orientation, to check how well they overlap in the space. Similar structures will have atoms in almost the same positions, while different structures will be more distant. This similarity between the superimposed proteins can be used to calculate the RMSD, which is the average distance between the superimposed atoms in the chains. A value close to zero indicates a perfect fit of the structures. RMSD will increase when the differences between the protein structures increase.

Due to all this, RMSD must also be a filter to account for those structures that are different, even when we had considered them homologous in previous steps.

Analysis of the energy levels Lastly, it is important to consider the final energy levels of the complex. A good model should have minimum energy, as functional, naturally occurring complexes are the ones with the lowest energy among the set of possible foldings. Situations such as two atoms very close to each other, aminoacids with hydrophobic residues located in the external part of the macromolecule and several other cases can increase the final energy of the complex, which should be then corrected.

After that, it is possible that the position of the atoms inside the model is not the most adequate. Sometimes, despite having evidence of interaction between two chains, collisions or clashes appear when they are joined in the structure. Taking this into account avoids impossible models, such as those with chains crossing with each other, which will be thermodynamically unstable and unlikely to happen.

Features

The most striking features of Promod are:

1. Building of macromolecular complexes from basic input data (pairs of interactions between molecules).
2. Optimization of the final model using MODELLER.
3. Graphical user interface (GUI), with the same functionalities as the command line interface (CLI).

Implementation

The Promod package is implemented in Python3. It strongly depends on the Biopython library, which is an installation requirement for the correct execution of this software. The GUI has been developed under the Tkinter framework, providing a user-friendly interface and, thus, avoiding the use of the command line. The implementation of the GUI and the CLI are independent, so each one can be executed on its own. This software works with well-established bioinformatics formats, such as FASTA and PDB files, so no atypical formatting of input data shall be conducted. Additionally, two scripts are provided to divide a given PDB file in separate pairs of chains and join them if there is an interaction between the molecules. Some examples are also included for the testing of the program. Promod is freely available from the following Github repository. URL: https://github.com/Fabian-RY/SBI-Python-project.

The formatted documentation of the package includes dependencies and installation instructions, examples of use and a full tutorial with sample code. Each function has got its own documentation that instructs on the particularities when importing them to be used independently. The users could learn all knowledge of the library by looking up the detailed documentation. The main functionalities are further explained in the tutorial section.

Discussion

Promod is a basic tool that relies on information from both the sequence and the tridimensional structure of pairs of interacting molecules to build a final complex model. As we can see in the analysed examples, it is successful with small complexes and it can even handle nucleic acids interaction with proteins. The MODELLER final step adds an additional refinement of the proposed model, returning to the user a good quality structure. All this methodology is built in an easy-to-use package, well documented and with a GUI to save unpleasant command line work for the standard user.

However, the main objective of modelling software is to provide new knowledge without the need of huge amount of experimental data, such as the determination of all paired interactions, in this case. A limitation of our program is the necessity of input protein-protein or protein-nucleic acid interaction. This type of software is being broadly developed by means of various approaches, which are proving to be very successful. Both docking and homology modelling paradigms are being applied to achieve this goal.

One of the strategies being exploited is evidence combining methods. The core of this strategy is integrating evidence from multiple sources, including them in comprehensive databases for later integration. They are called gold standard databases and they contain information for both training and testing of the new methods. The annotation of paired protein interactions goes beyond structural features. For instance, evolutionary relationships, functional features, network topologies, sequence-based signatures, structure-based signatures, and text mining information is recorded in these databases. Finally, machine learning algorithms are fed with subsets of data and performance is measured to propose the better candidates, depending on target species, data sources, demand of accuracy and coverage. These evidence combining methods are performed repeatedly to find converging results with different input data and classifiers.

As we can see, template-based methods are leading the in silico modelling techniques. It is reasonable to believe that there are a limited number of possible interactions and that, once we have big enough databases and curated interaction catalogs, most of the PPIs should be easy to model with high confidence. However, *ab initio* modelling has also yielded promising results, mainly from competitions such as CASP, CAPRI or CAMEO, where world-class groups bring their latest developments to test their performance with real problems. Of course, these solutions are computationally expensive and require a long time to return a final model. In addition, complexes with weak interactions where the conformational state changes upon binding are a big challenge for docking software.

A large number of information is nowadays available from high throughput experimental techniques and a lot of structural bioinformatics software has been developed to integrate these data. The best performance of in silico techniques has been achieved at the tertiary structure level of proteins. Nevertheless, quaternary structures are the ones responsible for the majority of biological functions and their knowledge is essential to disentangle protein interaction networks in both physiological and pathological scenarios. It is clear that hybrid approaches, joining atomic-level experimental structures, database information and the newest machine learning techniques, will give promising results in the near future.

Future perspectives

As we have already mentioned, there are diverse strategies for modelling PPIs. The development of refined algorithms to perform this task is far beyond our current knowledge in structural bioinformatics. Therefore, if we had time and resources to expand our project, we would focus on improving the user experience and the accessibility to existing data.

First, we would like to implement different ways of modelling. It would be useful for the user to choose whether to use a template-based or a template-free approach. We could also use existing Biopython packages to perform certain operations, but further research and testing would be needed to decide the better candidates.

Next, the result of the modelling operations could return automatic reports about the process. Not only the final model would be reported, with the tridimensional structure and their characteristics, but also information about the reasons why our software has discarded certain possible models. Therefore, direct visualization of these other options and energy plots for user reference should be displayed. A good option for the format of this report would be a Jupyter notebook, so the user can interactively check all the available resources.

Another functionality that could be feasible to achieve is the automatic download of structures from public databases, such as PDB. Both sequences and structures can be obtained using Biopython modules and some keyword and ID search should be easy to implement, as well. These would be the input data for the additional scripts that build the pairs of interacting molecules to run the core Promod builder.

Finally, it would be ideal to develop integration options of biological information to help build better complexes, in line with trending research in the field. We are not aware of Python packages that would allow to directly implement this kind of work, but maybe external tools are capable of doing it. However, a broad review of the latest literature and a challenging programming development would be needed to know the most promising approximations.

Conclusion

Although computational approaches for building macromolecular complexes are far to be perfect at the moment, a lot of effort is being made to develop strategies to overcome the pitfalls in this field. We have provided a software that, despite not using any novel strategy, achieves notable results when using already defined structures. Further steps using machine learning approaches and a broader range of training data would improve the performance and confidence of this type of programs. Finally, the integration of different types of biological data will also be a key step in the progress of in silico modelling of protein-protein interactions. The achievement of better methodologies will definitely have an impact in applied fields, such as drug discovery, biomedical research or food industry.

Bibliography

Chang, J., Zhou, Y., Qamar, M. T. U., *et al.* Prediction of protein–protein interactions by evidence combining methods. *International Journal of Molecular Sciences* **17**, 1946 (2016). doi: 10.3390/ijms17111946

Ding, Z., Kihara, D. Computational identification of protein-protein interactions in model plant proteomes. *Scientific Reports* **9**, 8740 (2019). doi: 10.1038/s41598-019-45072-8

Hayes, S., Malacrida, B. , Kiely, M., Kiely, P. A. Studying protein–protein interactions: progress, pitfalls and solutions. *Biochemical Society Transactions* **44**, 994-1004. doi: 10.1042/BST20160092

Liu, S., Liu, C., Deng, L. Machine learning approaches for protein–protein interaction hot spot prediction: progress and comparative assessment. *Molecules* **23**, 2535 (2018). doi: 10.3390/molecules23102535

Nealon, J. O., Philomina, L. S., McGuffin, L. J. Predictive and experimental approaches for elucidating protein–protein interactions and quaternary structures. *International Journal of Molecular Sciences* **18**, 2623 (2017). doi: 10.3390/ijms18122623

Sarkar, S., Gulati, K., Kairamkonda, M., *et al.* Elucidating protein-protein interactions through computational approaches and designing small molecule inhibitors against them for various diseases. *Current Topics in Medicinal Chemistry* **18**, 1-18 (2018). doi: 10.2174/1568026618666181025114903

Keskin, O. , Tuncbag, N. , Gursoy, A. Predicting protein-protein interactions from the molecular to the proteome level. *Chemical Reviews* **116**, 4884-4909 (2016). doi: 10.1021/acs.chemrev.5b00683

Tutorial

Installing dependencies

In order to make Promod work, there are some dependencies that must be installed before executing the software: Python3 and Biopython.

If you are using Windows, you can download Python3 from its website. If you have got Linux, it is likely that you already have got it installed.

Independently of the operating system you run, you can install Biopython using `pip`.

```
pip3 install biopython
```

Installation

The recommended way to install Promod is by using `pip`, exactly with the same command as we indicated with Biopython above.

```
pip3 install promod
```

You can install Promod easily by downloading it from the Github repository and running the next command in the downloaded folder. That is the recommended installation procedure.

```
pip3 install .
```

Alternatively, you can also run the `setup` installation script, as follows:

```
python3 setup.py install
```

Hands on: how to use Promod

Promod has a command line interface which explains briefly how to use it before execution, when using the `-h` flag.

```
promod -h
```

There are 3 mandatory commands, while the rest are optional. The mandatory ones are related to the input files and output folder, required for the program to work. They are `-i`, the input folder; `-o`, the output folder; and `-f`, the fasta file of the sequences. All of them will be covered later in this manual.

It also has a graphical interface which accepts the same parameters, but graphically with message boxes and graphical aid. It is an independent executable, so the CLI can be used for automation by itself. The command to start the GUI is simple:

```
promod-tk
```

Input The input must be a folder with two or more files: each file containing two proteins, a protein and a DNA or RNA strand or 2 single DNA strands (that do not necessarily form a double strand), which represent an interaction between those chains. Two structures are considered to be interacting if the minimum distance between them is in a range of a few Angstroms (1-10 Å) . However, in shorter distances, the forces between atoms are strong enough to produce changes between the chains and promote a different conformation, and, thus, a realistic model must keep the residues at an adequate distance to consider it correct.

Output The output is mainly one PDB file with all the possible chains joined in the selected folder. However, if the optimized option was selected, several PDB files will be created in the same directory. These are different approaches made by MODELLER to optimize the energies of the model and the files it used. The model will be saved as `final_model.pdb`.

If the model is required with minimum energy, and thus the `-optimize` flag is used, then several files will be saved. MODELLER will save some mid-step PDB files as `final-model.DXXXXX.pdb` and the fully optimized complex will be saved as `optimized.pdb`.

Parameters

Input folder This is a mandatory argument and should be indicated with the `-i` or `--input-folder` tags. The input folder should contain, at least, 2 PDB files. This folder may contain other files or subfolders. However, the program will ignore other files and will not check subfolders to find more PDB files. If you wanted to include some other PDB, you should copy it in the folder before running the program.

Output folder This is a mandatory argument and should be indicated with the `-o` or `--output-folder` tags. The output folder is the directory where the output files are written. At the moment, the computed model file is saved as `final_model.pdb` and, thus, any other file with the same filename will be overwritten.

Fasta file This is a mandatory argument and should be indicated with the `-f` or `--fasta` tags. The fasta file contains the sequences of the chains and the identifiers for the stoichiometry. It is a critical file, as the sequences must be homologous to the chain in the PDB file. We can hold up to 5% of differences (default threshold value which can be modified using the `-t` parameter). However, if one of the chains (if stoichiometry is not provided) or one of the chains in the indicated stoichiometry differs largely, the program will exit. The reason for this behaviour is to avoid guessing possible chains that could randomly match, as that would produce unexpected results.

Note: The sequences in the fasta file should be of a similar length than their counterparts in the PDB file. For instance, the fasta files and the PDB files downloaded directly from Protein Data Bank may differ in the initial and ending residues, as they are difficult to model, so they are usually removed from PDB entries.

Distance This parameter can be indicated with the `-d` or `--distance` tags. You can indicate the minimum distance to consider that two proteins do not clash. The model will discard interactions with too many atoms at lower distance than the value indicated in Angstroms. Being too restrictive may discard some meaningful interactions, but being too flexible can force chains to overlap and produce interactions that are not energetically favorable.

Threshold This parameter can be indicated with the `-t` or `--threshold` tags. The threshold is the minimum score an alignment must reach in order to insert the chain in the model and determine its homologous protein. A very high value (0.9 or more) is usually recommended to ensure that the chains are correctly identified. However, for sequences with less identity, it might be needed to decrease this value.

Stoichiometry file This parameter can be indicated with the `-s` or `--stoichiometry` tags. An additional file with the stoichiometry can be provided to make sure that the final protein will contain a specific number of chains. This file must be properly formatted, one value per line, as follows:

```
chain_id_in_fasta,number_of_columns
```

For example, this is the stoichiometry file for the example number 1:

```
3e0d_A,2
3e0d_B,2
3e0d_C,2
```

Starting PDB This parameter can be indicated with the `-start` or `--start` tags. Promod's result is highly dependent on the first PDB used to build the model. In fact, results can vary a lot when different starting PDB files are used. Therefore, we provide an optional argument to select the starting PDB file to account for this variation. This allows to use the same PDB file with the same parameters, to obtain the same model. By default, the PDB files are sorted alphabetically and the first one in the list is selected.

Uninstalling

You can uninstall this software by using the following command, if it was installed using `pip3`:

```
pip3 uninstall promod
```

However, if it was installed via `setup.py`, all files must be deleted manually.

Analysis of examples

Example 1 (3e0d)

3e0d is a small complex formed by 2 protein chains and 2 double DNA strands. It is a subunit of the eubacterial DNA polimerase, but it is perfect for an initial testing of our project. We can test how the program works, the time it takes to run and if the result is similar to the original one. In this case, we have 6 different interactions, as each DNA strand is counted as a single one interacting with another one and bound to the protein chain by one of them. In fact, this complex can be thought as a dimer: two monomers formed by a protein and a double strand DNA, which interacts with some residues in their protein chains.

The first test to our program consists of joining the different PDB files into one single structure, without further information about the stoichiometry. Therefore, the program will try to build the protein with only the information provided by the PDB files and the fasta file. The command to build it is the next one:

```
promod -i examples/example_1/chains/pairs -o examples/example_1/results  \
        -f examples/example_1/3e0d.fa
```

We only need to provide the folder with the input PDB files (`-i`), the desired output folder (`-o`) and the fasta file with the sequences of the proteins (`-f`). The program will take the different PDB files on the folder (by alphabetical order, to make it reproducible) and the result is incomplete:

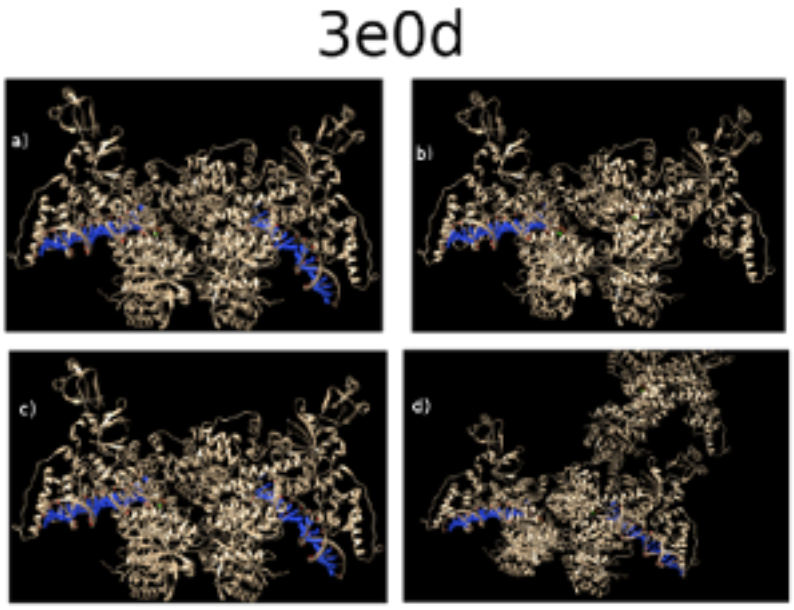

Figure 1: 3e0d structure obtained from: a) rcsb.com b) promod without any extra info c) promod with starting point and selected stoichiometry d) without stoichiometry control

We can observe that there is a missing double strand DNA. So, we will take a more elaborate approach to complete the model. As there is a missing protein, we should force the program to include it. There are two different ways of indicating this. The easiest one is selecting the starting complex of our model (which contains those missing parts). Different starting PDB files can result in different models, so, it is important to select a suitable one. By default, PDB filenames are sorted alphabetically and the first one is chosen as the starting point. As we said before, this is an important choice that is evidenced when we compare not choosing a starting point (Figure 1.b), choosing an adequate file (Figure 1.c) or choosing another one (Figure 1.d), which in this case results in models with additional or less chains than expected.

The other available option is to indicate the stoichiometry of the complex. We will focus on selection the starting PDB file, so the running command would be:

```
promod -i examples/example_1/chains/pairs -o examples/example_1/results \
       -f examples/example_1/3e0d.fa -start examples/example_1/chains/pairs/3e0d_ZG.pdb
```

This is more similar to the original structure, very close to the initial model. However, this is a very simple protein, and bigger complexes may be harder to build. However, from this example, we have learnt that:

- The program can build a model without any indication. However, it might be incomplete or it might have additional chains.
- The starting PDB file is an important choice that can influence the final model. Thus, it is important to remember which starting point gave which result. The same starting PDB file will give the same output.

Example 2 (6gmh)

6gmh is a very big complex formed by 24 different chains, with a double strand of DNA. There are several chains that interact between them, but there are not repeated sequences: each monomer is unique.

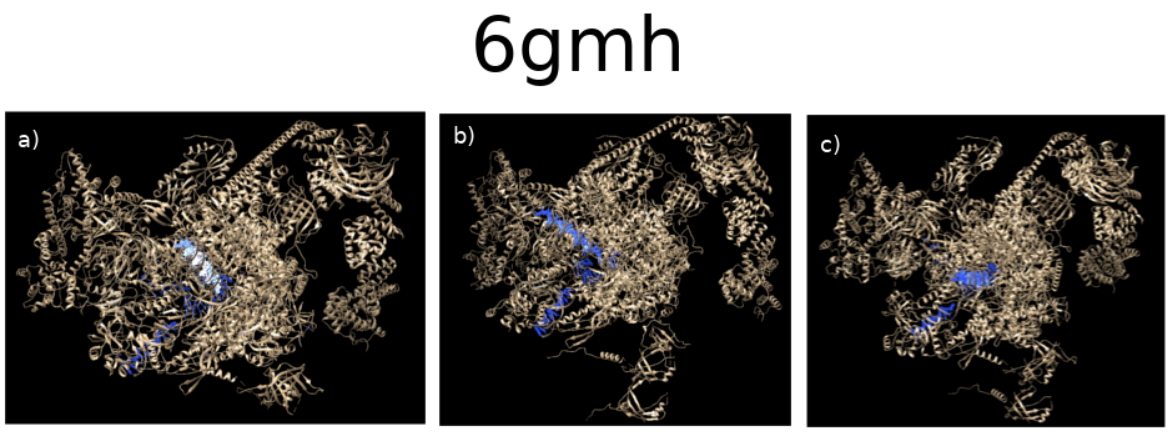

Figure 2: 6gmh structure obtained from a) rcsb.com b) builded without assistance c) selecting starting point and stoichiometry

Our first try is the following one:

```
promod -i examples/example_2/chains/pairs -o examples/example_2/results  \
        -f examples/example_2/6gmh.fa
```

In this case, there are several chains that has not been added to the model, and some others are repeated. We should select a different starting point and limit the chains with the stoichiometry, so there is one copy of each chain, at most.

```
promod -i examples/example_2/chains/pairs -o examples/example_2/results  \
        -f examples/example_2/6gmh.fa -start examples/example_2/chains/pairs/6ghm_VA.pdb \
        -s examples/example_2/6gmh.stoic
```

The `6gmh.stoic` file contains the stoichiometry of the protein. Basically, it is a csv file, without header, in which each line follows the same structure: ,. Each chain_name must be in the fasta as a sequence identifier.

This starting PDB file was not in the model in our first attempt, so now we force it to be included in the computation, and start building from there. This way, the result contains more chains in the model and it is more similar to the original one.

Example 3 (5nss)

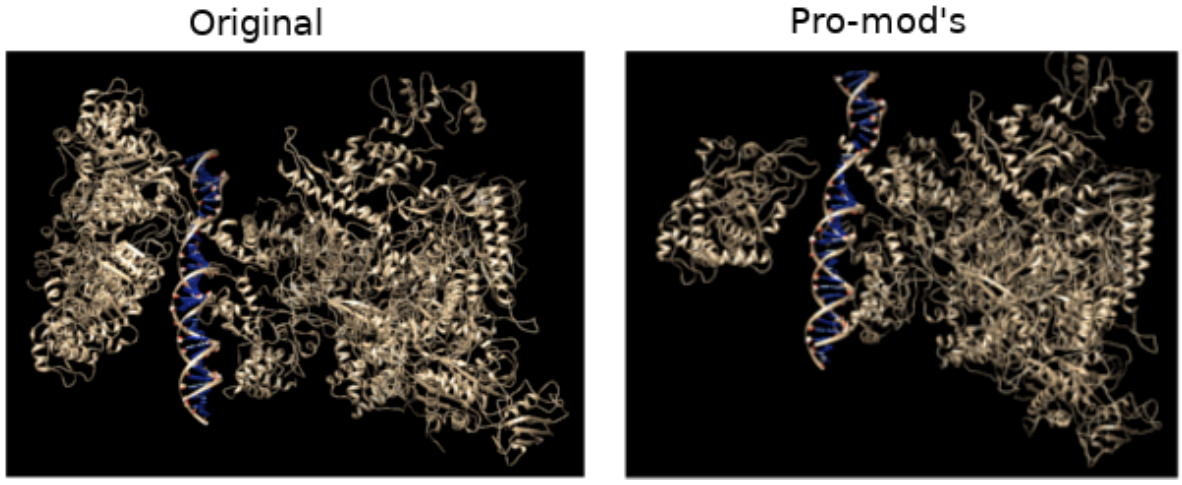

Figure 3: 5nss structure obtained from rcsb.org vs composed one

Here we want to show how to use two parameters we have not shown yet: distance and threshold. The distance is a value that avoids the situation when two atoms collide, which is not allowed because two very close atoms would have high energy. So, if the separation between them is less than the indicated distance, it is considered invalid. However, to avoid some errors, up to 10 colliding atoms are allowed. The threshold makes reference to the identity percentage between the PDB structures and the fasta sequences. This allows certain flexibility in sequences with some mutations that may alter the conformation of the chain.

```
promod -i examples/example_3/chains/pairs -o examples/example_3/results  \
        -f examples/example_3/5nss.fa -d 2 -t 0.9
        -start examples/example_3/chains/pairs/5nss_NG.pdb
```

In this case, the result is quite good, but an important issue is the speed of the program. As the number of interactions grow, the time it takes the program to complete the modelling also increases. However, there are a few considerations about that:

1. Adding a new chain that satisfies all the restrictions is the most computationally expensive situation. The number of attempts increases with the number of chains in the model.
2. It takes less time to discard an interaction that has not an homologous molecule already in the model than adding a new chain. During the first steps of the program, until the model has grown enough, this is the most common case. The number of calculated alignments is relatively low as there are a few chains in the model, but this may become an issue at certain point.
3. Once the model has a considerable number of chains, and particularly if there are several copies of a monomer, discarding an homologous protein because it does not fit with the current parameters is a very time-consuming process. The program will try to introduce the chain using all the possible alignments whose scores are higher than the threshold. This is an important problem for proteins with a lot of copies of several monomers.
4. The stoichiometry must be carefully selected to reach the desired structure. It may happen that the program cannot construct a model with the desire stoichiometry, but it is also possible that the stoichiometry was not appropriate for such case.

Example 4 (6om3)

In this last example, we will cover the software behaviour in relation to the number of input PDB files and the influence of stoichiometric restrictions.

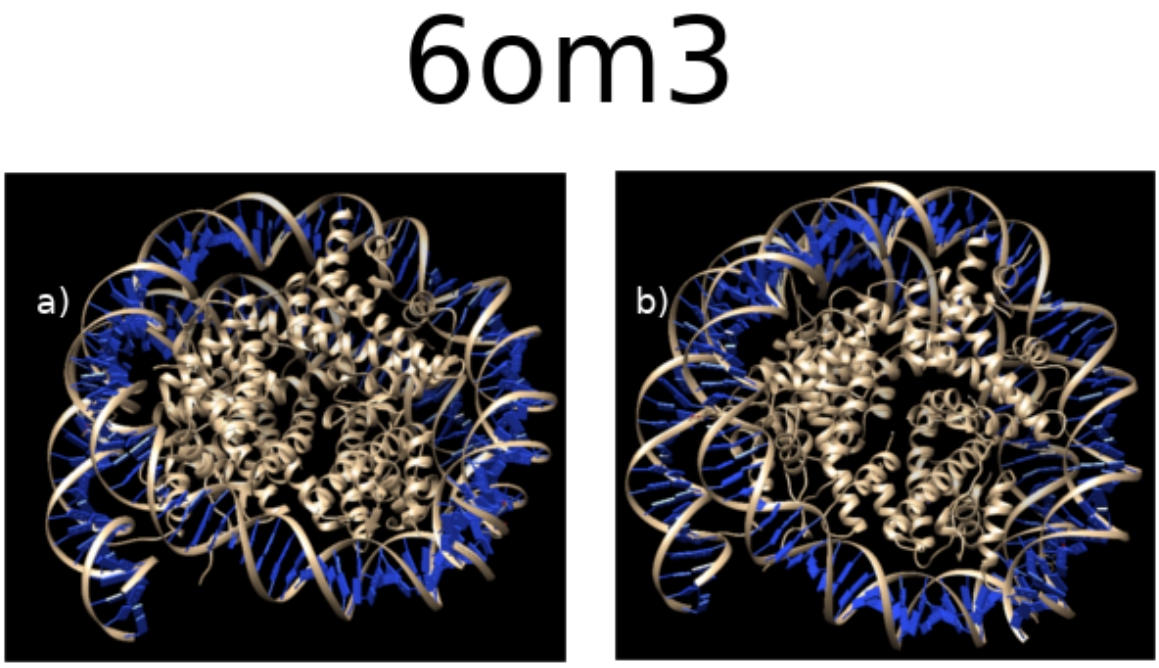

Figure 4: 6om3 builded from: a) rcsb.com b) promod with stoichiometry

```
promod -i examples/example_4/chains/pairs -o examples/example_4/results  \
        -f examples/example_4/6om3.fa -d 0.3 -t 0.95 \
        -s examples/example_4/stoic.csv
```

We can see that there are some absent chains in the protein complex built with Promod, which can be added using custom values of distance or an appropriate starting PDB file. Nevertheless, we will focus on another topic that has not been commented yet.

If you provide a stoichiometry file but it does not exist, the program will abort. We could have changed it to a non-stoichiometry builder, but we considered a better approach to explicitly state this situation and let the user manage this type of error.

An important thing about the stoichiometry is that it is critical to make sure that all the model is built following the user preferences. For example, without any assistance, the modelling in this example is good. However, with a selected stoichiometry, the result has some missing chains. Here we want to stress that the user should carefully check the stoichiometry file to make sure the chains are correctly selected and the number of them is adequate to our purpose.

Limitations

1. The sequences in the PDB files and the sequences in the fasta files must be similar. The program admits a certain degree of tolerance (which can be adjusted by the user using the `-t` parameter), but selecting sequences in the fasta file which are significantly different from the PDB structures will not

allow the assembly of the complex. This is particularly difficult for PDB files in which protein tails are not properly modelled and, therefore, they are not present in the structure, but they are contained in the fasta. The pairwise alignment with the tail will yield a low score, resulting in an incorrectly failed homology identification or even returning a fake corresponding sequence in the fasta just by chance.

Our recommendation is to prepare a fasta file as similar as possible to the structural sequence to avoid this kind of errors. We provide an additional script, `pdbsplit.py`, which can extract the sequences of the chains inside the pdb file, preventing this kind of issues up to certain degree. So, if you are investigating the effect of certain mutations in a target protein, it might be adequate to prepare the fasta file according to the PDB sequences.

2. The running time of the program is proportional to the number of PDB files selected and it is also affected by the order of the structures and the starting point. The most critical step is reading all the PDB sequences. However, as the number of interactions increase, more chains might be added to the complex. This is a trade-off between the number of interactions and computational speed.

3. Selecting different starting points can produce different models. It is uncertain to state which is the most adequate starting point, as this may vary according to the goal of the assembly.

4. The fasta files require to be specially prepared for this application. There should not be repeated sequences (with at least an alignment score equal or higher than selected threshold. This also applies to sequences with unknown aminoacids, which are usually marked as X in the fasta file and will cause a mismatch with the sequence in the PDB file. Those sequences will be usually missing in the final model.