BIOINFORMATICS APPLICATIONS NOTE Vol. 18 no. 11 2002
Pages 1542–1543

GeneCards™ 2002: towards a complete, object-oriented, human gene compendium

Marilyn Safran[2,], Irina Solomon[2], Orit Shmueli[1], Michal Lapidot[1], Shai Shen-Orr[3], Avital Adato[1], Uri Ben-Dor[1], Nir Esterman[1], Naomi Rosen[1], Inga Peter[1], Tsviya Olender[1], Vered Chalifa-Caspi[2] and Doron Lancet[1]*

[1]*Departments of Molecular Genetics,* [2]*Biological Services (Bioinformatics Unit) and* [3]*Molecular Cell Biology, The Weizmann Institute of Science, Rehovot, Israel*

Received on March 26, 2002; revised on May 28, 2002; accepted on June 6, 2002

ABSTRACT

Motivation: In the post-genomic era, functional analysis of genes requires a sophisticated interdisciplinary arsenal. Comprehensive resources are challenged to provide consistently improving, state-of-the-art tools.

Results: GeneCards (Rebhan *et al.*, 1998) has made innovative strides: (a) regular updates and enhancements incorporating new genes enriched with sequences, genomic locations, cDNA assemblies, orthologies, medical information, 3D protein structures, gene expression, and focused SNP summaries; (b) restructured software using object-oriented Perl, migration to schema-driven XML, and (c) pilot studies, introducing methods to produce cards for novel and predicted genes.

Availability: Freely available for educational and research purposes by non-profit institutions at http://bioinfo.weizmann.ac.il/cards/ and academic mirror sites. Commercial usage requires a license.

Contact: marilyn.safran@weizmann.ac.il

Supplementary Information: http://bioinfo.weizmann.ac.il/cards/9pageGC2002Bioinformatics.doc http://bioinfo.weizmann.ac.il/cards/GeneCardByResource.xsd and http://bioinfo.weizmann.ac.il/cards/GeneCardByFunction.xsd

INTRODUCTION

GeneCards is a system of human genes, proteins, and diseases that integrates, searches, and displays gene-centered human genome information, focusing on comprehensiveness versus compactness, presenting *just the right mix* of detail and links. With over two million hits at home, and mirroring by 26 academic sites, it has made major strides in biological data mining, integration, and infrastructure.

FEATURES AND ALGORITHMS

SNP summaries. Genetic variation and allelic association studies are challenges of the human genome project for pharmacogenomic applications and the elucidation of multigenic diseases. Biologists at our Genome Center work with the most common type of genetic variation, Single Nucleotide Polymorphisms (SNPs) regularly, and contributed to the design of the GeneCards SNP subset summaries. SNP information is currently extracted from dbSNP (ftp://ncbi.nih.gov/snp/human/XML). We filter to include only those that are not *artifacts*, not connected to *gene duplication*, fully *specified*, without *ambiguous locations* or low *map quality*, and having single LocusLink and contig ids. A gene's SNPs are prioritized by *location type*: (*coding non synonymous*, *coding synonymous*, *coding*, *splice site*, *mRNA-UTR*, *intron*, *locus*). Each displayed line includes SNP-related, contig-related, and expression level data sections. The initial number of top-priority summaries shown is expandable.

Scoring relevant assemblies. To map genes to entries from databases that assemble ESTs and mRNAs into consensus sequences, relevant GeneCards accession numbers are matched against accessions in the assemblies. 1 point is scored for a matched EST, and 3 for an mRNA. Higher scoring clusters are shown first.

Unigene associations. While most Unigene records associated with a gene are labeled as such, a fair number of unnamed clusters can be mapped to a particular gene even if not so named. If a cluster is not found via gene symbol, associated SWISS-PROT ids are used as search criteria.

Quality checks. OMIM was dropped as a source for location, yet retained for its medical data. Cards are not built for LocusLink interim symbols that conflict with HUGO approved symbols.

INFRASTRUCTURE

The goals of version 3.0 include improving flexibility, supporting partial updates, providing an Application

*To whom correspondence should be addressed.

Programming Interface for incorporation of private data, standardization, maintaining a stable id for each card and improving software maintainability, testability, and quality while retaining the current look and functionality. Procedural Perl was initially chosen for GeneCards because of its simplicity, text-processing strengths, and web friendliness. As the project grew and the need for migrating to industrial strength software became apparent, we chose the evolutionary approach of remaining with Perl, reusing some of the existing code, and redesigning to provide modularity and flexibility.

The XML-Based Database. A plain text file format has served GeneCards well, providing easy packaging for mirror sites, independence from expensive DBMS solutions, and effective searching. A natural progression was to continue to use text files, but to standardize. XML is a meta-language that supports customized tags for describing and providing semantic meaning to structured data. Its typed elements can be arranged to form a nested hierarchy. This exquisitely maps to the data presented by GeneCards, where each source can provide such a hierarchy, possibly containing data elements also found within others. Currently, both versions (text and XML) of the files are maintained, as users become familiar with the new format and start to migrate their data integration scripts. The XML version of the data is about 2.27 times as large as the original (753 255 versus 338 215 KB).

The Schema-Driven Display Software. ***GeneCardByResource.xsd*** defines the format for card text files; top elements define mined sources (e.g. *HUGO*, *SWISSPROT*). ***GeneCardbyFunction.xsd*** depicts the layout of a card on the web; top elements define the functions shown in the display boxes (e.g. *Synonyms*, *Proteins*). To facilitate display automation, we are implementing rule-based code to transform a *ByResource* XML file to a *ByFunction* display. Source URLs are customizable; mirror sites can be configured to link to local pages.

Perl's object orientation is simple and well integrated, but not fundamental. Type safety, proper encapsulation and aggregation are not enforced. An object-oriented implementation in Perl will usually be 20–50% slower than the equivalent non-object-oriented version. The benefits of OOP include simpler analysis methods, cleaner and more compact code, greater modularity, easier debugging, more comprehensible interfaces to modules, better abstraction, less namespace pollution, greater code reusability, scalability, and better marketability (Conway, 2000). Version 3.0 generation software combines an object-oriented skeleton with some non-object-oriented internals. The large data structure of gene-based data is implemented as a hash of hashes, and not as objects, avoiding numerous costly instantiations. All the other parts are implemented in an object-oriented manner, providing gains in modularity, scalability and clarity.

APPLICABILITY

A related project, coined ChipCards, correlates expression data from the Affymetrix GeneChip arrays system with GeneCards genes, enhancing experimental results with rich annotations and links. Similar efforts have been undertaken at GeneCards mirror sites at NIH and at the Curie Institute. Examples of applications of GeneCards usage include: serving as a gateway to web resources for analyzing deafness genes, saving lab time by identifying that certain mutations were polymorphisms and not disease-causing, investigating adult-onset diabetes without obesity in India, converging on the gene related to the PVT heart disease of Beduins, and providing a foundation for homework in undergraduate biology courses.

CONCLUSION AND FUTURE DIRECTIONS

GeneCards has been an impetus for similar efforts (Pruitt *et al.*, 2000; Lenhard *et al.*, 2001; Gilbert, 2002) and has advanced in features, algorithms and structure. We look forward to continuing this adventure. On the software engineering front, we plan to delve into XML schema validation and search mechanisms. In the features arena, we plan to add richer expression data, more orthologies, and visualizing exons on 3D protein structures. The challenge of delineating, locating, and categorizing *all* of the human genes—known, predicted, and novel—continues to drive our research. We are currently engaged in integrating data to produce cards for genes that don't yet have a symbol, as well as to name each gene with a unique, persistent GeneCards (GC) identifier that reflects its chromosomal location.

ACKNOWLEDGMENTS

This work was supported by the Weizmann Institute Crown Human Genome Center and by DoubleTwist Inc.

REFERENCES

Conway,D. (2000) *Object Oriented Perl*, 1st edn, Manning Publications Co.

Gilbert,D.G. (2002) euGenes: a eukaryote genome information system. *Nucleic Acids Res.*, **30**, 145–148.

Lenhard,B., Hayes,W.S. and Wasserman,W.W. (2001) GeneLynx: A Gene-Centric Portal to the Human Genome. *Genome Res.*, **11**, 2151–2157.

Pruitt,K.D., Katz,K.S., Sicotte,H. and Maglott,D.R. (2000) Introducing RefSeq and LocusLink: Curated human genome resources at the NCBI. *Trends Genet.*, **16**, 44–47.

Rebhan,M., Chalifa-Caspi,V., Prilusky,J. and Lancet,D. (1998) GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support. *Bioinformatics*, **14**, 656–664.