A Full-Text Search System Covering the Whole GenomeNet

Takao Kataoka **Kenji Satou**
tkataoka@jaist.ac.jp ken@jaist.ac.jp

School of Knowledge Science, Japan Advanced Institute of Science and Technology,
1-1 Asahidai, Tatsunokuchi, Ishikawa 923-1292, Japan

1 Introduction

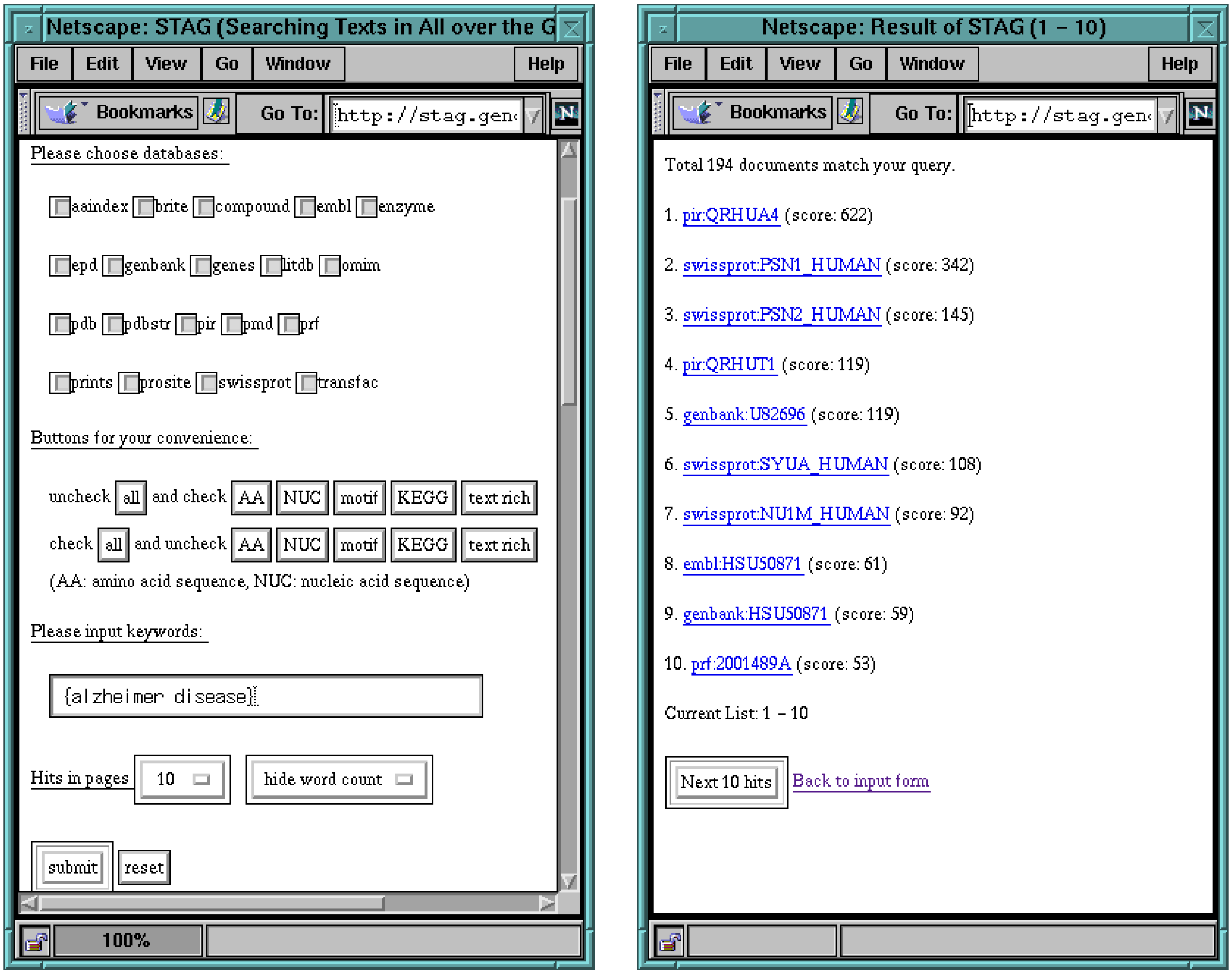
Retrieval of entries by keywords is one of the most traditional services for genome databases. DBGET [1] service in GenomeNet [2] is a typical one which provides keyword search and entry retrieval against about 20 heterogeneous databases (GenBank, EMBL, SWISS-PROT, PDB, KEGG, etc.). In general, such a service is based on keyword indices to entries, which is generated from restricted fields of entry headers, that is, entry identifiers, names of genes and proteins, and so on. Besides this type of approach has advantages in retrieval with high precision, its recall remains low since all the keywords which only occur in the ignored fields do not hit. Furthermore, specialized indexing process, which is sensitive to the description style of fields, tends to need long time computation for generating indices. On the other hand, as WWW grows and grows, technology of full-text search is drastically progressing in these years. Since GenomeNet contains approximately 10 millions of entries, it might be promising to use freely distributed search engines for indexing and keyword search against the whole GenomeNet in practical and sufficient response time. In this poster, a result of such application study is reported.

2 System and Methods

Using Namazu(`http://openlab.ring.gr.jp/namazu/`), which is the most popular search engine freeware in Japan, we indexed most of the entries in GenomeNet. Unlike DBGET, all the fields in entries were processed by the indexer of Namazu except ATOM, CONNECT, and HETATM fields in PDB since they contain huge amount of numeric words harmful for the indexer. In advance with indexing, all the data files were splitted so that one file for one entry. Since GenBank and EMBL is too large, they were processed in parallel. For instance, GenBank has approximately 80 of *.seq files and each of them contains up to 134,494 entries. Entry splitting and indexing were performed in parallel using Sun Enterprise 3000 (4CPU) and PC cluster (18CPU). As to the indexing, it took about 9 hours to process 100 thousands of entries by one CPU. In this system (Fig. 1), 9,358,847 entries are currently indexed and searchable at the following URL.

`http://stag.genome.ad.jp/`

In the case of simple keyword search (e.g. alzheimer) against the whole GenomeNet, this system returns all the answers within 20 seconds (if GenBank and EMBL are omitted, within 5 seconds). Usually it returns much answers since the fields ignored in DBGET are also searched in this system. In other words, there must be a shorthand in this system that answers might include garbages which a user did not want to retrieve. It suggests that this system and DBGET can be coexist in complementary relationship.



Please choose databases:

□aaindex □brite □compound □embl □enzyme

□epd □genbank □genes □litdb □omim

□pdb □pdbstr □pir □pmd □prf

□prints □prosite □swissprot □transfac

Buttons for your convenience:

uncheck all and check AA NUC motif KEGG text rich

check all and uncheck AA NUC motif KEGG text rich

(AA: amino acid sequence, NUC: nucleic acid sequence)

Please input keywords:

{alzheimer disease}

Hits in pages 10 hide word count

submit reset

Total 194 documents match your query.

1. pir:QRHUA4 (score: 622)

2. swissprot:PSN1_HUMAN (score: 342)

3. swissprot:PSN2_HUMAN (score: 145)

4. pir:QRHUT1 (score: 119)

5. genbank:U82696 (score: 119)

6. swissprot:SYUA_HUMAN (score: 108)

7. swissprot:NU1M_HUMAN (score: 92)

8. embl:HSU50871 (score: 61)

9. genbank:HSU50871 (score: 59)

10. prf:2001489A (score: 53)

Current List: 1 – 10

Next 10 hits Back to input form

Figure 1: Screenshots of the input form and the search result.

Acknowledgments

This work was supported in part by Grant-in-Aid for Scientific Research on Priority Areas, "Genome Science", from the Ministry of Education, Science, Sports and Culture of Japan.

References

[1] Fujibuchi,W., Goto,S., Migimatsu,H., Uchiyama,I., Ogiwara,A., Akiyama,Y., and Kanehisa,M., DBGET/LINKDB: an integrated database retrieval system, *Pac. Symp. Biocomput. '98*, 683–694, 1998.

[2] Kanehisa, M., Linking databases and organisms: GenomeNet resources in Japan, *Trends Biochem. Sci.*, 22:442–444, 1997.