**Bioinformation** by Biomedical Informatics Publishing Group *open access*
www.bioinformation.net **Software**

# DAVID gene ID conversion tool

**Da Wei Huang[1, $], Brad T. Sherman[1, $], Robert Stephens[2], Michael W. Baseler[3], H. Clifford Lane[4], Richard A. Lempicki[1, *]**

[1]Laboratory of Immunopathogenesis and Bioinformatics; [2]Advanced Biomedical Computing Center, SAIC-Frederick, Inc., National Cancer Institute at Frederick, MD 21702; [3]Clinical Services Program, SAIC-Frederick, Inc., National Cancer Institute at Frederick, MD 21702; [4]Laboratory of Immuno-regulation, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD, 20892; [$] Both the authors contributed equally; Richard A Lempicki* - E-mail: rlempicki@mail.nih.gov; * Corresponding author

received June 30, 2008; accepted July 05, 2008; published July 30, 2008

**Abstract:**
Our current biological knowledge is spread over many independent bioinformatics databases where many different types of gene and protein identifiers are used. The heterogeneous and redundant nature of these identifiers limits data analysis across different bioinformatics resources. It is an even more serious bottleneck of data analysis for larger datasets, such as gene lists derived from microarray and proteomic experiments. The DAVID Gene ID Conversion Tool (DICT), a web-based application, is able to convert user's input gene or gene product identifiers from one type to another in a more comprehensive and high-throughput manner with a uniquely enhanced ID-ID mapping database.

**Availability:** http://david.abcc.ncifcrf.gov/conversion.jsp

**Keywords:** gene; protein; microarray; proteome; database

**Background:**
Our current biological knowledge is spread over many independent bioinformatics databases, containing both novel and redundant data. Many different types of gene or gene product identifiers (IDs) are selectively used by these different databases and platforms. To leverage heterogeneous annotations across different bio-sources during data analysis, one of the immediate tasks is to convert users' IDs from one type to another as required by the individual source. For example, after an Affymetrix microarray experiment, one must typically translate Affymetrix IDs to gene names, GenBank accessions, RefSeq accessions, UniProt IDs, etc. for further analysis. While this is a time-consuming and tedious process, more importantly, an incomplete or inaccurate translation may easily result in the loss of key information during data analysis.

NCBI's Entrez Gene **[1]** is a popular bioinformatics source for the translation of gene IDs from one type to another. In addition, several ID translation tools also offer this service in a high-throughput fashion **[2-6]** (supplementary file 1), based either on Entrez Gene or on the UniProt/PIR mapping databases **[7]**. The research goal of the DAVID Gene ID Conversion Tool (DICT), one of the components in the DAVID Bioinformatics Resources **[8, 9]**, is to provide a more comprehensive means for batch translations among common gene/protein ID types.

The important features and advances of the DICT are: 1) Enhanced translation capability over other similar tools. 2) Extensive ID type coverage, including more than 20 main and secondary ID types. 3) A batch mode interface in support of one-to-one, one-to-many and many-to-many ID relationships. 4) Hyperlinks to in-depth information about genes are provided for users to exam any potential translation errors. 5) A summary table of the overall translation which is generated for quality control purposes. 6) Capability to handle a mixture of ID types as well as a 'not sure' type.

Approximately 130,000 ID conversion jobs have been conducted with the DICT since 2007 (based on survey on web log file). The usefulness of the tool motivates us to write this application note paper, which intends to introduce the availability, enhanced conversion capability and interface features of the tool to more researchers who have ID conversion needs. However, the technical details behind the features will not be discussed here, but can be found in our other related works for which references are provided in the appropriate sections.

**Methodology:**

**A unique backend database for ID-ID mapping information**

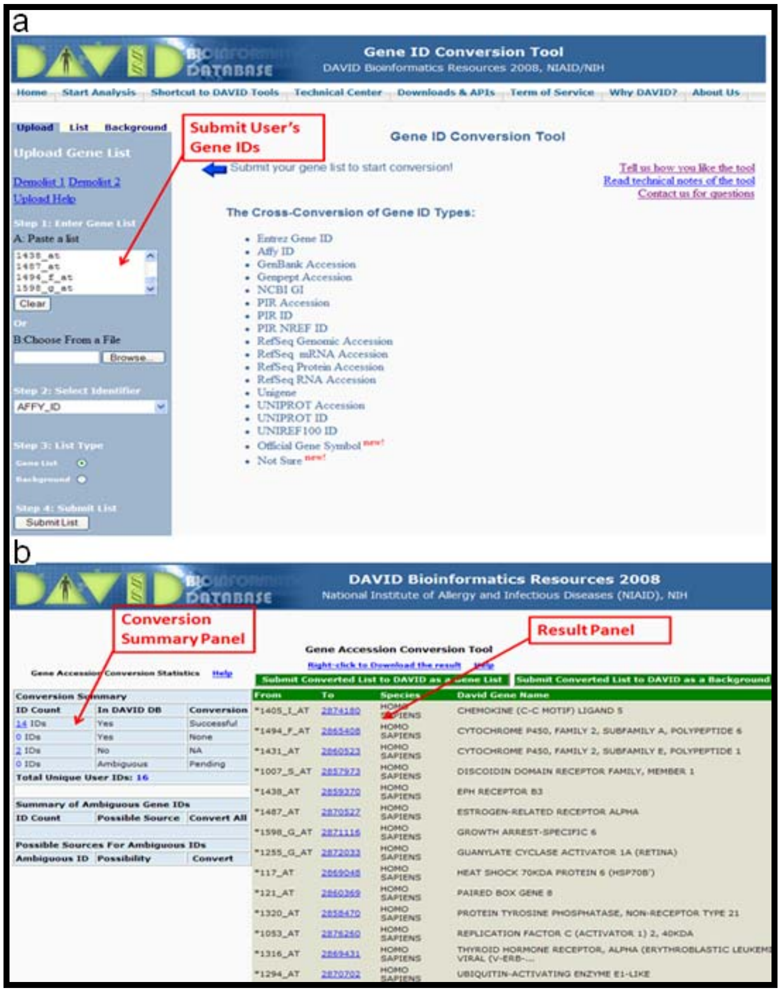
A comprehensive backend ID-ID mapping database is the most important foundation for a better ID-ID translation. The unique advance of the DICT is that its backend ID mapping database, the DAVID Knowledgebase **[10]**, does not simply adopt the popular NCBI Entrez Gene or UniProt ID mapping information as other similar tools do. The DAVID Knowledgebase was specially constructed by comprehensively re-agglomerating ID-ID relationships with a unique procedure, called the DAVID Gene Concept **[10]**. Such a procedure is able to maximally extend additional ID-ID links that were missed in the original systems (e.g. NCBI, PIR and UniProt systems). The newly identified ID-ID links,

© 2008 Biomedical Informatics Publishing Group

as well as the existing ID-ID mapping information from the original systems are stored in the tables of a relational database, where heavy table indexing and a specialized schema are used to enhance the performance of the database query.

### Interactive web-based interface

The DICT is a web based application which does not require any configuration and installation in the client's computers. The output of the program can be described as two panels, i.e. left and right panels (Figure 1). The left panel provides the translation summary and options for ambiguous IDs. The right panel displays the final translation result. Various hyperlinks to in-depth gene information are provided for users to exam any potential errors or alternative translation choices. The results can be either copied over to other tools, such as a spreadsheet, or downloaded as a tab-delimited text file. An additional button is provided for users to directly import the translation result into DAVID for further analysis with other DAVID analytic functions **[9]**.



1438_at
1487_at
1494_f_at
1598_g_at

AFFY_ID

The Cross-Conversion of Gene ID Types:

- Entrez Gene ID
- Affy ID
- GenBank Accession
- Genpept Accession
- NCBI GI
- PIR Accession
- PIR ID
- PIR NREF ID
- RefSeq Genomic Accession
- RefSeq mRNA Accession
- RefSeq Protein Accession
- RefSeq RNA Accession
- Unigene
- UNIPROT Accession
- UNIPROT ID
- UNIREF100 ID
- Official Gene Symbol new!
- Not Sure new!

| Conversion Summary | | |
|---|---|---|
| ID Count | In DAVID DB | Conversion |
| 14 IDs | Yes | Successful |
| 0 IDs | Yes | None |
| 2 IDs | No | NA |
| 0 IDs | Ambiguous | Pending |
| Total Unique User IDs: 16 | | |

| Summary of Ambiguous Gene IDs | | |
|---|---|---|
| ID Count | Possible Source | Convert All |

| Possible Sources For Ambiguous IDs | | |
|---|---|---|
| Ambiguous ID | Possibility | Convert |

| From | To | Species | David Gene Name |
|---|---|---|---|
| *1405_I_AT | 2874180 | HOMO SAPIENS | CHEMOKINE (C-C MOTIF) LIGAND 5 |
| *1494_F_AT | 2865408 | HOMO SAPIENS | CYTOCHROME P450, FAMILY 2, SUBFAMILY A, POLYPEPTIDE 6 |
| *1431_AT | 2860523 | HOMO SAPIENS | CYTOCHROME P450, FAMILY 2, SUBFAMILY E, POLYPEPTIDE 1 |
| *1007_S_AT | 2857973 | HOMO SAPIENS | DISCOIDIN DOMAIN RECEPTOR FAMILY, MEMBER 1 |
| *1438_AT | 2859370 | HOMO SAPIENS | EPH RECEPTOR B3 |
| *1487_AT | 2870527 | HOMO SAPIENS | ESTROGEN-RELATED RECEPTOR ALPHA |
| *1598_G_AT | 2871116 | HOMO SAPIENS | GROWTH ARREST-SPECIFIC 6 |
| *1255_G_AT | 2872033 | HOMO SAPIENS | GUANYLATE CYCLASE ACTIVATOR 1A (RETINA) |
| *117_AT | 2869048 | HOMO SAPIENS | HEAT SHOCK 70KDA PROTEIN 6 (HSP70B') |
| *121_AT | 2860369 | HOMO SAPIENS | PAIRED BOX GENE 8 |
| *1320_AT | 2858470 | HOMO SAPIENS | PROTEIN TYROSINE PHOSPHATASE, NON-RECEPTOR TYPE 21 |
| *1053_AT | 2876260 | HOMO SAPIENS | REPLICATION FACTOR C (ACTIVATOR 1) 2, 40KDA |
| *1316_AT | 2869431 | HOMO SAPIENS | THYROID HORMONE RECEPTOR, ALPHA (ERYTHROBLASTIC LEUKEMI VIRAL (V-ERB-... |
| *1294_AT | 2870702 | HOMO SAPIENS | UBIQUITIN-ACTIVATING ENZYME E1-LIKE |

**Figure 1:** Layouts of the submission page (a) and result page (b) of the DICT.

## DICT Features

### The improved ID-ID conversion capability and the extensive ID coverage

The DICT covers dozens of commonly used types of gene and protein identifiers (Table 1 under supplementary material). Importantly, all types of IDs are fully cross convertible to each other by DICT. In addition, the DICT introduces a special type, i.e. 'not sure'. The 'not sure' type is provided as an aid to users that may not be sure about the type of identifiers that their list contains or that contain a mixture of many types. In both cases, the tool will systematically search all possible identifier types and suggest appropriate choices to the user.

Most importantly, with the uniquely constructed ID-ID mapping information in the DAVID Knowledgebase, the cross-reference capability of ID-to-ID is largely improved **[10]**. Accordingly, the conversion quality and success rate of

Bioinformation, an open access forum
© 2008 Biomedical Informatics Publishing Group

the tool is enhanced as compared to other similar tools. In the supplementary file 2 and 3, the translation results from nine ID translation tools (e.g. ONTO-Translate, MatchMiner, AliasServer, IDConverter, etc.), based on the same set of example IDs, were compared side-by-side. For the particular examples, the DICT is able to handle various combinations of translation tasks in a more comprehensive way than other similar tools.

**The high-throughput capability and entire database download**

The DICT is able to efficiently convert up to three thousand gene IDs at-a-time, which is sufficient for the need of typical high-throughput data analysis. Moreover, if users want to convert IDs for genome-wide genes, such as all Affy IDs to RefSeq, the entire DAVID Knowledgebase is available for download **[10]**.

**Acknowledgement:**

The project has been funded in whole with federal funds from the National Cancer Institute, National Institutes of Health, under contract N01-CO-12400. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

**References:**

**[01]** D. Maglott, *et al.*, *Nucleic Acids Res.*, 33: D54 (2005)

**[02]** A. Alibes, *et al.*, *BMC Bioinformatics,* 8: 9 (2007)

**[03]** K. J. Bussey, *et al.*, *Genome Biol.*, 4: R27 (2003) [PMID: 12702208]

**[04]** M. Diehn, *et al.*, *Nucleic Acids Res.*, 31: 219 (2003) [PMID: 12519986]

**[05]** S. Draghici, *et al.*, *Bioinformatics*, 22: 2934 (2006) [PMID: 17068090]

**[06]** F. Iragne, *et al.*, *Bioinformatics*, 20: 2331 (2004) [PMID: 15059813]

**[07]** R. Apweiler, *et al*: *Nucleic Acids Res.*, 32: D115 (2004) [PMID: 14681372]

**[08]** G. Dennis, *et al.*, *Genome Biol.*, 4: P3 (2003) [PMID: 12734009]

**[09]** W. Huang da, *et al.*, *Nucleic Acids Res.*, 35: W169 (2007) [PMID: 17576678]

**[10]** B. T. Sherman, *et al.*, *BMC Bioinformatics*, 8: 426 (2007) [PMID: 17980028]

**Edited by P. Kangueane**

**Citation: Huang *et al.***, Bioinformation 2(10): 428-430 (2008)

**License statement:** This is an open-access article, which permits unrestricted use, distribution, and reproduction in any medium, for non-commercial purposes, provided the original author and source are credited.

## Supplementary material

| ID type | Count | ID type | Count |
|---|---|---|---|
| AFFY_ID | 2254679 | PIR_NREF_ID | 3355759 |
| ENSEMBL_ID | 76978 | REFSEQ_GENOMIC | 1866800 |
| ENTREZ_GENE_ID | 1734858 | REFSEQ_MRNA | 645831 |
| GENBANK_ACCESSION | 16828735 | REFSEQ_PROTEIN | 1644632 |
| GENE_SYMBOL | 1693151 | REFSEQ_RNA | 1364 |
| GENEBANK_ID | 20291282 | UNIGENE | 161138 |
| GENPEPT_ACCESSION | 4065385 | UNIGENE | 161138 |
| 'NOT SURE' | all | UNIPROT_ACCESSION | 2864344 |
| PIR_ACCESSION | 282281 | UNIPROT_ID | 2789453 |
| PIR_ID | 308092 | UNIREF100_ID | 2552342 |

**Table 1:** The coverage of gene or protein ID types in DICT.

Other supplementary files can be found at http://david.abcc.ncifcrf.gov/manuscripts/conversion/.

© 2008 Biomedical Informatics Publishing Group