Searching Entrez using Boolean logic

What is a "Boolean"?
This tutorial introduces the concept of Boolean logic as it is used in general search engines as well as it is implemented at Swiss-Prot and NCBI's Entrez.

Boolean logic takes its name from British mathematician George Boole (1815-1864), who wrote about a system of logic designed to produce better search results by formulating precise queries. He called it the "calculus of thought." From his writings, we have derived Boolean logic and its operators: AND, OR, and NOT, which we use to link words and phrases for more precise queries.

BOOLEAN "AND"

The Boolean AND narrows your search by retrieving only documents that contain every one of the keywords you enter. The more terms you enter, the narrower your search becomes.

EXAMPLE: truth AND justice
EXAMPLE: truth AND justice AND ethics AND congress

BOOLEAN "OR"

The Boolean OR expands your search by returning documents in which either or both keywords appear. Since the OR operator is usually used for keywords that are similar or synonymous, the more keywords you enter, the more documents you will retrieve.

EXAMPLE: college OR university
EXAMPLE: college OR university OR campus OR higher education

BOOLEAN "NOT" / "AND NOT"

The Boolean NOT or AND NOT limits your search by returning only your first keyword but not the second, even if the first word appears in that document, too.

EXAMPLE: bulimia AND NOT anorexia
EXAMPLE: cowboys AND NOT dallas

NESTING -- WITH BOOLEAN OPERATORS

Nesting, i.e., using parentheses, is an effective way to combine several search statements into one search statement. Use parentheses to separate keywords when you are using more than one operator and three or more keywords.

EXAMPLE: (scotch OR bourbon) AND NOT (beer OR wine)
(For best results, always enclose OR statements in parentheses.)

BOOLEAN LOGIC REDUX

Boolean logic is not always simple or easy. Different search engines handle Boolean operators differently. For example, some accept NOT, while one accepts ANDNOT as one word, others AND NOT as two words. Some require the operators to be typed in capital letters while others do not.
Some search engines use drop-down menu options to spell out the Boolean logic in short phrases. For example, "All of the words" or "Must contain" equates to AND; "Any of the words" or "Should contain" equates to OR; and "Must not contain" equates to NOT.

IMPLIED BOOLEAN OPERATORS

Implied Boolean operators use the plus (+) and minus (-) symbols in place of the full Boolean operators, AND and NOT. Typing a (+) or (-) sign in front of a word will force the inclusion or exclusion of that word in the search statement.

EXAMPLE: +bulimia -anorexia

Similarly, putting double quotation marks (" ") around two or more words will force them to be searched as a phrase in that exact order.

EXAMPLE: "green tea"

While full Boolean operators are usually accepted only in the advanced search option of search engines, implied Boolean operators are accepted in the basic search options of most search engines.

PROXIMITY OPERATORS

Proximity, or positional, operators (NEAR, ADJ, SAME, FBY) are not really part of Boolean logic, but they serve a similar function in formulating search statements.

Not all search engines accept proximity operators, but a few accept NEAR in their advanced search option. The NEAR operator allows you to search for terms situated within a specified distance of each other in any order. The closer they are, the higher the document appears in the results list. Using NEAR, when possible, in place of the Boolean AND usually returns more relevant results.

EXAMPLE: phylogeny NEAR ontogeny
EXAMPLE: de Vere NEAR Shakespeare

Even fewer search engines accept ADJ (adjacent to). ADJ works as a phrase except that the two terms, which must appear adjacent to each other in the webpage, can appear in any order.

EXAMPLE: Ernest ADJ Hemingway
EXAMPLE: endangered ADJ species
returns both Ernest Hemingway and Hemingway Ernest; endangered species and species endangered.

Other proximity operators, such as SAME (keywords found in the same field) and FBY (followed by), are used as advanced searching techniques in library and other specialized databases that contain bibliographic citations or references to journal articles, but are not yet employed by search engines.

The major biological databases use only a subset of the Boolean logic operators.

If you want to see how Google implements Boolean logic, click on the Advanced search link just to the right of the text box on the Google homepage.
Writing Advanced Search Statements in Entrez

Complex search statements can be written and executed directly from the query box of any of the five databases, as long as you follow some simple rules and use the correct syntax.

A search is performed by:
- specifying the search terms
- the fields to be searched
- the boolean operations to perform on the term.

Use the following syntax:

```
term [field] OPERATOR term [field]
```

The **term** is the search term. There can be more than one search term.

The **field(s)** are the [Search Fields and Qualifiers](#) (listed below). These represent fields within the Genbank record that can be searched. Imagine that you want to search for “protein kinase c”. If you use “protein kinase c” as the search term and do not restrict it to a particular field, it will search the entire Genbank record for any instance of “protein kinase c.” This means it can return records that pertain to genes that interact with protein kinase c or that for some reason have the “protein kinase c” somewhere in its record. If you restrict it to the title or gene name field, you will be more likely to find the records that represent the protein kinase c gene.

The **OPERATOR(s)** are the [Boolean Operators](#). Remember that Boolean operators are normally processed left to right. If you wish part of your Boolean expression to be processed out of order, enclose it in parentheses. They are always in **ALL CAPS**.

**Example:** Find all human nucleotide sequences with D-loop annotations.

In the Nucleotide database use the following expression -

```
D-loop[FKEY] AND human[ORGN]
```

**Example:** Find all human protein sequences with lengths between 50 and 60 amino acids that were entered into the database during 1999.

In the Protein database use the following expression -

```
```

**Example:** Find drosophila population studies published in the Journal of Molecular Evolution.

In the PopSet database use the following expression -

```
j mol evol[JOUR] AND drosophila[ORGN]
```
Boolean Operators used in Entrez are:

**AND:** To AND two search terms together instructs Entrez to find all documents that contain BOTH terms.

**OR:** To OR two search terms together instructs Entrez to find all documents that contain EITHER term.

**NOT:** To NOT two search terms together instructs Entrez to find all documents that contain search term 1 BUT NOT search term 2.

The Entrez search rules and syntax for using Boolean operators are:

1. Boolean operators, AND, OR, NOT must be entered in UPPERCASE (e.g., promoters OR response elements).
2. Entrez processes all Boolean operators in a left-to-right sequence. The order in which Entrez processes a search statement can be changed by enclosing individual concepts in parentheses. The terms inside the parentheses are processed first as a unit and then incorporated into the overall strategy. For example, the search statement: g1p3 AND (response element OR promoter) is processed by Entrez by ORing the terms response element OR promoter first and then ANDing the resulting set of documents with g1p3.
3. Click on the Details button to see how Entrez translated and executed your search strategy.

**Further reading:**


This is a pretty good introduction to the basics of internet web searching strategy.

<table>
<thead>
<tr>
<th>Search Field</th>
<th>Definition</th>
<th>Qualifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accession</td>
<td>Contains the unique accession number of the sequence or record, assigned to the nucleotide, protein, structure, genome record, or PopSet by a sequence database builder. The Structure database accession index contains the PDB IDs but not the MMDB IDs.</td>
<td>[ACCN]</td>
</tr>
<tr>
<td>All Fields</td>
<td>Contains all terms from all searchable database fields in the database.</td>
<td>[ALL]</td>
</tr>
<tr>
<td>Author Name</td>
<td>Contains all authors from all references in the database records. The format is last name space first initial(s), without punctuation (e.g., marley jf).</td>
<td>[AUTH]</td>
</tr>
<tr>
<td>EC/RN</td>
<td>Number assigned by the Enzyme Commission or Chemical</td>
<td>[ECNO]</td>
</tr>
<tr>
<td>Number</td>
<td>Abstract Service (CAS) to designate a particular enzyme or chemical, respectively.</td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>Filter</td>
<td>Contains predetermined or filtered subsets of the various databases. These subsets or filters are created by grouping records that are commonly linked to other Entrez databases or within the same database. For example, the PopSet database Filter index includes PopSet all, PopSet medline, PopSet nucleotide, and PopSet protein. The PopSet medline filter includes all PopSet records with links to PubMed; the PopSet nucleotide filter includes all PopSet records with links to the nucleotide database; and, the PopSet protein filter includes all PopSet records with links to the protein database. The PopSet all filter includes all PopSet records. The Nucleotide database Filter index contains a great deal more filters because the database records are linked to numerous external links. For more information see <a href="http://www.ncbi.nlm.nih.gov/collab/FT/index.html">Link Out</a>.</td>
<td></td>
</tr>
<tr>
<td>Gene Name</td>
<td>Contains the standard and common names of genes found in the database records. This field is not available in Structure database.</td>
<td></td>
</tr>
<tr>
<td>Issue</td>
<td>Contains the issue number of the journal in which the data were published.</td>
<td></td>
</tr>
<tr>
<td>Journal Name</td>
<td>Contains the name of the journal in which the data were published. Journal names are indexed in the database in abbreviated form (e.g., J Biol Chem). Journals are also indexed by their by ISSN. Browse the index if you do not know the ISSN or are not sure how a particular journal name is abbreviated.</td>
<td></td>
</tr>
<tr>
<td>Keyword</td>
<td>Contains special index terms from the controlled vocabularies associated with the GenBank, EMBL, DDBJ, SWISS-Prot, PIR, PRF, or PDB databases. Browse the Keyword indexes of the individual databases to become familiar with these vocabularies. A Keyword index is not available in the Structure database.</td>
<td></td>
</tr>
<tr>
<td>Modification Date</td>
<td>Contains the date that the most recent modification to that record is indexed in Entrez, in the format YYYY/MM/DD (e.g., 1999/08/05). A year alone, (e.g., 1999) will retrieve all records modified for that year; a year and month (e.g., 1999/03) retrieves all records modified for that month that are indexed in Entrez.</td>
<td></td>
</tr>
<tr>
<td>Molecular Weight</td>
<td>Molecular weight of a protein, in Daltons (Da), calculated by the method described in the <a href="http://www.ncbi.nlm.nih.gov/collab/FT/index.html">Searching by Molecular Weight</a> section.</td>
<td></td>
</tr>
</tbody>
</table>
of the Entrez help document. Note that molecular weight must be entered as a fixed 6 digit field, filled with leading zeros (not letter O), e.g., 002002 [MOI.WT]

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organism</td>
<td>Contains the scientific and common names for the organisms associated with protein and nucleotide sequences.</td>
<td>[ORGN]</td>
</tr>
<tr>
<td>Page Number</td>
<td>Contains the number of the first journal page of the article in which the data were published.</td>
<td>[PAGE]</td>
</tr>
<tr>
<td>Primary Accession</td>
<td>Contains the primary accession number of the sequence or record, assigned to the nucleotide, protein, structure, genome record, or PopSet by a sequence database builder. A Primary Accession index is not available in the Structure database.</td>
<td>[PACC]</td>
</tr>
<tr>
<td>Properties</td>
<td>Contains properties of the nucleotide or protein sequence. For example, the Nucleotide database's Properties index includes molecule types, publication status, molecule locations, and GenBank divisions. A Properties index is not available in the Structure database.</td>
<td>[PROP]</td>
</tr>
<tr>
<td>Protein Name</td>
<td>Contains the standard names of proteins found in database records. Common names may not be indexed in this field so it is best to also consider All Fields or Text Words. A Protein Name index is not available in the Structure database.</td>
<td>[PROT]</td>
</tr>
<tr>
<td>Publication Date</td>
<td>Contains the date that records are released into Entrez, in the format YYYY/MM/DD (e.g., 1999/08/05). It is the date the entry first appeared in GenBank explicitly indexed in Entrez. A year alone, (e.g., 1999) will retrieve all records for that year; a year and month (e.g., 1999/03) will retrieve all records released into GenBank for that month.</td>
<td>[PDAT]</td>
</tr>
<tr>
<td>SeqID String</td>
<td>Contains the special string identifier, similar to a FASTA identifier, for a given sequence. A SeqID String index is not available in the Structure database.</td>
<td>[SQID]</td>
</tr>
<tr>
<td>Sequence Length</td>
<td>Contains the total length of the sequence. Sequence Length indexes are not available in the Structure or PopSet databases.</td>
<td>[SLEN]</td>
</tr>
<tr>
<td>Substance Name</td>
<td>Contains the names of any chemicals associated with this record from the CAS registry and the MEDLINE Name of Substance field. Substance Name indexes are not available in the Genome or PopSet databases.</td>
<td>[SUBS]</td>
</tr>
<tr>
<td>Text Word</td>
<td>Contains all of the &quot;free text&quot; associated with a record.</td>
<td>[WORD]</td>
</tr>
<tr>
<td>Title Word</td>
<td>Includes only those words found in the definition line of a record. The definition line summarizes the biology of the sequence and is carefully constructed by database staff. A standard definition line will include the organism, product name, gene symbol, molecule</td>
<td>[TITL]</td>
</tr>
<tr>
<td><strong>Uid</strong></td>
<td>Contains the Medline unique identifier for records that contain published references that are linked to PubMed. The Uid index is not browsable.</td>
<td>[UID]</td>
</tr>
<tr>
<td><strong>Volume</strong></td>
<td>Contains the volume number of the journal in which the data were published.</td>
<td>[VOL]</td>
</tr>
</tbody>
</table>