МЕЖДУНАРОДНЫЙ УНИВЕРСИТЕТ АСТАНА

Высшая школа естественных наук

ОСНОВЫ БИОИНФОРМАТИКИ
(Практикум)

Поиск последовательностей

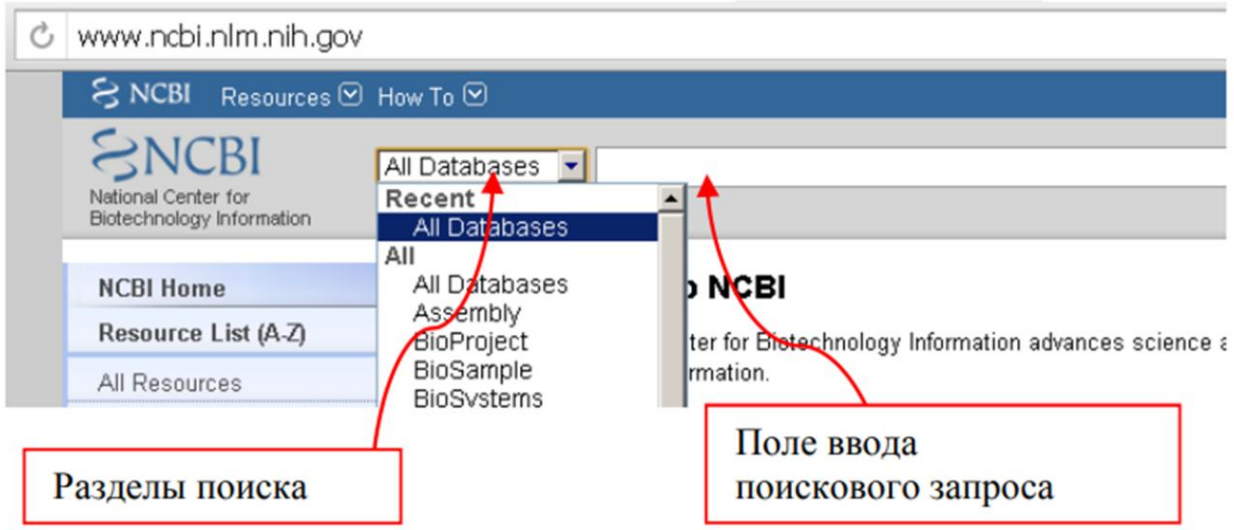
ОРАЗОВ АЙДЫН ЕРҒАЛИҰЛЫ
PhD, и.о. ассоциированного профессора Высшей школы естественных наук международного университета Астана,
Заведующий научно-исследовательской лаборатории изучения окружающей среды «NatureLaB»

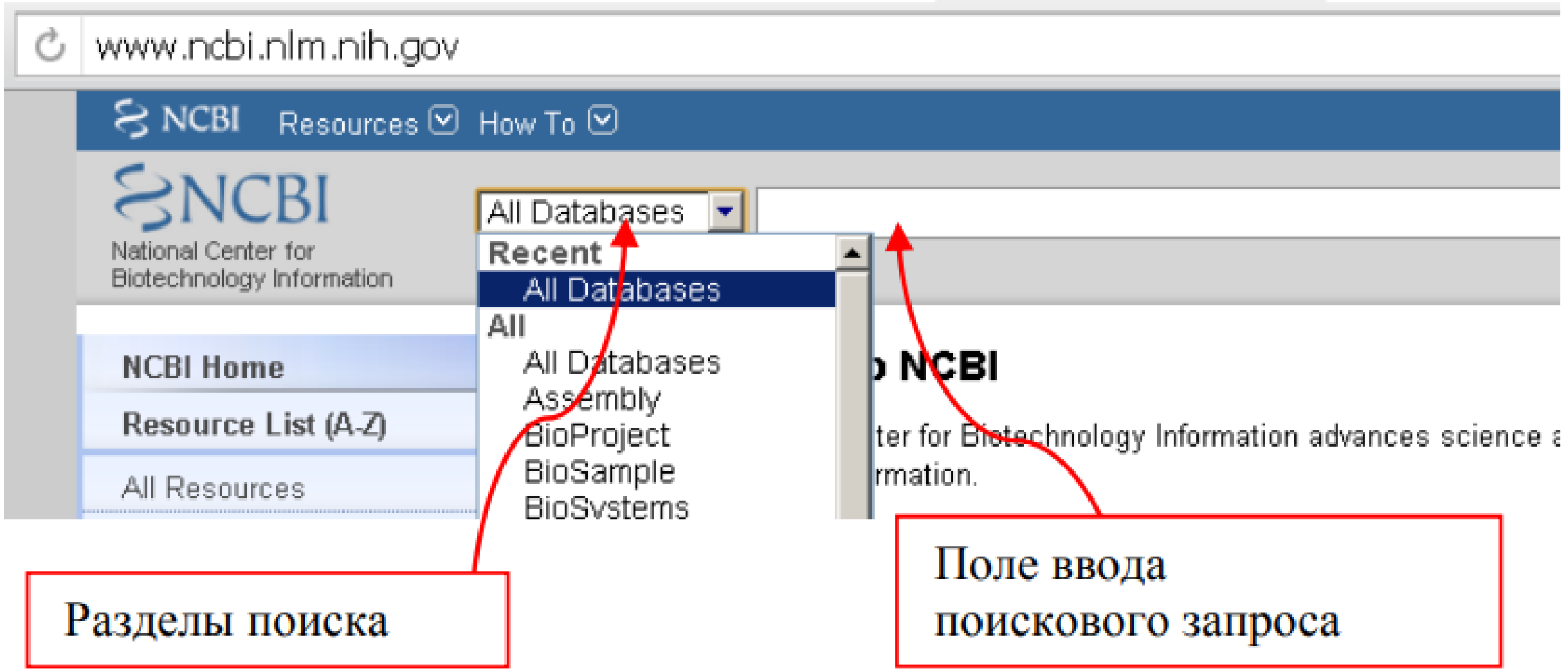
АСТАНА, 2025

Данный ресурс является гигантским сборником разнообразнейших инструментов анализа биологической информации, порой способных заменить десятки специализированных программ – вы убедитесь в этом в ходе дальнейшей работы. Тем не менее, на главной странице вы видите единственную строку поиска – поскольку первым шагом к работе с биологическими данными является поиск информации в специальных базах и банках данных. Рассматриваемый в этой работе тип поиска представляет собой т.н. поиск по ключевым словам: система будет искать записи баз данных, содержащие введенные нами слова.

Одним из центральных ресурсов поиска информации в биологии является портал Национального центра биотехнологической информации США:

NCBI

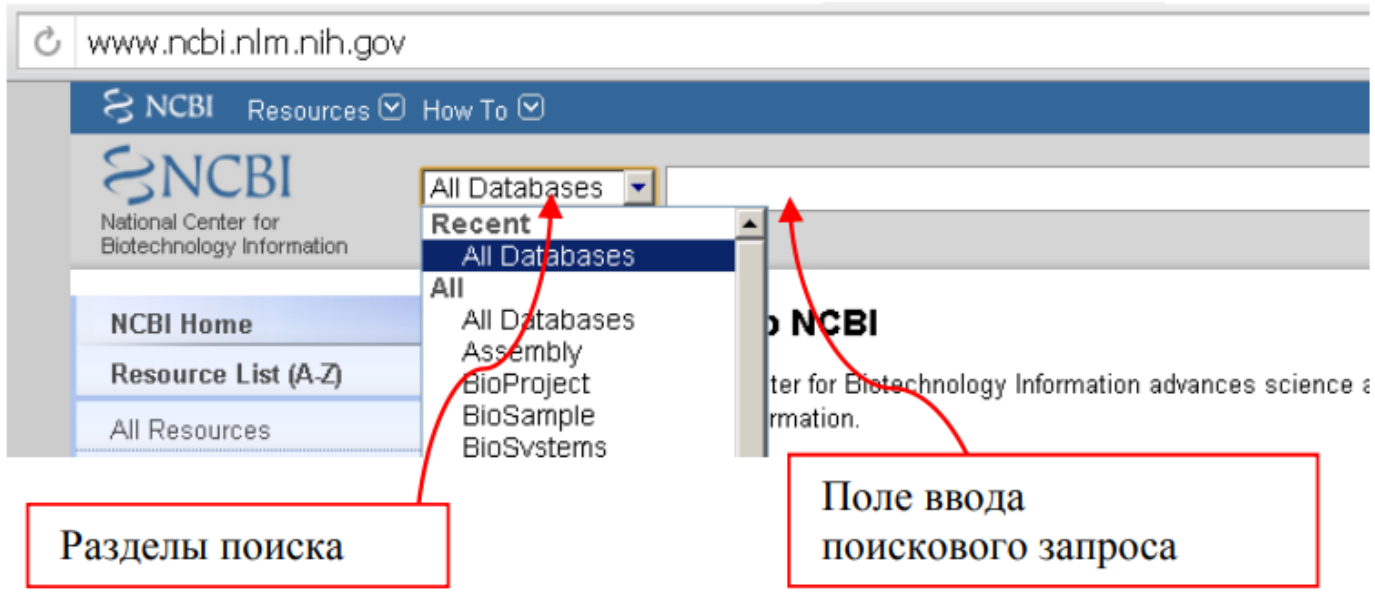




Перейдите в веб-браузере по указанному адресу
ncbi.nlm.nih.gov

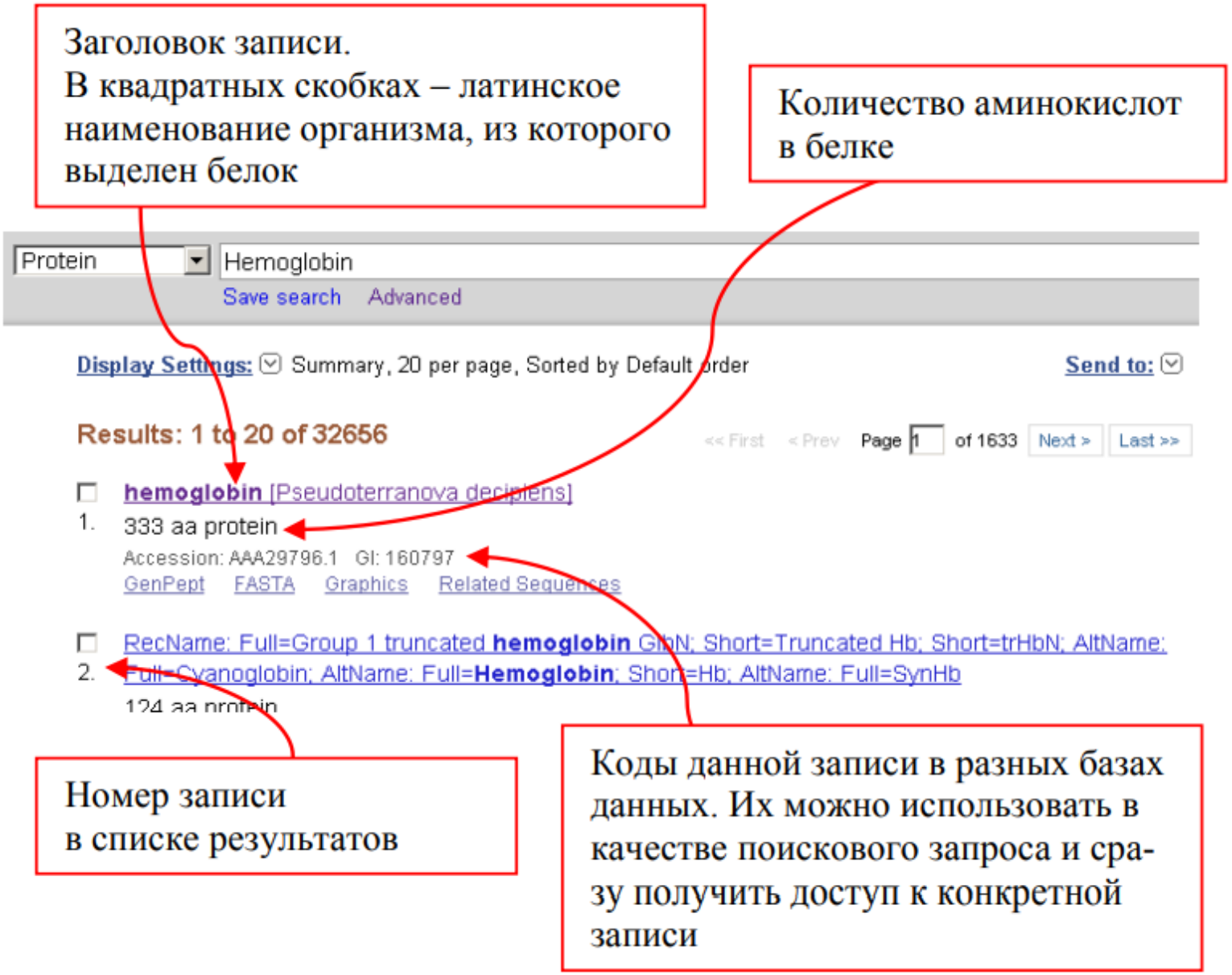
Выпадающий список позволяет выбрать базу данных нужного типа (например,
Protein – для поиска последовательностей белков,
Nucleotide – нуклеиновых кислот,
Structure – пространственных структур биомолекул,
Pubmed – поиск научных публикаций).

Nucleotide



Поиск по ключевым словам

Найдите записи, относящиеся к белку гемоглобину. Для этого в выпадающем списке выберите раздел Protein, впишите в поле ввода Hemoglobin и нажмите Search или Enter. В результате на экран будут выведены ссылки на все записи, содержащие данное слово.



Запись состоит из полей, обозначенных заглавными буквами, которые содержат информацию о последовательности. Ниже приведено описание некоторых полей.

LOCUS	Данное поле содержит несколько элементов, описанных ниже.
Locus Name	Например, **AAA29796**. Уникальный идентификатор (код) записи или последовательности.
Sequence Length	Например, **333 aa**. Длина последовательности.
GenBank Division	Раздел банка данных **GenBank**, к которому относится последовательность. Например, **INV**. Разделы отражают происхождение последовательности из определенного организма, технологию определения последовательности или другие особенности:

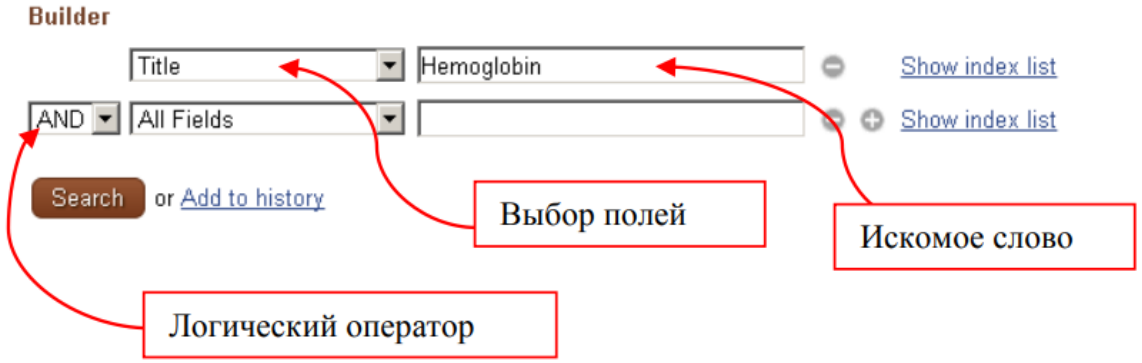
PRI – primate sequences	Приматы
ROD – rodent sequences	Грызуны
MAM – other mammalian sequences	Другие млекопитающие
VRT – other vertebrate sequences	Другие позвоночные
INV – invertebrate sequences	Беспозвоночные
PLN – plant, fungal, and algal sequences	Растения, грибы, водоросли
BCT – bacterial sequences	Бактерии
VRL – viral sequences	Вирусы
PHG – bacteriophage sequences	Бактериофаги
SYN – synthetic sequences	Синтетическая последовательность
UNA – unannotated sequences	Неописанная последовательность
PAT – patent sequences	Последовательность, включенная в патент
ENV – environmental sampling sequences	Последовательность из образцов окружающей среды

Modification Date	Дата последнего изменения записи. Например, **26-APR-1993**.
DEFINITION	Краткое описание последовательности.
ACCESSION	Уникальный идентификатор записи. Формат зависит от типа последовательности. Например, **AAA29796**.
VERSION	Уникальный идентификатор последовательности. Например, **AAA29796.1**. Изменяется при внесении поправок в существующую запись с сохранением **Accesion**.
GI - "GenInfo Identifier"	Еще один идентификатор последовательности. Изменяется при любой правке последовательности, поэтому входит в раздел **Version**.
KEYWORDS	Ключевые слова, описывающие последовательность. При отсутствии таковых ставится точка.
SOURCE	Неформализованное название организма – источника последовательности.
ORGANISM	Формализованное наименование организма и его таксономия.
REFERENCE	Публикации, содержащие информацию из данной записи.
AUTHORS	Список авторов.
TITLE	Название работы.
JOURNAL	Название журнала.
FEATURES	Описание генов, продуктов их экспрессии, а также участков последовательности.
ORIGIN	Последовательность в формате "GenBank".

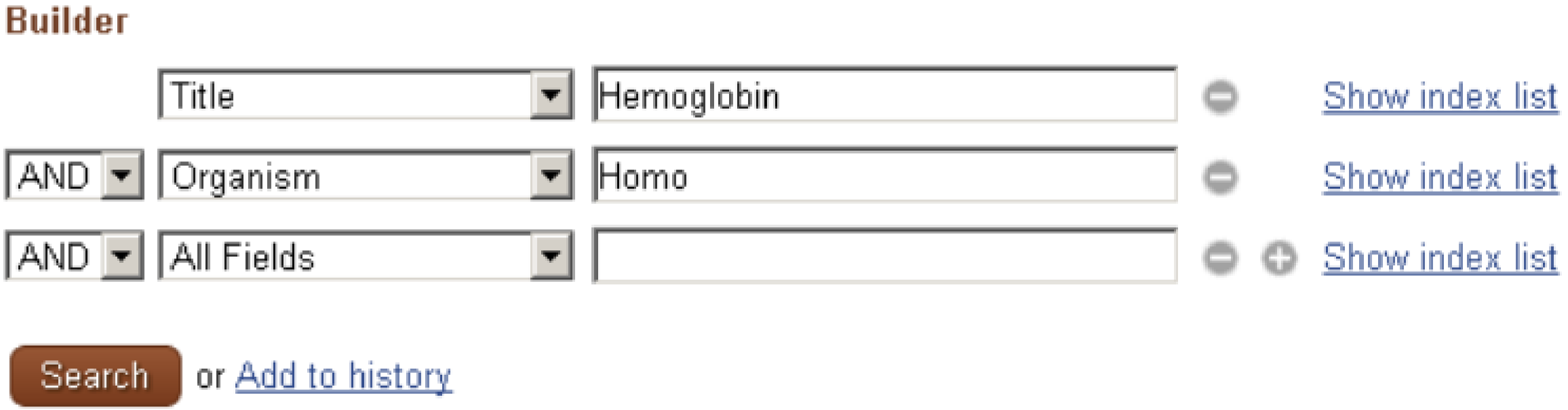
Расширенный режим поиска

В данном случае системой были найдены записи, в которых слово Hemoglobin встречается в любом из полей. При таком подходе часто бывает сложно найти нужную последовательность среди множества нерелевантных результатов. Более гибкие возможности поиска предоставляет расширенный режим, в который можно перейти с помощью ссылки Advanced под полем ввода (откройте ее в новой вкладке).

В расширенном режиме можно указывать, в каких полях должны встречаться искомые слова. Пусть нам нужно найти записи, содержащие слово Hemoglobin исключительно в заголовке. В первом выпадающем списке выберите Title, в поле ввода рядом с ним впишите Hemoglobin. Нажмите Search или Enter, чтобы увидеть результаты поиска. Сравните количество найденных записей в этом случае и в предыдущем.



В расширенном режиме выполните поиск записей, соответствующих гемоглобину человека. Используйте дополнительную строку поиска и логический оператор AND.



Перед каждой дополнительной строкой можно установить один из трех логических операторов:

При использовании…	Будут найдены записи, содержащие слова…
AND	строго из обоих строк – из данной и из предыдущей
OR	хотя бы из одной строки – данной или предыдущей
NOT	любые, кроме указанных в данной строке

Самостоятельно выполните поиск записей, соответствующих гемоглобину всех эукариот, кроме человека. (Попробуйте самостоятельно догадаться, где подсмотреть правильное написание слова «эукариот».) Корректно выполненный поиск выдаст около 10000 результатов.

Те же результаты можно получить в обычном режиме, используя специальный синтаксис поискового запроса (его можно увидеть в строке поиска при отображении результатов запроса), например, такой:

(Hemoglobin[Title]) AND Homo[Organism]

Названия полей указываются в квадратных скобках сразу после ключевых слов, круглые скобки, наряду с операторами, служат для логической группировки элементов. Использование данного синтаксиса часто бывает проще и быстрее, чем «мышеориентированного» интерфейса, а иногда и вовсе незаменимо, как можно убедиться, выполнив следующее задание.

С использованием поискового синтаксиса составьте запрос для поиска белка глобина (слово globin в названии записи) человека и быка (Bos). Обратите внимание на правильное использование логических операторов.

Затем проделайте то же самое в расширенном режиме (по ссылке Advanced). Сопоставьте результаты. Найдите причину различий. Корректно выполненный поиск выдаст около 200 результатов.

Ссылки:

https://www.ncbi.nlm.nih.gov/

https://ru.wikipedia.org/wiki/%D0%9F%D0%BE%D0%BB%D1%8B%D0%BD%D1%8C

https://ru.wikipedia.org/wiki/%D0%9E%D0%B4%D1%83%D0%B2%D0%B0%D0%BD%D1%87%D0%B8%D0%BA

https://xn--80abvyzg.xn--p1ai/%D0%B1%D0%B8%D0%BE%D0%B8%D0%BD%D1%84%D0%BE%D1%80%D0%BC%D0%B0%D1%82%D0%B8%D0%BA%D0%B0/#11