News section

PROGRAMMING-FREE TEXT PARSING: CONTENTMASTER

More and more biologists today are encountering problems with data sets becoming too large and complex to manipulate. While copy-pasting goes quite a long way, in many cases the need to parse and transform text in an automated fashion becomes inevitable. In an effort to ameliorate this situation, Itemfield[1] has provided a solution with ContentMaster, a programming-free text parsing tool.

Whereas some bioinformaticians may welcome the possibility to hack out some code using regular expressions in a Perl script, many biologists may be at a loss. Even for a trained programmer, writing regular expressions is quite demanding, and it is often considered to be an art form in itself. ContentMaster invokes 'Parsing-by-Example', in which the user interactively defines what data should be extracted from the document by graphically selecting the desired parts of the text. In other words, there is less use of regular expressions, and more 'pointing and clicking'. Knowing how to write regular expressions may still be necessary at times, but one may get pretty far without writing any. The program also requires that the extracted information be mapped to an XML schema, and provides a relatively intuitive interface for developing this.

Even though it is colourful and mouse-clickable, ContentMaster may seem intimidating at first glance. It does take a while to learn how to use, but there are very thorough and redeeming built-in tutorials and an online demo that come with the software. We found that by using them, learning the software to a useful level is pretty easy – definitely faster than learning how to program. Most of the training is required for the schema design part. Some predefined schemas are included out-of-the-box, already supporting protocols from healthcare informatics. We can only hope that schemas will soon be offered to represent important life-science data types. An editor is included for construction of schemas.

To test the applicability of the software for bioinformatics tasks, we downloaded a FastA format text file from Genbank and parsed it, looking for sequence name. It took a considerable amount of time to create our first XML schema, but we assume this would shorten considerably with experience. The schema construction is very flexible, and could easily accommodate our needs. ContentMaster thus offers an alternative solution to learning a programming language to some users in a biological laboratory.

Josiah Altschuler
Bauer Center for Genomics Research
Harvard University

RECENT INNOVATIONS IN THE GENECARDS SUITE

The GeneCards project was initiated in 1997 to help biologists grapple with the plethora of heterogeneous and complex web-based genomic resources in the wake of the expanding Human Genome Project. Aiming to provide some order out of the chaos, GeneCards (http://bioinfo.weizmann.ac.il/genecards/) offers an integrated database of human genes, their products and their functional annotation. It displays *just the right mix* of textual summaries about each human gene, together with detailed hyperlinks for more in-depth analyses. Simple yet powerful global query capabilities include free text searches, disease-gene lists, and viewing randomly selected genes. Over the years, GeneCards has consistently been upgraded to include new data sources and graphical representations. In March 2004, Version 2.29 was released, and distributed to over 20 academic

mirror sites worldwide. GeneCards is provided free of charge to academic or non-profit institutions, and has recently become available for commercial use via XenneX Inc (Cambridge, USA). This release continues the process of augmenting GeneCards to encompass a suite of add-on applications, including GeneAnnot[3], which improves the mapping of Affymetrix microarray probesets to genes, and GeneLoc[4], which integrates several gene location resources. A unique member of the GeneCards suite is GeneNote, which presents, for most GeneCards[5] genes, an in-house-generated transcription pattern for 12 normal human tissues, along with comparison to electronic Northern and SAGE data.

Version 2.29 brings many new features to the core resource, including: links to pathway information;[6] the addition of two new gene categories 'potentially expressed sequence' and 'RNA gene', thereby adding 844 categorised genes; links to transcript and alignment information;[7] links to the Genetic Association website;[8] improved rendering of 3D protein structures via an upgraded PDB[9] facility; and enhancements to the GeneLoc algorithm to accommodate haplotype-related discrepancies and data source conflicts. In the expression arena, a novel re-normalisation algorithm was implemented for optimal comparison of transcriptome data obtained by diverse technologies (microarrays, e-Northern and SAGE), and GeneNote has been enhanced to include a variety of gene expression ranking scores (Tissue Specificity Indices – TSIs), to better quantify the degree of tissue specificity, which ranges from housekeeping to fully tissue-specific.

Release 2.29 also updates all of the traditional GeneCards features including: the HUGO nomenclature symbol and a comprehensive list of aliases and descriptors, as well as rational gene categorisation; chromosomal cytogenetic and megabase coordinates, embodied in a unique and meaningful gene identifier; protein information such as sequences, domains and families; functional attributes and ontologies; DNA sequence information which includes orthologues, single nucleotide polymorphisms (SNPs) and mutations; disease-related information, bibliography links and links to other specialised and genome-wide databases. These data are collated from 47 sources including LocusLink, Unigene, PubMed, Ensembl, UCSC's Golden Path and SwissProt/TrEmbl, and are offered in both plain text and XML formats. The XML representation of the data is a feature that developers find useful for integrating GeneCards information with other repositories and tools.

GeneCards evolution is not over. New projects include expanding the XML capabilities to serve as a basis for more sophisticated searches, continuing to improve the quality of the GeneCards suite via a semi-automated tool coined GeneQArds, and aiming towards comprehensive expressed sequence tage (EST) assignment to existing and *de novo* GeneCards genes via a new Terra Incognita Discovery Endeavor (TIDE).

Marilyn Safran and Doron Lancet
Department of Biological Services and Department of Molecular Genetics
Weizmann Institute of Science

NCBI-ENTREZ CROSS-DATABASE SEARCH

When it comes to getting information on sequences and databases, many paths lead to the National Center for Biotechnology Information (NCBI).[10] The ENTREZ databases and querying system developed at NCBI offers a powerful and relatively simple way to search most of NCBI's numerous databases. The main drawback with ENTREZ was that only one database could be accessed at a time. Unfortunately, often a researcher is interested in finding all there is to know about a favourite entity – be it a protein, a cDNA molecule or a SNP. In such cases, they were forced to