Previous Contents Next

Issues in Science and Technology Librarianship Fall 2007
DOI:10.5062/F47P8WBF

URLs in this document have been updated. Links enclosed in **{curly brackets}** have been changed. If a replacement link was located, the new URL was added and the link is active; if a new site could not be identified, the broken link was removed.

Refereed article

Entrez and BLAST: Precision and Recall in Searches of NCBI Databases

Tina O'Grady
Science Librarian
University of New Orleans
New Orleans, Louisiana
cogrady@uno.edu

Abstract

This project analyzes the results of searches for genes and proteins in the NCBI databases Gene, RefSeq RNA and RefSeq Protein. Corresponding searches were performed using the search programs Entrez and BLAST, and search recall and precision were calculated. The findings demonstrate the different types of result sets that can be expected from using different search programs and settings. Also, some unexpected results indicate that the default search settings are not optimal for all searches; an important aspect of searching which information professionals should remember and communicate to researchers.

Introduction

As the number of genomes, genes and proteins sequenced by researchers continues to grow, the ability to effectively access and use sequence data grows in importance for biologists and other researchers. Almost every biological research laboratory makes use of sequence data in some way, and skill in searching the resources available for these data is becoming a necessary component of biomedical information literacy.

The {ACRL Information Literacy Standards for Science and Engineering/Technology} refer to the wide variety of formats of scientific and technical information. These standards do not apply to bibliographic information alone; librarians are becoming increasingly aware of bioinformatics resources and the needs of researchers using them. Though the format and search methods of biological sequence databases are different from traditional library databases, these information resources carry a great deal of

potential for library involvement. Geer (2006) notes the difficulty end-users have in keeping up with rapid change in these tools and their lack of awareness of available search features. Many librarians are seeking an appropriate role, and some libraries are moving forward with initiatives such as bioinformatics specialists, workshops, lectures, consultation services, web portals and software provided through the library (Messersmith 2006; Chattopadhyay 2006; Alpi 2001).

The basis for many bioinformatics tools, and the data driving the current information boom in biological science, is sequence data -- the sequences of nucleic acids (DNA and RNA) and proteins. DNA, or deoxyribonucleic acid, is the genetic material for virtually all organisms and consists of long chains made from four molecular bases -- adenine, guanine, cytosine, and thymine, abbreviated in written sequences as A, G, C and T. RNA (ribonucleic acid), an intermediate between DNA and proteins, is also made up of four bases: A, G, C, and uracil (U) instead of thymine. Proteins consist of sequences of amino acid molecules that fold into often complicated shapes, based on the type and order of the amino acids in the sequence. Genes are sections of the DNA sequence that encode a protein; the DNA in these sections is first "transcribed" to an RNA molecule, then the RNA is "translated" to a protein molecule. The DNA and RNA sequences form a "code" that indicate which amino acids are incorporated into the protein. For example, the DNA sequence ATGACTGACTTC would be used as a template to create an RNA molecule with the sequence AUGACUGACUUC. The RNA molecule would then be used as a template to create a protein with the amino acid sequence Methionine-Threonine-Aspartic Acid-Phenylalanine. Thus, the sequences of DNA, RNA and protein are related to one another for each gene, based on this "code". More information about this "central dogma" of molecular biology can be found on the {NCBI's Science Primer}.

Groups of genes and proteins can be classified into families: slight changes in the DNA sequence produce different but related genes and proteins. The human genome has many gene and protein families. Comparison of gene or protein sequences can illustrate relationships between individual genes, as well as between individual organisms or species. Scientists use sequence information to classify genes and organisms, elucidate relationships between them, and identify genes or proteins. This has important implications for new discoveries in science and medicine.

The major provider of sequence data in the United States is the National Center for Biotechnology Information (NCBI), a division of the National Library of Medicine. NCBI makes available over 30 databases and other resources on its web site, http://www.ncbi.nlm.nih.gov/, and offers various ways to search for and navigate through the information. These databases, ranging from the bibliographic database PubMed to the DNA sequence database GenBank, can be searched simultaneously and contain cross-database links between related records.

The two tools for searching NCBI databases are Entrez, a keyword-searching program, and BLAST (Basic Local Alignment Search Tool), a program that uses a nucleic acid or protein sequence as a query and retrieves sequences that are similar to the query. Several variations of the BLAST program have been designed to search in different ways and return different result sets. Both Entrez and BLAST offer various search limits and settings that can be modified by the user. The default settings are meant to give the "best all-around results" for a wide variety of searches (McGinnis 2001), but many modifications can be made to searches to obtain different result sets.

This project uses sample searches based on sequences from a gene family in the human genome to test the performance of default settings when searching for genes in that family, and to analyze and compare search results obtained from various Entrez and BLAST searches. The results illustrate how the choice of search program influences results, and evaluates the performance of these search programs for a particular family of human genes.

Three related datasets were searched in this project: the Gene database, the RefSeq RNA collection and the RefSeq Protein collection. The Gene database, searched via the Entrez interface, contains records for individual genes from fully-sequenced genomes (Maglott 2007). Records include gene names and symbols, unique GeneID numbers, and links to nucleic acid and protein sequences in the RefSeq RNA and Protein collections. The RefSeq collections are curated sets of nucleic acid and protein sequences that can be searched directly using BLAST or indirectly using Entrez via Gene or other databases (Pruitt 2007). RefSeq sequences, unlike other sequences in GenBank, are reviewed and checked for problems, and the collection is non-redundant (GenBank as a whole is often redundant, containing several records for the same gene submitted by different researchers). Links between records in all three datasets allow easy navigation between related records: for example a user searching with Entrez can retrieve a Gene database record that links to the RefSeq RNA record for the gene sequence, and the RefSeq Protein record for the corresponding protein sequence. The links facilitate information discovery and allow comparison of retrieval when searching different but related datasets.

To compare retrieval based on different search methods, a large family of genes in the human genome was selected: the Intermediate Filament family. Searches of the Gene, RefSeq RNA and RefSeq Protein databases were performed, and the results analyzed to determine search precision (here defined as the number of Intermediate Filament records returned divided by the total number of records returned) and recall (here defined as the number of Intermediate Filament records returned divided by the total number of Intermediate Filament records in the database). Because this is an extensively studied and well-characterized gene family, a complete list of genes in the human Intermediate Filament family could be compiled and assumed to represent the complete list of human Intermediate Filament records in the database for the purposes of measuring recall.

The Intermediate Filament family consists of 70 genes that encode proteins of several types. These include structural proteins such as the keratins in hair and nails, and proteins found in the lens of the eye. Some Intermediate Filament genes have transcript variants, meaning that a single gene (DNA sequence) encodes for two or more slightly different RNA sequences or proteins. Some Intermediate Filament genes also have pseudogenes, which are almost exact duplicates of the gene that do not encode proteins. The presence of transcript variants and pseudogenes has implications for searching. For the purposes of precision and recall measurement, transcript variants were included if they have RefSeq records, but pseudogenes were excluded as there is no consensus on how many Intermediate Filament pseudogenes exist (many RefSeq records for pseudogenes are labeled "model"; that is, hypothetical), so a complete and reliable list could not be obtained or compiled.

Methods

A list of genes in the Intermediate Filament family (Table 1) was produced by comparing and compiling lists in the published literature (Hesse 2001; Schweizer 2006). This involved combining incomplete or partially developed lists and controlling for synonyms (for example, the protein synemin is also called desmuslin). Once the list was compiled, records for each of the 70 genes were retrieved from the Gene database to confirm their presence there, obtain the official names and symbols used by NCBI, and find the number of transcript variants. This complete list of Intermediate Filament genes and proteins was then used to calculate the precision and recall of the searches.

Table 1. Characterized Human Intermediate Filaments

Keratins			
Type I (Acidic)		**Type II (Basic)**	
Keratin 9	Keratin 19	Keratin 1	Keratin 71

Keratin 10	Keratin 20	Keratin 2	Keratin 72
Keratin 12	Keratin 23	Keratin 3	Keratin 73
Keratin 13	Keratin 24	Keratin 4	Keratin 74
Keratin 14	Keratin 25	Keratin 5	Keratin 75
Keratin 15	Keratin 26	Keratin 6A	Keratin 76
Keratin 16	Keratin 27	Keratin 6B	Keratin 77
Keratin 17	Keratin 28	Keratin 6C	Keratin 78
Keratin 18		Keratin 7	Keratin 79
		Keratin 8	Keratin 80
Type I Hair		**Type II Hair**	
Keratin 31	Keratin 36	Keratin 81	
Keratin 32	Keratin 37	Keratin 82	
Keratin 33A	Keratin 38	Keratin 83	
Keratin 33B	Keratin 39	Keratin 84	
Keratin 34	Keratin 40	Keratin 85	
Keratin 35		Keratin 86	
Other Intermediate Filaments			
Type III (Vimentin-like)		**Type IV (Neurofilaments)**	
Vimentin		alpha-Internexin	NF-L
Desmin		Syncoilin	NF-M
GFAP		Nestin	NF-H
Peripherin		Synemin	
Type V (Lamins)		**Eye Proteins**	
Lamin A/C		Filensin	
Lamin B1		Phakinin	
Lamin B2			

Keratin list from Schweizer (2006)
Others from Hesse 2001

Entrez Searches

The first series of searches was performed in the Gene database using the Entrez keyword search interface. All searches were limited to the species *Homo sapiens* to allow analysis of the results (complete lists of Intermediate Filament proteins have not been established for all species represented in the Gene database, making recall determination for other species difficult or impossible). First, a search was performed for the phrase "intermediate filament". No truncation, limits (aside from species) or other advanced search techniques were used in order to perform the search as most users would (70% of Entrez searches are simple, unmodified queries -- Geer 2006). Results were tallied and the recall and precision of the search determined. Next, a search was performed for the only Intermediate Filament subtype of a large enough size to be useful -- the Keratins. The term "keratin" was used to query the Gene database via Entrez, again limiting the search to *Homo sapiens* but using no other advanced search techniques.

BLAST Searches

The second part of the project involved searching the RefSeq RNA and Protein sequence databases by inputting sequence queries via the BLAST program (Altschul 1997). Three genes were selected from the Intermediate Filament family to act as queries: Keratin 18, Lamin A and Syncoilin 1 (Table 2). Keratin 18 is known to have a large number of pseudogenes (Schweizer 2006) and Lamin A is known to have 3

transcript variants. Syncoilin 1 was chosen based on results of searches with the other two genes.

Table 2. Query Sequences

	GenBank Accession Number	
	Nucleic Acid	**Protein**
Keratin 18	NM_000224	NP_000215.1
Lamin A	NM_170707.1	NP_733821
Syncoilin 1	NM_030786.1	NP_110413.1

Five BLAST searches were performed using each gene sequence or its corresponding protein sequence as a query. Nucleic acid sequences were used as queries with the BLAST programs megaBLAST, discontiguous megaBLAST, and blastn. Protein sequences were used as queries with the BLAST program blastp, using two different scoring matrices (BLOSUM62 and PAM30). Results were limited to *Homo sapiens*, but no other changes were made to the default search settings, again under the assumption that the average user does not normally use advanced search techniques, and following McGinnis' (2001) assertion that the default settings "should give the best all-round results." All results from BLAST searches are ranked by a score called the e-value, a number that functions as a predictor of the expected false-positive rate for a given query. Smaller e-values represent a statistically lower chance of a false-positive result, and thus a better sequence match (NCBI, *The Statistics of Sequence Similarity Scores*). For this project, results with an e-value of less than 1 were tabulated to calculate recall and precision for the Intermediate Filament family.

Results and Discussion

Precision and recall measurements are calculated based on the list of Intermediate Filaments in Table 1; that is, named, characterized protein-coding genes in the Intermediate Filament family. Pseudogenes and genes that haven't been officially named (usually hypothetical genes based on rough genome scans), though they may be of interest to a researcher, are excluded in calculations because of the difficulties involved in determining the total number of records for these genes in the database. "Relevance" of results is always a difficult determination in measures of precision and recall and in this case if pseudogenes and hypothetical genes are considered "relevant", the precision of these searches can be assumed to be higher.

Entrez Searches

The Entrez search for "intermediate filament" returned 150 hits from the Gene database (Table 3). This search found all of the genes in the list of known Intermediate Filament genes – a recall rate of 100%. Of the 150 total hits, 70 were Intermediate Filament genes on the list in Table 1: a precision rate of 47%. The remaining 80 hits were of several types. Nineteen hits were for discontinued records, presumably for genes that had been predicted by scans of the human genome but ultimately discarded. An observant user could easily remove these discontinued records from the display by selecting the "Current Only" tab at the top of the results list. Other hits were unrelated genes that, for various reasons, contained the term "intermediate filament" somewhere in their annotation (the proteins may interact with Intermediate Filaments, for example).

Table 3. Entrez Search Results

	Search Term

Retrieved Record Type	Intermediate Filament	Keratin
Intermediate Filament	70	55
Discontinued Records	19	443
Intermediate Filament Pseudogene	2	42
Hypothetical Intermediate Filament	0	131
Unrelated Gene	59	169
Total	**150**	**840**
Recall	**100%**	**100%**
Precision	**47%**	**7%**

The more specific Entrez search for "keratin" returned a much larger answer set of 840 hits. Discontinued records were a much larger problem with this search, forming 53% of the answer set. Only 55 (7%) of the results were verified Intermediate Filament genes. Also included were Keratin pseudogenes, hypothetical genes designated as "Keratin-like" or "similar to Keratin" and a large number of unrelated genes. Recall for the named Keratin genes was 100%, and one non-keratin Intermediate Filament gene was found for an overall Intermediate Filament recall of 79%.

For both of these searches the recall rate was perfect, retrieving records for 100% of the Intermediate Filament or Keratin genes, respectively. This is a benefit of searching for genes in a relatively well-studied and well-characterized family in the human genome; annotation for these records in the Gene database tends to be fairly thorough and standard. Lesser known genes may not be annotated as well, and vocabulary for more recently discovered genes and families is often less standardized. This is evident to some extent even in these searches; the "keratin" search returned 173 hypothetical Keratin genes and pseudogenes that were not retrieved by the "intermediate filament" search (Table 3), though as Keratins they are members of the Intermediate Filament family. If these particular hypothetical genes are further studied and confirmed as genes, their annotation will be expanded and they should be retrieved in a more general "intermediate filament" search. It is important to remember when searching the Gene database that annotation can vary, especially for genes that are currently hypothetical, and using synonyms or related keywords in a query can yield different results.

One difficulty in interpreting the results of these searches was that the relatively low precision rates could lead to a lot of "wading" through irrelevant records in the result sets. Entrez does offer features to help improve precision such as limits and Boolean search capabilities. Some of these features are quite user-friendly (e.g. the "Current Only" tab in the results list) and might be less intimidating for end-users reluctant to use Boolean operators or field codes. For information professionals and other savvy users however, the Entrez interface offers many features that can be exploited to improve searches.

BLAST Searches

BLAST searches based on the three selected queries showed great variation in their result sets (Table 4a-c). For example, a megaBLAST search for Keratin 18 yielded a recall of only 4% and precision of only 3%. On the other hand, using Keratin 18 as a query for a blastp search with the PAM30 matrix resulted in a recall of 92% and precision of 78%. That searches using the same query can lead to such disparate result sets indicates the level of thought and familiarity with the database needed for efficient searching with BLAST.

Table 4a. BLAST Query Results – Keratin 18

	Nucleic Acid BLAST	**Protein BLAST**

Subtype	megaBLAST	discontiguous megaBLAST	blastn	BLOSUM62 matrix	PAM30 matrix
Type I Recall	16%	74%	84%	95%	95%
Type I Hair Recall	0	55%	91%	100%	100%
Type II Recall	0	0	24%	100%	100%
Type II Hair Recall	0	0	100%	100%	100%
Type III Recall	0	0	0	100%	100%
Type IV Recall	0	0	0	88%	50%
Type V Recall	0	0	0	100%	100%
Eye Protein Recall	0	0	0	100%	50%
Total Hits	**106**	**129**	**151**	**162**	**88**
Total Recall	**4%**	**26%**	**49%**	**97%**	**92%**
Precision	**3%**	**15%**	**25%**	**46%**	**78%**

Table 4b. BLAST Query Results – Lamin A

	Nucleic Acid BLAST			**Protein BLAST**	
Subtype	megaBLAST	discontiguous megaBLAST	blastn	BLOSUM62 matrix	PAM30 matrix
Type I Recall	0	0	0	95%	95%
Type I Hair Recall	0	0	0	100%	100%
Type II Recall	0	57%	95%	100%	100%
Type II Hair Recall	0	17%	17%	100%	100%
Type III Recall	0	75%	100%	100%	100%
Type IV Recall	0	38%	0	88%	75%
Type V Recall	60%	100%	100%	100%	100%
Eye Protein Recall	0	0	0	100%	0%
Total Hits	**3**	**48**	**96**	**327**	**92**
Total Recall	**4%**	**32%**	**39%**	**97%**	**93%**
Precision	**100%**	**50%**	**31%**	**23%**	**77%**

Table 4c. Syncoilin 1 Percent Recall

	Nucleic Acid BLAST			**Protein BLAST**	

Subtype	megaBLAST	discontiguous megaBLAST	blastn	BLOSUM62 matrix	PAM30 matrix
Type I Recall	0	0	0	84%	0
Type I Hair Recall	0	0	0	36%	0
Type II Recall	0	0	0	62%	0
Type II Hair Recall	0	0	0	83%	0
Type III Recall	0	0	0	75%	0
Type IV Recall	13%	13%	13%	25%	25%
Type V Recall	0	0	0	0	0
Eye Protein Recall	0	0	0	100%	0
Total Hits	**4**	**4**	**7**	**76**	**13**
Total Recall	**1%**	**1%**	**1%**	**59%**	**3%**
Precision	**25%**	**25%**	**14%**	**59%**	**15%**

All BLAST searches work in approximately the same way, by breaking the query sequence into short series of letters or "words", searching for matching words in all the sequences in the database, and then determining how far the region of matching sequence extends. Different forms and settings of the algorithm can be used to vary parameters, such as whether the initial query is a nucleic acid or protein sequence, the size of the initial word and how similarity is scored (Altschul 1997; McGinnis 2004). These different BLAST programs are meant to be used for different purposes, and the NCBI provides guides to the selection and use of the different programs in their BLAST Program Selection Guide at www.ncbi.nlm.nih.gov/blast/producttable.shtml.

Nucleic Acid BLAST Searches

Three BLAST programs that work with nucleic acid sequences are megaBLAST, discontiguous megaBLAST and blastn. These programs accept nucleic acid sequences as queries and were used in this project to search the RefSeq RNA database. According to the BLAST Program Selection Guide, megaBLAST is best used for identifying unknown sequences. MegaBLAST searches for long stretches of similarity with very few differences between the sequences – it finds only the very closest matches to the query sequence. This is useful if a researcher has experimentally obtained sequence data from an unknown gene; the sequence can be input into megaBLAST, which will ideally find the exact sequence or a closely related one in the database, identified and annotated. This worked very well in the search for Lamin A: megaBLAST returned only the three transcript variants of Lamin A (Table 4b). With Keratin 18, however, megaBLAST returned 106 results including the two Keratin 18 transcript variants, Keratin 14, 92 uncharacterized model RefSeq sequences, and 11 results that were not Intermediate Filaments at all (Table 4a). Although the Keratin 18 results were listed first, this large result set does cloud sequence identification. Keratin 18 is known to have a large number of pseudogenes (Schweizer 2006) and many of the megaBLAST results are records for hypothetical Keratin 18 pseudogenes. Other genes with a large number of pseudogenes may present similar problems with megaBLAST searches. Interestingly, the megaBLAST search for the Syncoilin 1 sequence returned the Syncoilin 1 record and 3 other hits, all related to Retinoblastoma Binding Protein 4, and not in the Intermediate Filament family.

In contrast to megaBLAST, blastn and discontiguous megaBLAST are designed to search for nucleic acid sequences that are related, but not necessarily nearly identical to the query sequence. Blastn works as a typical BLAST search, breaking the query into words and searching for matches. It requires less sequence similarity than megaBLAST, and should find more distantly related (or less similar) sequences. Discontiguous megaBLAST is designed to be more sensitive than blastn, finding sequences that can show a greater degree of variation. It does this by not simply searching for matching words, but allowing the words to be "discontiguous"; that is, searching for the letters of a word, in order but not necessarily immediately adjacent to one another, within a certain window of sequence length (BLAST Program Selection Guide). Because they are designed to return more widely varying sequences, both blastn and discontiguous megaBLAST should have greater recall than megaBLAST, but perhaps lower precision. Using the default settings, the more sensitive discontiguous megaBLAST should have a higher recall than blastn and possibly lower precision.

In fact, blastn and discontiguous megaBLAST did usually provide results with higher recall than megaBLAST did. Unexpectedly, for each of the three query sequences blastn returned a higher number of results than discontiguous megaBLAST; for Keratin 18 and Lamin A blastn also returned more relevant results (i.e. had a higher recall - Table 4a-c). It is possible that these unanticipated results stem from searching within the genome of a single species, as related genes in other species are presumably more likely to have the type of variations that discontiguous megaBLAST is designed to cope with. The blastn search also had higher precision than discontiguous megaBLAST with the Keratin 18 query, but lower precision with the Lamin A and Syncoilin 1 queries. The number of results returned ranged from 4 (Syncoilin 1/discontiguous megaBLAST) to 151 (Keratin 18/blastn). These are not unwieldy numbers even when precision is low, and so in this case the blastn searches, which had recall rates equal to or better than the discontiguous megaBLAST searches, were probably the more useful searches despite sometimes having lower precision. Based on information in the BLAST Program Selection Guide this is surprising, and one would do well to remember that NCBI defaults and recommendations are meant to give generally good results in a wide range of searches (McGinnis 2004) but may not be ideal for a user's specific search needs.

Protein BLAST Searches

When searching for more distantly related sequences, BLAST searches using proteins (blastp) should generally be more useful than those using nucleic acids (BLAST Program Selection Guide 2007). Searching the RefSeq Protein database using the protein sequences of Keratin 18, Lamin A and Syncoilin 1 instead of their nucleic acid sequences should increase recall but also possibly decrease precision, as more distantly related sequences are retrieved. BLAST scoring of the similarity between protein sequences, one of the parameters used in calculating e-values, is based on scoring matrices. BLAST supports the use of several scoring matrices, designated BLOSUM (BLOcks of Amino Acid SUbstitution Matrix) or PAM (Point Accepted Mutation) and followed by a number indicating the ideal range of sequence similarity between query and results. The default matrix is BLOSUM62, which has been shown to outperform other matrices for most protein groups (Henikoff & Henikoff 1992). It was expected that all BLAST searches with protein sequences would have greater recall than the BLAST searches using nucleic acid sequences, and that blastp searches with BLOSUM62 would have better recall than blastp searches with PAM30.

In each case, blastp searches did have greater recall than nucleic acid BLAST searches for the corresponding sequence (Table 4a-c). For both Lamin A and Keratin 18 the recall rates were above 90% for blastp searches (compared to 4-49% using nucleic acid sequences), and the recall rate for the Syncoilin 1 search increased as well. BLOSUM62 did yield higher recall than the PAM30 matrix for each query, but the

difference tended to be fairly small (e.g. 97% recall for Lamin A/BLOSUM62 and 93% for Lamin A/PAM30). For both Keratin 18 and Lamin A though, the PAM30 matrix returned results with much higher precision than BLOSUM62. This has strong implications for usability. For example, the BLOSUM62/Lamin A search had the highest recall rate at 97%; however, it returned a total of 327 hits, with a precision rate of only 23%. In contrast, the same search using the PAM30 matrix had slightly lower recall at 93%, but returned only 92 results, with a precision rate of 77%. Depending on the needs of the searcher, the PAM30 matrix could be considered the better performer. Very similar results were seen with the blastn searches for Keratin 18. For Syncoilin 1, however, BLOSUM62 was clearly the better matrix, returning results with much higher recall and precision.

For almost every BLAST search both recall and precision were lower when using Syncoilin 1 as the query than when using Keratin 18 or Lamin A (Table 4c). This was not entirely unexpected; Syncoilin 1 was chosen as a query when it was noticed that it was never retrieved in BLAST searches using Keratin 18 or Lamin A as the query. By looking only at results from Syncoilin 1 BLAST searches, one might conclude that Syncoilin 1 was not an Intermediate Filament at all – results such as Retinoblastoma Binding Protein 4 appeared much more consistently than Intermediate Filament records. This indicates that official gene and protein families are based on more than strict sequence similarity, and in this case there is substantial variation in sequence among members of the same family.

Conclusions

Entrez Searching vs. BLAST Searching

Entrez and BLAST provide different ways of searching and accessing the same information (in this case RefSeq sequence records and their corresponding entries in the Gene database). Depending on the search technique, the results retrieved can be very different. Usually the two approaches are used by researchers in very different situations -- a searcher seeking information about a particular gene or gene family will likely use the Entrez interface, while a searcher in possession of sequence data from an unknown gene, or a searcher looking for sequence relationships to a known gene, can search with BLAST. The many links provided between database records allow searchers to find a record in the Gene database by following links from a sequence record returned by a BLAST search, and vice versa.

The results of this project illustrate some of the advantages and challenges of both interfaces. Entrez uses the familiar format of keyword searching, but without a consistent controlled vocabulary, record annotation can vary and end-users may need to be reminded to use a variety of keywords. One advantage of BLAST sequence searches is that they do not have the inherent synonym and nomenclature problems of keyword searches, but different BLAST programs and settings can lead to very different results. Knowledge of the functions of different BLAST programs and the needs of the researcher (i.e., is the researcher trying to identify which gene a particular sequence comes from? To obtain a large number of related sequences for analysis?) is important for effective searching with BLAST.

Largely, the two interfaces returned similar result sets in this project, including many Intermediate Filaments and members of other apparently similar families such as the Myosins. There were also some key differences. For example, the Syncoilin 1 record was retrieved in an Entrez search for "intermediate filament" but not in any of the BLAST searches using Keratin 18 or Lamin A as queries. Conversely, the hypothetical protein FLJ40504 is a predicted protein that is consistently returned with low e-values in BLAST searches using Keratin 18; if its existence is confirmed it is likely a keratin, but its Gene record is not annotated as such and it is not retrieved by an Entrez search

for "keratin". Generally, Entrez queries are more likely to return better annotated records (that may or may not have a sequence similarity to the query), while BLAST searches can return results that have similar sequences but are not confirmed members of a gene family.

BLAST programs

The various BLAST programs performed roughly as expected, with a few surprising results. Generally, blastn and discontiguous megaBLAST results had greater recall and lower precision than megaBLAST, which was expected as these tools were designed to return records for nucleic acid sequences with more variation from the query sequence. Likewise, BLAST searches using protein sequence queries had higher recall than BLAST searches using nucleic acid sequence queries. If a researcher has both nucleic acid and protein sequence data, nucleic acid BLAST searches can be used when seeking other closely related sequences, and protein BLAST searches can be used when seeking more distantly related sequences. Surprisingly, blastn results had a higher recall than discontiguous megaBLAST, and the PAM30 matrix could be interpreted as a better matrix than BLOSUM62 for most searches in this project. These unexpected results indicate that, as with searching bibliographic databases, search strategy sometimes needs to be adjusted based on the specific search, and using multiple searches can produce a more comprehensive list of results.

The National Center for Biotechnology Information provides access to a large and valuable set of databases with powerful accompanying search tools. Default settings and recommendations tend to work well but a strong knowledge of database content, search tools and input data characteristics can increase the efficiency and effectiveness of searches. Knowledge or assumptions about the search query (e.g. which gene family it might belong to, and how that family is related to other families) can inform the choice of keywords for Entrez searches and which BLAST program to use. Searchers can more closely tailor their search strategies to their needs with a good understanding of both the query data and the search tools. Librarians mediating searches of these resources may not be able to gain intimate knowledge of the data researchers are using as queries, but can make patrons aware that different queries may behave differently in the system, and be prepared to offer suggestions for varying search parameters to improve results.

Acknowledgements

The research for this article originated as a project for the University of Maryland class CMBG 688O -- Bioinformatics and Molecular Systematics. The author would like to thank the instructor of that class, Dr. Charles Delwiche, as well as Paul St-Pierre, Reference Librarian (Science) at Tulane University, for his comments on earlier drafts.

References

***ACRL Information Literacy Standards for Science and Engineering/Technology*.** 2006. [Online]. Available: {http://www.ala.org/ala/mgrps/divs/acrl/standards/infolitscitech.cfm} [Accessed August 6, 2007].

Alpi, K. 2003. Bioinformatics training by librarians and for librarians: developing the skills needed to support molecular biology and clinical genetics information instruction. *Issues in Science and Technology Librarianship* 37 [Online]. Available: http://www.istl.org/03-spring/article1.html [Accessed November 1, 2007].

Altschul, S. F., et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* 25(17): 3389-3402.

BLAST Program Selection Guide. 2007. [Online]. Available: http://www.ncbi.nlm.nih.gov/blast/producttable.shtml [Accessed November 1, 2007].

Chattopadhyay, A., et al. 2006. Design and implementation of a library-based information service in molecular biology and genetics at the University of Pittsburgh. *Journal of the Medical Library Association* 94(3):307-313.

Geer, R. C. 2006. Broad issues to consider for library involvement in bioinformatics. *Journal of the Medical Library Association* 94(3):286-298.

Henikoff, S. and Henikoff, J. G. 1992. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America* 89:10915-10919.

Hesse, M., Magin, T. M. and Weber, K. 2001. Genes for intermediate filament proteins and the draft sequence for the human genome: novel keratin genes and a surprisingly high number of pseudogenes related to keratin genes 8 and 18. *Journal of Cell Science* 114: 2569-2575.

Maglott, D., et al. 2007. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research* 35(Database Issue):D5-D12.

Messersmith, D. J. et al. 2006. A Web-based assessment of bioinformatics end-user support services at US universities. *Journal of the Medical Library Association* 94(3):299-305.

McGinnis, S. and Madden, T. L. 2004. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Research* 32(Web Server Issue):W20-W25.

NCBI. *A Science primer*. Retreived June 5, 2007 from http://www.ncbi.nlm.nih.gov/About/primer/genetics_genome.html.
NCBI. *The Statistics of Sequence Similarity Scores*. Retrieved Aug. 6, 2007 from http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html.

Pruitt, K. D., Tatusova, T. and Maglott, D. R. 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research* 35(Database Issue):D61-D65.

Schweizer, J. *et al*. 2006. New consensus nomenclature for mammalian keratins. *Journal of Cell Biology* 174(2): 169-174.

Previous Contents Next