|  |
| --- |
| ФББ МГУ, Москва |
| Описание программы PDBParser |
| Версия программы 1.3.1 |
|  |
| **Автор Котлов Никита** |
| **31.12.2014** |

|  |
| --- |
| [Введите аннотацию документа. Аннотация обычно представляет собой краткий обзор содержимого документа. Введите аннотацию документа. Аннотация обычно представляет собой краткий обзор содержимого документа.] |

Оглавление

[Назначение программы 3](#_Toc407822500)

[Условные обозначения и сокращения 3](#_Toc407822501)

[Описание программы 3](#_Toc407822502)

[Запуск программы 4](#_Toc407822503)

[Обычная работа с PDB файлами. 4](#_Toc407822504)

[Обновление базы сортировки резидью. 7](#_Toc407822505)

[Сбор результатов докинга. 7](#_Toc407822506)

[Создание файла лиганда из результатов докинга 7](#_Toc407822507)

[Корректирование лигандов. 8](#_Toc407822508)

[Вызов описания параметров. 8](#_Toc407822509)

[Параметры, применимые почти для любого запуска 8](#_Toc407822510)

[Параметры программы 10](#_Toc407822511)

[Примеры разных задач 13](#_Toc407822512)

# Назначение программы

### Условные обозначения и сокращения

**Лиганд** – молекула низкомолекулярного вещества (обычно неорганического), находящаяся в контакте с белком. Обычно лигандом в терминах докинга называют докируемое вещество. Название лиганда – название соответствующего резидью (см. ниже). Атомы лигандов должны быть записаны в поле HETATM.

**Рецептор** – молекула исследуемого белка.

**Структура белка** - файл, содержащий координаты рецептора. Обычно в файлах присутствуют координаты молекул растворителя, лигандов. Также записи нестандартных связей, которые по умолчанию не предусмотрены в стандартных аминокислотах и пептидной связи. Например, координационные связи с металлами.

**Цепь** - мономер из структуры, имеет свою букву (заглавную латинскую).

**Резидью** - остаток. Для белка это аминокислоты, для других веществ – это 1 молекула такого вещества. Каждый резидью имеет уникальный номер внутри одной цепи структуры белка.

**Парсер** – сокращенное название программы **PDBParser**.

### Описание программы

PDBParser – программа для работы с файлами в формате PDB3.1 в консольном режиме (без графического интерфейса). Основная цель парсера – облегчить работу пользователя с файлами во время подготовки докинга и после него. Функционал программы позволяет делать большое число разных действий, которые часто необходимы при подготовке докинга и просто при работе с файлами в формате PDB.

Новизна программы PDBParser в том, что большой функционал собран весь в одном месте.

# Запуск программы

Парсер можно запускать в одном из 5 режимов.

Обычная работа с PDB файлами.

Обязателен флаг «–s file». Задает входной PDB файл. Считанный файл записывается в специальную структуру данных, позволяющую удобно хранить параметры молекул, которые содержатся в PDB файлах. Структура pdb\_file разбита логически на сегменты (специфические структуры, позволяющие разделять логически PDB файлы по биологическим единицам и т. п., обычно не используется). Сегменты разбиты на цепи (в PDB файлах может быть несколько цепей). Цепи разбиты на поля ATOM и HETATM (логическое отделение атомов белка от атомов других химических веществ, таких как растворители, лиганды и пр.). Каждое из полей содержит множество резидью. А резидью разбиты на атомы.

PDB файл

Цепь 2

Цепь 1

Цепь 3

Сегмент 1

Сегмент 2

ATOM

HETATM

Резидью 1

Резидью 2

Резидью 3

Резидью 1

Резидью 2

Резидью 3

Резидью 4

Атомы

В момент считывания и сохранения происходит группировка элементов, принадлежащих одному родительскому элементу вместе (атомов, принадлежащих одному резидью или резидью, принадлежащих одной цепи), что может привести к изменению порядка следования атомов относительно исходной нумерации серийных номеров. Данная оптимизация позволяет сделать файлы более удобными и понятными.

При вводе записей связей (CONECT) происходит объединение записей, относящихся к одним и тем же атомам, что приводит к структуризации и компактизации области CONECT.

Записи ANISOW игнорируются.

Записи, идущие в начале файла (до координатной секции), могут игнорироваться или переноситься в выходной файл без изменений. Пока удалятся всегда.

Большая часть других флагов работает вместе с флагом «-s file».

«-os file» - выходной файл PDB. После всех изменений в него будет записаны новые параметры структуры.

«-og file» - выходной файл лиганда. Для того, чтобы задать, какой лиганд нужно записать в этот файл, необходим флаг «-hf file».

«-hf file» - входной файл HETATM. HETATM – файл параметров для парсера. Он должен содержать одну или более строк. Каждая строка должна начинаться со специального трехбуквенного обозначения. Пустые строки игнорируются.

* **LIG**
* **STD**
* **DEL**
* **REN**
* **CHA**

**LIG** – указание названия нестандартного лиганда. Данная запись может быть только одна, а при наличии нескольких будет использоваться последняя. После LIG должен быть 1 пробел и 3-4х буквенное название лиганда. Через пробел может быть написана большая буква – название цепи. Если цепь не указана, будет использована цепь A. Если вместо названия лиганда будет написано ключевое слово FIND, парсер предпримет попытку найти «нестандартный» лиганд в соответствующий цепи. Поиск и вырезание лиганда осуществляется исключительно в поле HETATM. Найденный лиганд будет записан в файл, заданный параметром «-og file».

Пример:

* **LIG** FLP A #нестандартный лиганд FLP в цепи A

Пример:

* **LIG** FIND C #найти нестандартный лиганд в цепи C

STD – указание считать резидью стандартным. Данных записей может быть много. После этого слова через пробел должно быть одно название резидью. Те лиганды, имена которых будут перечислены в строках, начинающихся с STD, будут считаться стандартными и не будут рассматриваться, как нестандартные (при указании ключевого слова FIND после LIG). Нестандартный лиганд будет определен по остаточному принципу. В случае нескольких будет использован первый найденный.

Пример:

* **STD** HOH #вода – стандартный лиганд
* **STD** BOG #BOG – растворитель, тоже стандартный лиганд

DEL – удалить резидью из поля HETATM. Данных записей может быть много. После ключевого слова через пробел должно быть одно название резидью. Через пробел должна быть указана цепь. Если вместо буквы цепи указан символ «!», резидью будет удален из всех цепей. Резидью из поля HETATM, имена которых будут указаны в строках, наинающихся с DEL, не будут записаны в выходной файл рецептора для соответствующих цепей. Если вместо названия резидью указано ключевое слово FIND, то если задан поиск лиганда (при наличии ключевого слова FIND после LIG), найденный нестандартный лиганд будет удален из соответствующей цепи.

Пример:

* **DEL** HOH ! # удалить воду из всех цепей
* **DEL** FIND A # удалить найденный лиганд из цепи A

Символ «#» указывает на пользовательские комментарии. Все, написанное после него, и до конца строки будет проигнорировано.

Если есть LIG начинающаяся строка и нет «-og file» флага, то лиганд не будет выведен. Если есть флаг «-og file», но нет LIG начинающейся строки, будет ошибка.

**REN** - переименовать резидью. После через пробел вводятся 2 названия резидью, сначала исходное, потом то, на которое нужно переименовать. Необходим флаг «-renres».

Пример:

* **REN** HEME HEM
* **REN** CL CLA

**CHA** - добавить заряд атому в резидью. Сначала вводится название резидью, потом атома, потом целое число.

* **CHA** ZN2 ZN 2

«-bf file» - флаг, задающий файл базы сортировки. Для всех атомов должны быть указаны полные и правильные названия атомов. Формат файла базы: несколько строк, состоящих из названия резидью, пробела, перечисления всех названий атомов в нужном порядке, через точку.

Пример:

* ALA N.HN.CA.HA.CB.HB1.HB2.HB3.C.O
* ASP N.HN.CA.HA.CB.HB1.HB2.CG.OD1.OD2.C.O
* ASN N.HN.CA.HA.CB.HB1.HB2.CG.OD1.ND2.HD21.HD22.C.O

«-st» - флаг, указывающий, что нужно провести сортировку по шаблону, указанному после флага «-bf file».

«-rr [number]» - флаг, указывающий парсеру на то, что нужно задать новые номера всем резидью, начиная с 1 или с нужного числа, если оно указано после флага.

«-rs [number]» - флаг, указывающий парсеру на то, что нужно задать новые номера всем атомам, начиная с 1 или с нужного числа, если оно указано после флага.

«-ct» - флаг, переводящий парсер в подрежим разрезки структур на цепи. Создается для каждой цепи по файлу, содержащему атомы цепи. Если указан линагд в HETFILe, для каждой цепи еще и создается файл лиганда, а из цепи он удаляется. Имена файлов будут взяты из параметров «-os file» «-og file», перед расширением будет дописан суффикс \_№, где № - название соответствующей цепи. Расшиерение, если указано, будет сохранено, в противном случае будет «.pdb».

 «-fasta» - создание файла в фаста формате с последовательностью белка.

Обновление базы сортировки резидью.

Обновление на основе структуры белка. Необходимы «-s file», «-bf file» и «-ub» параметры. Если в заданном файле уже были записаны некоторые резидью, то они будут обновлены (Если в них меньше атомов, перезаписаны. Если в них столько же атомов, но другой порядок, то останется предыдущая версия). Если в структуре встречаются резидью с разным количеством и/или порядком атомов, то будет сохранен самый длинный первый встретившийся порядок. Файл будет перезаписан.

Сбор результатов докинга.

Необходим входной файл, содержащий полные/относительные пути к DLG файлам флаг «-dp file», содержащим результаты докинга, проведенного с помощью программы Autodock4 версии 2.5.1. В выходной файл записываются следующие параметры результатов: название файла, ядро, время счета, параметры докинга, количество кластеров. Для лучшего кластера и для второго, если есть:количество структур, рмсд лучшей от исходного положения лиганда(не объективно для лигандов, которые не были вырезаны из структуры), энергия лучшей находки в кластере. Результаты суммируются в виде псевдоэксель таблицы – текстового файла, который интерпретируется программой excel, как таблица).

«-ndoe» не писать результаты, содержащие ошибки в файл с результатами (обычно относится к незавершенным DLG файлам).

Создание файла лиганда из результатов докинга для лучших находок из кластеров.

Флаг «-dlglig file1 [file2 number]», принимает от 1 до 3х аргументов: входной DLG файл, выходной файл (если не задан, выходной файл будет создан в папке, содержащей DLG файл, тем же именем и расширением PDB, число – количество кластеров, из которых нужно создать файлы (если не задано, будет создан всего 1 файл для лучшего кластера, иначе будет создано от 1 до максимума из количества кластеров и этого числа файлов лигандов для лучшей находки из каждого кластера). Если указано число > 1, то каждый файл будет создан с суффиксом перед расширением «\_№», где № - номер кластера.

### Корректирование лигандов.

Необходим флаг «-ligcor file1 file2 file3». Был добавлен для исправления файлов лигандов после обработки программой MOPAC. Требует ввода 3х файлов: файл шаблона (исходный), файл с новыми координатами (результат работы MOPAC), выходной файл. Файлы должны содержать 1 лиганд в поле HETEATM (в некоторых случаях на вход может быть подан файл, содержащий такое же количество атомов в том же порядке, но разбитых на несколько резидью или не в поле хетатм), иметь одинаковое количество атомов и их порядок. Координаты в первом файле будут заменены на координаты, взятые из второго файла. А результат будет записан в третий файл.

### Вызов описания параметров.

 Флаг «-h», другие флаги не работают.

Параметры, применимые почти для любого запуска.

«-nc [number]» не выводить заряд. Некоторые программы не понимают или неправильно работают с зарядами, которые по стандарту PDB должны быть после типа атома (79-80). Если число не указано или указанное равно 0, то заряд будет записан только тем атомам, у которых он был в исходном файле. Если задано число 1, заряд не будет записан ни одному атому. Если этот параметр не задан, то всем атомам, не имевшим заряд, будет дописан заряд 0+.

«-nw» без предупреждений. Некритичные ошибки и предупреждения не будут выводиться на экран. Если этот параметр не указан, то уведомления о всех предупреждениях и ошибках, позволяющих продолжить работу программы, будут выводиться в поток ошибок синим цветом.

«-silent» тихий режим. Лог программы не будет выводиться на экран. Если этот параметр не задан, лог выводится на экран белым цветом.

«-suf string» суффикс. Добавляет к имени всех выходных файлов указанный суффикс в конце или перед расширением при его наличии.

«-bp» возможные ошибки программы OPEN BABEL. Исправляются некоторые известные ошибки вывода PDB файлов программы OPEN BABEL 2.3.2. Исправляемые ошибки: При добавлении водородов, они все добавляются в поле ATOM. Если у атома 4 коннекта, то будет добавлена строка, содержащая последний номер атома на предыдущей строке (5й лишний коннект). Если параметр задан, соответствующие предупреждающие сообщения не будут выводиться на экран.

«-vmd» изменение формата PDB, используемого программой NAMD, чтобы соответствовать общепринятому. Все сегменты, которые начинаются с P, становятся цепями белка с идентификаторами цепи соответствующими номеру (1 - A, 2 - B, 3 - C). Сегменты, названия которых начинаются на O, перемещаются в поле HETATM, но номер игнорируется, а идентификатор цепи присваивается по тому, атомы какой цепи из поля ATOM ближе всего. Остальные сегменты игнорируются.

«-ac» обрабатывать альтернативные координаты. Если параметр задан, в выходные фалы будут записаны только первые альтернативные положения (имеющие минимальную в лексикографическом порядке букву). В выходном файле они будут помечены, как основные положения, а остальные будут удалены.

# Параметры программы

**Таблица 1.** Список параметров

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Параметр** | **Required** | **Описание** | **Пример** | **FR** |  |
| -s | file | входной файл в PDB формате | -s 1eqh.pdb | \* | 1 |
| -os | file | выходной файл в PDB | -os rec.pdb | \* | 1 |
| -og | file | выходной файл лиганда в PDB формате | -og lig.pdb | \* | 1 |
| -bf | file | файл базы резидью с порядком атомов для сортировки | -bf basefile.dat | \* |  |
|  | Формат файла базы для сортировки: название резидью, пробел, список атомов в правильном порядке, разделенных точками. Пример:ALA N.HN.CA.HA.CB.HB1.HB2.HB3.C.O ASP N.HN.CA.HA.CB.HB1.HB2.CG.OD1.OD2.C.OASN N.HN.CA.HA.CB.HB1.HB2.CG.OD1.ND2.HD21.HD22.C.O |  |  |
| -st |  | сортировать порядок атомов в резидью по шаблону, -bf параметр необходим | -st |  |  |
| -ub |  | Флаг обновления базы для сортировки, параметр –bf обязателен, результат будет ДОБАВЛЕН в этот файл. Если в нем были уже записаны добавляемые резидью, то они будут обновлены | -ub |  | 1,2 |
| -hf | file | Файл параметров HETATM. Специальный файл, в который записываются указания программе, какой лиганд вырезать, какие лиганды считаются стандартными, какие лиганды удалить. | -hf hetfile.het | \* | 1 |
|  | Формат HETATM файла: #-комментарии, после них все игнорируется LIG, пробел, название лиганда, пробел, название цепи # указание вырезать лиганд из названной цепи в выходной файл лиганда. Если вместо названия лиганда написано FIND, указание программе найти нестандартный лиганд в указанной цепи. Если цепь не указана, будет проведен поиск по каждой цепи, если нужно вырезать только из одной, будет вырезан из первой найденной. Строка, начинающаяся с LIG должна быть только одна, в противном случае будет использована последняя DEL, пробел, название лиганда, пробел, название цепи # указание удалить лиганд из названной цепи. Если вместо цепи указан символ !, лиганд будет удален из всех цепей STD, пробел, название лиганда, пробел # указание программе считать лиганд стандартным Пример HETATM файла LIG FLP A # Вырезать лиганд FLP из цепи A в отдельный файл DEL HOH ! #Удалить воду из всех цепей DEL MAN A #Удалить лиганд MAN из цепи A REN HEME HEM #Переименовать HEME в HEM  CHA ZN2 ZN 2 #Задать заряд 2 атому ZN в резидью ZN2 |  |  |
| -rr | [num] | Флаг перенумеровки резидью. Если после него написано число, новая нумерация всех резидью начнется с этого числа для КАЖДОГО выходного файла | -rr 1 |  |  |
| -rs | [num] | Флаг перенумеровки серийных номеров атомов. Если после него написано число, новая нумерация всех атомов начнется с этого числа для КАЖДОГО выходного файла | -rs 1 |  |  |
| -ct |  | Режим разрезки на цепи. Каждая цепь будет выделена в отдельный файл. Если есть HETFILE, и в нем есть строка с нестандартным лигандом (наличие указания цепи игнорируется), то при наличии этого лиганда в соответствующей цепи, для нее будет сделан файл, содержащий этот лиганд, а из цепи он будет удален | -ct |  | 1 |
| -nc |  | Не писать зарядов. 0 – Заряд не будет записан для тех атомов, для которых его не было в исходной структуре. 1 – Заряды не будут записаны никаким атомам. Если флаг не задан, заряды будут записаны всем атомам. (0+, если не было в исходной структуре) | -nc 0 |  |  |
| -silent |  | Тихий режим. Комментарии не будут выводиться на экран. | -silent |  |  |
| -nw |  | Режим без предостережений. Некритичные ошибки не будут выводиться на экран (поток ошибок) | -nw |  |  |
| -dp |  | Режим суммирования запусков докинга. Необходим файл, содержащий в каждой строке путь до лог файла Autodock4. И выходной файл, в который будут записаны суммированные результаты (имеет текстовый формат, но может интерпретироваться excel, как таблица). | -db |  | 3 |
| -ndoe |  | Не выводить строки с ошибками в файл с результатами. | -ndoe |  | 3 |
| -pb |  | Возможные ошибки программы OpenBabel. На некоторые ошибки не будут выведены предупреждающие сообщения. Например: повтор серийного номера на новой строке, если их 4. Добавленные водороды в поле ATOM вместо HETATM будут перенесены в HETEATM | -pb |  |  |
| -vmd |  | Возможные ошибки программы VMD. После обработки VMD все лиганды попадают в поле ATOM, но имеют сегмент, начинающийся с O. По этому признаку они преобразуются в HETATM, добавляется цепь | -vmd |  |  |
| -clr |  | Переименовать резидью CL в CLA. По сути работает как аналогичная строчка в хетфайле | -clr |  |  |
| -zn |  | Переименовать резидью ZN в ZN2. По сути работает как аналогичная строчка в хетфайле | -zn |  |  |
| -ac |  | Обработать альтернативные координаты. 0 – не обрабатывать. 1 – удалить все, кроме первых в лексикографическом порядке. | -ac 1 |  | 1 |
| -ligcor | file1 file2 file3 | – Режим корректирования лигандов. Требует ввода 3х файлов: файл шаблон, файл с новыми координатами, выходной файл. Файлы должны содержать 1 лиганд в поле HETEATM, иметь одинаковое количество атомов и их порядок. Координаты в первом файле будут заменены на координаты, взятые из второго файла. А результаты записаны в третий. | -ligcor ligand.pdb ligand\_mop.pdb out\_lig.pdb | \* | 4 |
| -dlglig | file1 [file2 num] /[num] /[file2] | Создание лигандов из лог файла Autodock4. Требует указание файла dlg. Будет создан файл с лигандом из первого кластера в той же папке, что и dlg файл с тем же именем и расширением pdb. Если еще указано число, то будет создано столько файлов, сколько есть кластеров, если число больше их количества, или «число» кластеров в той же папке, что и dlg файл суффиксом \_№кластера и расширением pdb. Если еще будет задан выходной файл, то все файлы будут созданы с его именем \_№кластера (если задано число) и расширением pdb. | -dlglig res.dlg 2 out.pdb | \* | 5 |
| -suf | suffix | Добавить введенную строку как суффикс перед расширением для каждого выходного файла. | -suf \_ed |  |  |
| -evp | file1 file2 file3 | Анализ результатов докинга из DLG файлов из списка (file2), добавление к ним информации об изоформе (file1), создание ROC диаграмм. | -evp about.txt dlglist.txt img.jpg |  |  |
| -chc |  | Всем резидью из поля HETATM, у которых поле цепь пустое или там цепь, в которой нет ни одного резидью из поля ATOM, перемещаются в поле ATOM в цепь, атомы которой ближе всего к этому резидью | -chc |  |  |
| -frep |  | Принудительно перезаписывать файлы |  |  |  |
| -rotate | num1 [num2] [num3] | Поворот всех атомов структуры на угол по осям. Сначала X, потом Y, потом Z. 3 числа задают повороты по этим осям. Углы задаются от 0 до 360 | -rotate 90 90 90 |  |  |
| -renres |  | Переименовать резидью, как указано в хетфайле |  |  |  |
| -fasta | file1 | Создать файл в фаста формате с последовательностью белка | -fasta file.fasta |  |  |
| -igseg |  | Игнорировать поле сегмента | -igseg |  |  |
| -nooq |  | Не выводить поле oqqupancy | -nooq |  |  |
| -notf |  | Не выводить температурный фактор | -notf |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  | To be continued… |

Для тех параметров, для которых в четвертом столбике звездочка, необходим ввод параметров

Параметры, для которых в пятом столбце разные цифры, несовместимы друг с другом. Запускаются только отдельно друг от друга, а имеющие одинаковые цифры, обычно связаны друг с другом.

# Примеры разных задач

1. Пересортировка серийных номеров атомов с номера N
	* pdbparser -s file.pdb -os outfile.pdb -rs N
2. Пересортировка номеров резидью с номера M в каждой цепи
	* pdbparser -s file.pdb -os outfile.pdb -rr M
3. Создать файл в фаста формате с последовательностью белка
	* pdbparser -s file.pdb -fasta outfasta.fasta
4. Пересортировать атомы в каждом резидью в порядке, как указано в файле базы для сортировки
	* pdbparser -s file.pdb -os outfile.pdb -bf file -st
5. Создать/обновить файл базы для сортировки на основе pdb файла
	* pdbparser -s file.pdb -os outfile.pdb -ub
6. Создать файл в CSV формате с результатами докинга в программе Autodock4.2
	* pdbparser -dp list.txt results.txt.xls
	* pdbparser -dprls list.txt results.txt.xls
7. Создать файл в CSV формате с результатами докинга в программе Autodock Vina
	* pdbparser -vina\_dp list.txt results.txt.xls
	* pdbparser -vina\_dprls list.txt results.txt.xls
8. Исправить ошибки программы Open Babel
	* pdbparser -vina\_dp list.txt results.txt.xls
9. Удалить альтернативные координаты.
	* pdbparser -s file.pdb -os outfile.pdb -ac
10. Разделить структуру на мономеры
	* pdbparser -s file.pdb -os outfile.pdb -ct
11. Вырезать лиганд FLP из структуры в отдельный файл
	* + ------hetfile.txt-------
		+ LIG FLP A
	* pdbparser -s file.pdb -os outfile.pdb -og lig.pdb -hf hetfile.txt
12. Найти и вырезать нестандартный лиганд из структуры в отдельный файл
	* + ------hetfile.txt-------
		+ LIG FIND A
		+ STD HOH
		+ STD NAG
		+ STD BOG
		+ STD TIP3
	* pdbparser -s file.pdb -os outfile.pdb -og lig.pdb -hf hetfile.txt
13. Разделить структуру на мономеры, вырезав лиганды в отдельные файлы
	* pdbparser -s file.pdb -os outfile.pdb -og lig.pdb -hf hetfile.txt -ct
14. Удалить некоторые резидью из поля HETATM
	* + ------hetfile.txt-------
		+ DEL NAG
		+ DEL HOH
	* pdbparser -s file.pdb -os outfile.pdb -hf hetfile.txt
15. Переименовать некоторыерезидью
	* + ------hetfile.txt-------
		+ REN HEME HEM
		+ REN HSD HIS
		+ REN CL CLA
	* pdbparser -s file.pdb -os outfile.pdb -hf hetfile.txt
16. Повернуть атомы в структуре на 90 градусов по оси X
	* pdbparser -s file.pdb -os outfile.pdb -rotate 90

PDBPARSER – программа для работы с файлами в формате PDB3.1 в консольном режиме (без графического интерфейса). Основная цель парсера – облегчить работу пользователя с файлами во время подготовки докинга и после него.