Searching the World-Wide-Web using nucleotide and peptide sequences

Natarajan Ganesan[1], Nicholas F. Bennett, Bala Kalyanasundaram, Mahe Velauthapillai, and Richard Squier

Department of Computer Science, Bioinformatics & Computational Biosciences Unit,

240, Reiss building, Georgetown University, 37th & O Sts NW Washington DC 20057

Keywords: World-Wide-Web, search, nucleotide or peptide sequences

[1] Corresponding author: Natarajan Ganesan, ng6@georgetown.edu .

Abstract

Background - No approaches have yet been developed to allow instant searching of the World-Wide-Web by just entering a string of sequence data. Though general search engines can be tuned to accept ‘processed’ queries, the burden of preparing such ‘search strings’ simply defeats the purpose of quickly locating highly relevant information. Unlike ‘sequence similarity’ searches that employ dedicated algorithms (like BLAST) to compare an input sequence from defined databases, a direct ‘sequence based’ search simply locates quick and relevant information about a blunt piece of nucleotide or peptide sequence. This approach is particularly invaluable to all biomedical researchers who would often like to enter a sequence and quickly locate any pertinent information before proceeding to carry out detailed sequence alignment.

Results - Here, we describe the theory and implementation of a web-based front-end for a search engine, like Google, which accepts sequence fragments and interactively retrieves a collection of highly relevant links and documents, in real-time. e.g. flat files like patent records, privately hosted sequence documents and regular databases.

Conclusions - The importance of this simple yet highly relevant tool will be evident when with a little bit of tweaking, the tool can be engineered to carry out searches on all kinds of hosted documents in the World-Wide-Web.

Availability - Instaseq is free web based service that can be accessed by visiting the following hyperlink on the WWW

http://instaseq.georgetown.edu

Background

The amount of hosted sequence-data in biomedical sciences continues to double rapidly, and much of this information is accessible on the Web. There exists a wide variety of tools in the information pyramid (**Figure 1**), at specialized and sub-specialized levels for sequence analysis and manipulation e.g. BLAST, structure prediction servers , and profiling tools . These tools operate just above the ground-level mosaic of sequence records and other raw data present mainly in databases and other resources. At a higher tier are search engines like Entrez , Sequerome and Bioinformatic Harvester which provide information collected from a variety of databases. However, these engines confine their queries to a relatively small number of specific sources, and do not access the large amounts of data found in the Web, such as text documents and newly created data files in a variety of plain, flat-file formats hosted by special-interest groups. General search engines do access the Web at large, and if properly used, users can enter their sequences directly to locate relevant records. Unfortunately, such searches usually fail to produce the desired results because the user does not adequately prepare the search string. The burden of preparing the search string simply defeats the purpose of quickly locating highly relevant information.

The following example demonstrates the problem. Consider the single-line peptide sequence below,

```
ANLGCWMLVLFVATWSDLGLCKKRPKPGGWNTGGSRYPG
```

A user, unaware of the existence of its NCBI record, might try to grab any relevant information by pasting the sequence into a GOOGLE™ search window. This would not yield any information when, in fact, there exist many records on the web containing this sequence. Queries based directly on FASTA format rarely yield results. Alternately, the user could use BLAST to perform a sequence alignment, but this can be time consuming and requires careful attention to correctly adjust a variety of matrix and relevance parameters. Since neither of these methods obtains adequate results quickly, there exists a need for a simple and fast search tool to locate relevant information, in contrast to elaborate alignment and profiling tools. Despite its limitations, the sheer power and speed of Google search strongly appeals to end-users who would simply require a quick and crude method to locate bare minimum piece of information about an unknown sequence. One can then search for further references and perform specialized searches and detailed alignment analyses.

Implementation

One way to address the problem in the above example would be to insert a space into the sequence at every tenth character and submit this re-formatted sequence as a quoted query string to the search engine. In this case, our search retrieved records for a *Prion* protein with sources ranging from databases to shared flat-files . What happens is that each ten-letter subsequence is treated as a keyword by the Google, and it searches for exact matches to the entire phrase. Since many sequence records on the Web traditionally use this format, this approach yields useful results. However, the user's ten-letter formatted sequence could be frame-shifted with respect to the data source. This can be caused by the user starting with a truncated sequence or the data source having a record with gaps, for instance. Hence another solution would be to generate search strings using all ten possible frame shifts, which would greatly improve the search results. In our experience this is overkill; to assure a minimal, yet useful, search result containing an adequate number of references, it is sufficient to form search phrases by taking any two consecutive ten-character keywords from each of the ten frame shifts, then combining these phrases with the Boolean 'OR' operator. More often than not, such a search retrieves records describing the original source document along with other highly relevant information, if they exist. The reason for this specificity is that the probability of another sequence matching twenty characters as two consecutive ten-letter 'words' is extremely low. Even if the search hit is not specific, the matched sequence is more likely to bear a good 'e-value' with respect to the input sequence.

This simple idea is the basis behind the development and implementation of InstaSeq (**Figure 2**).

Results and Discussion

In our experiments, the retrieved documents were found to be well focused and often pointed to information derived from the input fragment. Databases covered included ExPASy, Uniprot, HPRD, FreshPatents, and text-format documents which included many shared flat-files from small sources. For example, pasting the following fragment of a protein sequence into InstaSeq,

“YLVTHLMGADLNNIVKCQKLTDDHVQFLIYQILRGLKYIHSADIIHRDLKPSNLAVNE”,

returned this link, among others,

http://pkr.sdsc.edu/html/3D/text/1p38/1p38_snap.html ,

which is a snapshot of a 3D structure of the “*mitogen activated protein kinase*” molecule from *Rattus norvegicus*.

Our current tool and approach represents only the tip of an iceberg of many possible extensions that leverage search engine functionality. One goal is to reduce the likelihood of failing to find relevant information when it exists. For instance, InstaSeq’s query string generation can be easily modified to also search for sequence data hosted in non-traditional forms such as protein sequences in three-letter format and DNA primer-sequences hosted by biotechnology companies. Also, nucleotide sequences could easily be translated into protein sequences using all three shift frames. A separate goal is to provide comprehensiveness in query results where relevant data does exist. One way of increasing the coverage of returned information is to use specific keywords directly related to the input sequence. Dedicated Web crawlers could search for documents containing sequence information and collect keywords found there. These keywords could then be indexed to their corresponding sequences in a database, and retrieved when a search for the sequence has been done. **Figure 3** shows a schematic approach of the method, with keywords being fed to a search engine to return a more comprehensive list of documents. Providing navigation aids for reviewing search results would be another opportunity for further development, given that specificity and simplicity are in competition with comprehensiveness.

Since the results from a Google query can vary significantly due to their patented page-ranking technology , InstaSeq may retrieve different document sets at different times for the same input sequence. However, we do not consider this to be a significant shortcoming as Web content itself is highly dynamic. Despite its shortcomings, when compared to a direct Google search on a batch of test sequences, InstaSeq yielded more hits which were all highly relevant. The search worked well in all tested browsers running under a variety of operating systems. InstaSeq is a publicly available, free, web-based service, and can be accessed via its homepage –

http://bioinformatics.georgetown.edu/InstaSeq.htm.

Availability and requirements: InstaSeq can be accessed freely and directly as a web-server at its homepage - http://bioinformatics.georgetown.edu/InstaSeq.htm.

Acknowledgements:

N.G. conceived and designed the project, and together with N.F.B coded the web based tool. M.V. and B.K. provided valuable inputs; N.G. and R.S co-wrote the paper.

Correspondence and requests for materials should be addressed to N.G. (ng6@georgeotwn.edu).

References

1. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. & Lipman,D.J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403-410 (1990).

2. Zuker,M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **31**, 3406-3415 (2003).

3. Schultz,J., Milpetz,F., Bork,P. & Ponting,C.P. SMART, a simple modular architecture research tool: identification of signaling domains. *Proc. Natl. Acad. Sci. U. S. A* **95**, 5857-5864 (1998).

4. Liebel,U., Kindler,B. & Pepperkok,R. 'Harvester': a fast meta search engine of human protein resources. *Bioinformatics.* **20**, 1962-1963 (2004).

5. Ganesan,N. *et al.* A web-based interface facilitating sequence to structure analysis of BLAST alignment reports. Biotechniques . 8-1-2005.

Ref Type: In Press

6. Stothard,P. The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. *Biotechniques* **28**, 1102, 1104 (2000).

7. Maglott,D., Ostell,J., Pruitt,K.D. & Tatusova,T. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.* **33**, D54-D58 (2005).

8. Ganesan,N. InstaSeq Sample Searches. http://bioinformatics.georgetown.edu/InstaSeq_sample_searches.htm . 6-15-2005.

Ref Type: Electronic Citation

9. Google. [#27870287] Test sequences - Old links vanishing. Ganesan, N. 5-15-2005.

Ref Type: Personal Communication

Supporting material The homepage accepts queries into the search box. Queries are best taken in FASTA format with length preferably > 30 characters. Upon submitting the query to InstaSeq, a small javascript on the homepage parses the input sequence into strings of ten-characters in all different frames, truncates at a particular length , takes the beginning two-words in each frame and finally combines all these generated 'words' with an 'OR' Boolean operator. This prepared string is then submitted to Google search (free source code) which then returns the records from the web. The default is a web search but users can also select a SEQUEROME - http://sequerome.georgetown.edu search, to perform a detailed BLAST profiling analysis. The file-types listed include documents enlisted as .doc, .pdf and .ppt thus allowing access to records stored in all formats. Since Google uses page-ranking technology to pull out documents from the web, the results of a same query can vary over time. The following link - http://bioinformatics.georgetown.edu/InstaSeq_sample_searches.htm - provides performance reports on the InstaSeq searches. Though the search yields all relevant links pointing to the input sequence, it is not designed to perform sequence alignment and pattern matching to return records that are similar to the input sequence, which is best carried out by tools like BLAST and its family of specialized tools. The search worked well in the browsers tested (MS Internet Explorer, Firefox, and Netscape Navigator) on several operating systems (Windows, Mac OS X, and Linux). InstaSeq is a publicly available, free, web-based service, and can be accessed via its homepage - http://bioinformatics.georgetown.edu/InstaSeq.htm).

Legends to Figures

Figure 1 shows the pyramid of gene/protein sequence data hosted on the World-Wide-Web and the existing approaches and set of informatic tools to deal with this specialized kind of information.

Figure 2 shows a brief schema behind the working of InstaSeq search tool. Queries are best taken in FASTA format with length preferably between > 30 characters. Upon submitting the query to InstaSeq, a small javascript on the homepage parses the input sequence into strings of 10 in all different frames, truncates at a particular length , take the beginning two-words in each frame and finally combines all these generated ‘words’ with an ‘OR’ Boolean operator. This prepared string in then submitted to Google search (free source code) which then returns the records from the web.

Figure 3 shows the schema of a possible approach by InstaSeq based searches to cover a wide variety of gene/protein sequence related flat-files hosted in the World- Wide Web. The main step lies creating mirrored in-house databases as flat-files and getting them crawled by dedicated crawlers OR WWW engines. After the user query locates the relevant flat-file, a suitable program would then generate a brief list of highly relevant keywords which would be then submitted to Google. This would result in a greatly enhanced set of relevant documents and links pertaining to the input gene/protein sequence.