Product Applications for the Sequence Analysis Collection

Pipeline Pilot™

Contents

Introduction

Hundreds of bioinformatics-related applications are available that handle various computational tasks, and domain experts are expected to maintain an ever-growing number of biological databases as a result. Many software tools can store, manage and retrieve data; and a variety of applications can analyze genomic sequences, SNPs, gene and protein expression results, protein-protein interactions, and pathway and metabolic information.

Bioinformatics software firms recognized a need to house all of these databases and tools into a single interface, and they tried to address this challenge by developing an all-encompassing suite of tools in a single application. They speculated that if a single monolithic suite of commercial tools was available, research organizations would be happy to embrace such a solution. Unfortunately, they were only partially correct.

It's true that biotechnology and pharmaceutical industries use a variety of tools and diverse databases for their discovery efforts, but they are reluctant to purchase them from a single software or content provider. Instead, companies and organizations prefer to use the "best of breed" approach when selecting software and database solutions, where these technologies originate from a combination of commercial vendors, open source communities, and internal development efforts.

Although a research organization may need an assortment of software tools to provide a particular solution, not all groups in an organization may use these tools in the same manner. Each team may have distinctive workflows they want to create, or may have a routine computational process that automates as a pipeline. Having a way to combine these tools and databases within a graphical interface allows an organization to quickly build a computational process that addresses a specific problem at hand.

Many informatics teams have a dual role of not only developing innovative computational algorithms and software approaches, but also supporting their organization's internal laboratory research initiatives. Because these informatics groups can help the research effort directly by creating pre-built workflows or computational pipelines that require a collection of software programs from a variety of sources, the Sequence Analysis Collection (along with other Pipeline Pilot collections) is an invaluable tool for research informatics tasks.

Pipeline Pilot and Bioinformatics

Sequence Searching with Profile HMM

A common workflow for identifying related proteins based on a sequence profile involves the following:

- Comparing a protein sequence against an in-house database of known protein sequences using BLASTp
- Creating a multiple sequence alignment of similar sequences
- Building a sequence profile using a program such as HMMER
- Using the profile as a query to search against a database of unrelated sequences

In the Pipeline Pilot protocol example shown below, a mouse sequence is first searched against a small non-redundant protein database using the *BLASTp* component.

Next, the sequences from the BLAST hit results are automatically sent to the *Multiple Sequence Alignment* component and then aligned.

Finally, the multiple sequence alignment results are viewed and sent directly to the *HMM Build* component, where the resulting HMM profile is used to search against a rat protein sequence database using the *HMM Search* component.

The HMM search results are then displayed in a *Similarity Search Viewer* component.

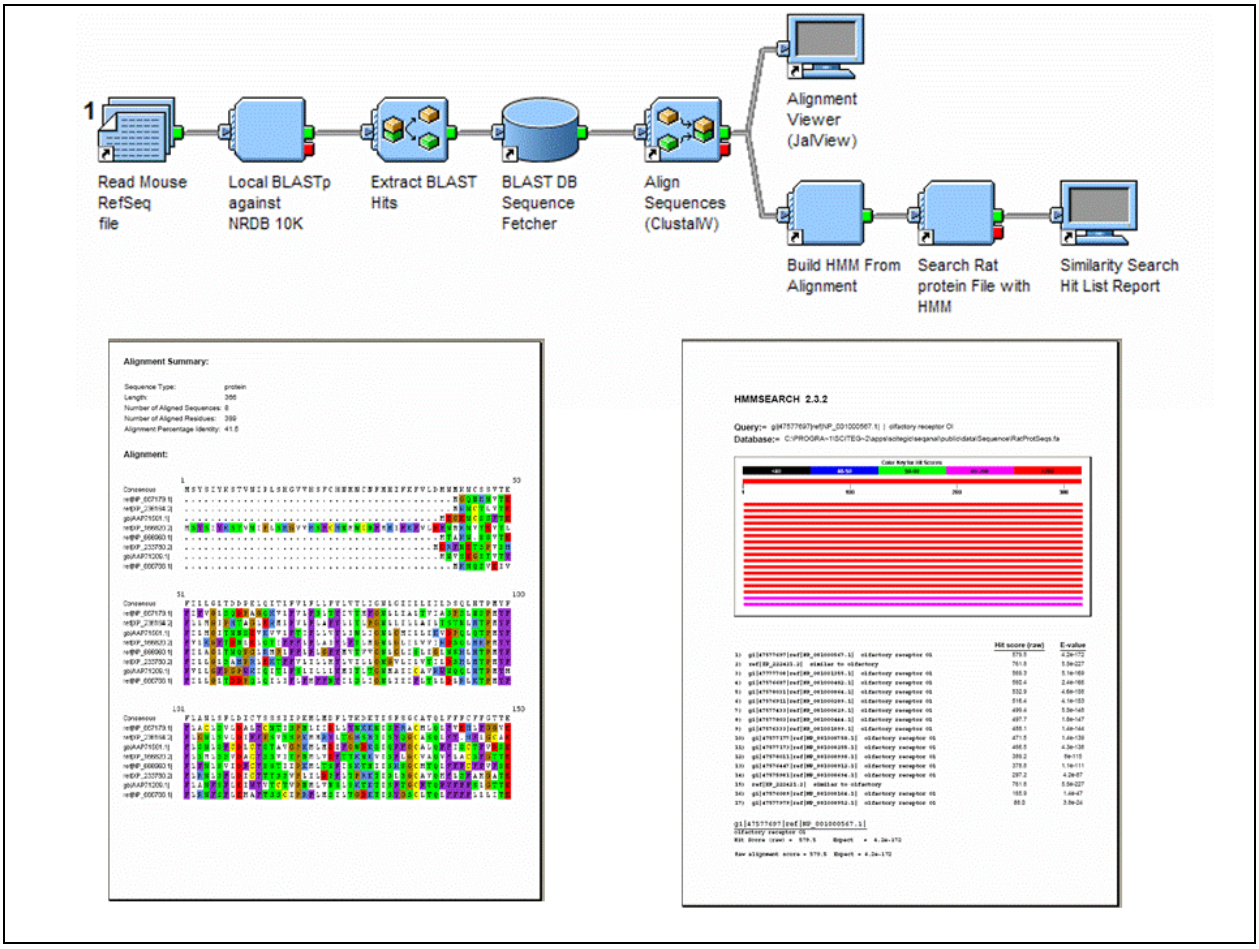

Figure 1: Using profile HMM for protein searching

Integrating Data in a Heterogeneous Protocol

The figure below shows the integration of a variety of analyses commonly used to annotate protein sequences. The protocol generates two items — a Swiss-Prot file and a Sequence Summary view. The *Swiss-Prot Sequence Writer* component creates a file with the new features added to the original sequence entry. It is saved in Swiss-Prot file format.

The Sequence Summary view shows the sequence location of the new features generated from the annotation pipeline, and it can be saved in PDF format.

BioPerl, Perl, BioJava, EMBOSS, and PROSITE seamlessly work together to add new features and annotations to a Swiss-Prot sequence.

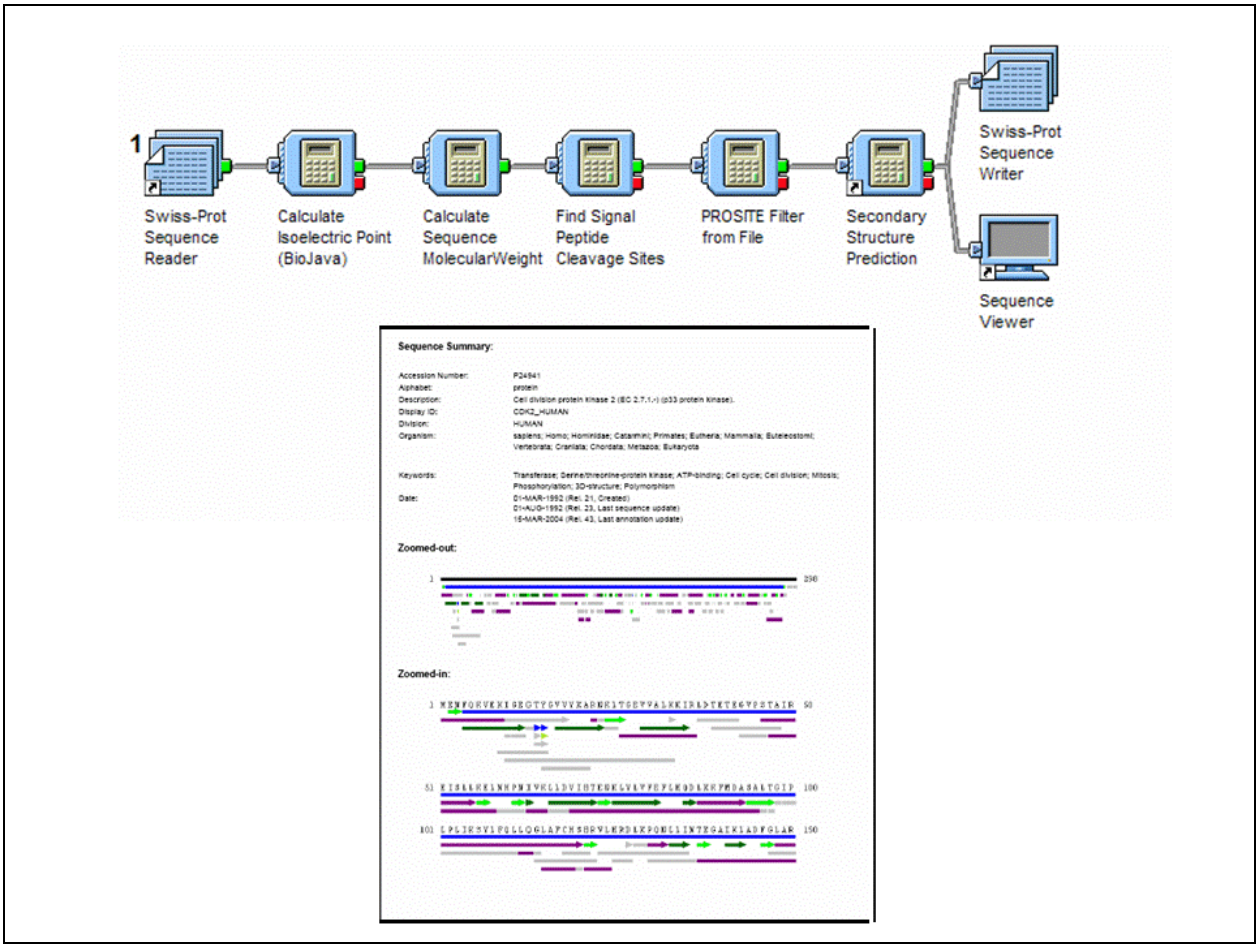

Figure 2: Integrating heterogeneous applications

Creating Custom Reports

The Reporting Collection enables you to create unique and interactive reports in HTML and PDF format. When combined with the components available in the Sequence Analysis Collection, you can design unique and powerful reports in Pipeline Pilot.

For example, you can create useful nucleotide sequence summary reports like the one shown in the example below (on the left). It contains information generated by a protocol that includes items such as GC Content, BLAST results, protein translations, and related references.

The report example on the right (Protein Sequence Summary Report) contains information generated by a custom protocol, and it includes items such as isoelectric point, molecular weight, a protein charge plot, a hydrophobic plot, and related references.

You can also link information in reports to other protocols; when you click a report link, Pipeline Pilot can launch hidden protocols that generate additional reports from other analyses or results.

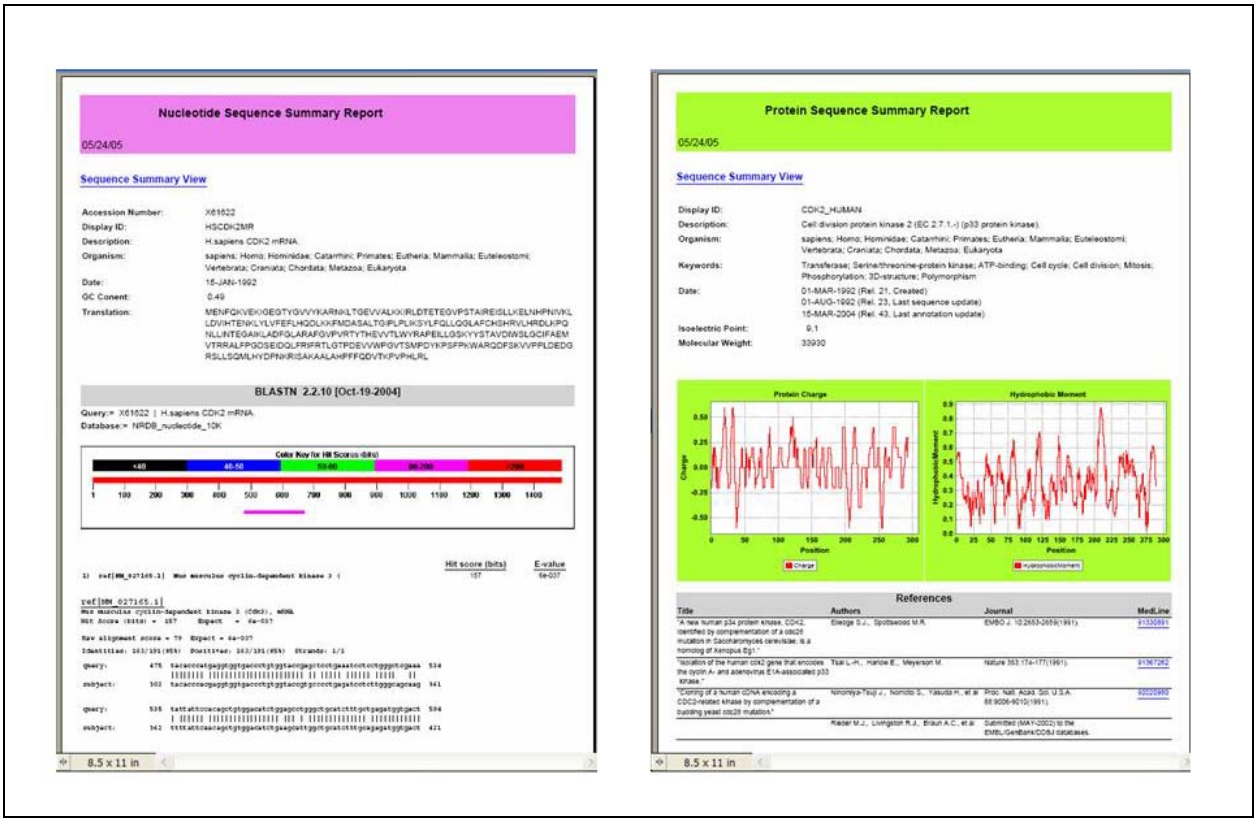

Figure 3: Generating custom reports

Generating Interactive Reports with Comparative Genomics

The example below shows you how you can use the Sequence Analysis and Reporting components to generate interactive reports. In this example, a few proteins from the human RefSeq database and the rat RefSeq database are compared using BLASTp.

Next, the results are filtered and sorted, and three reports are generated with respect to a user-specified "confidence score." Next, a hidden alignment protocol (which includes ClustalW) can be launched from each table report, and an alignment report is created dynamically.

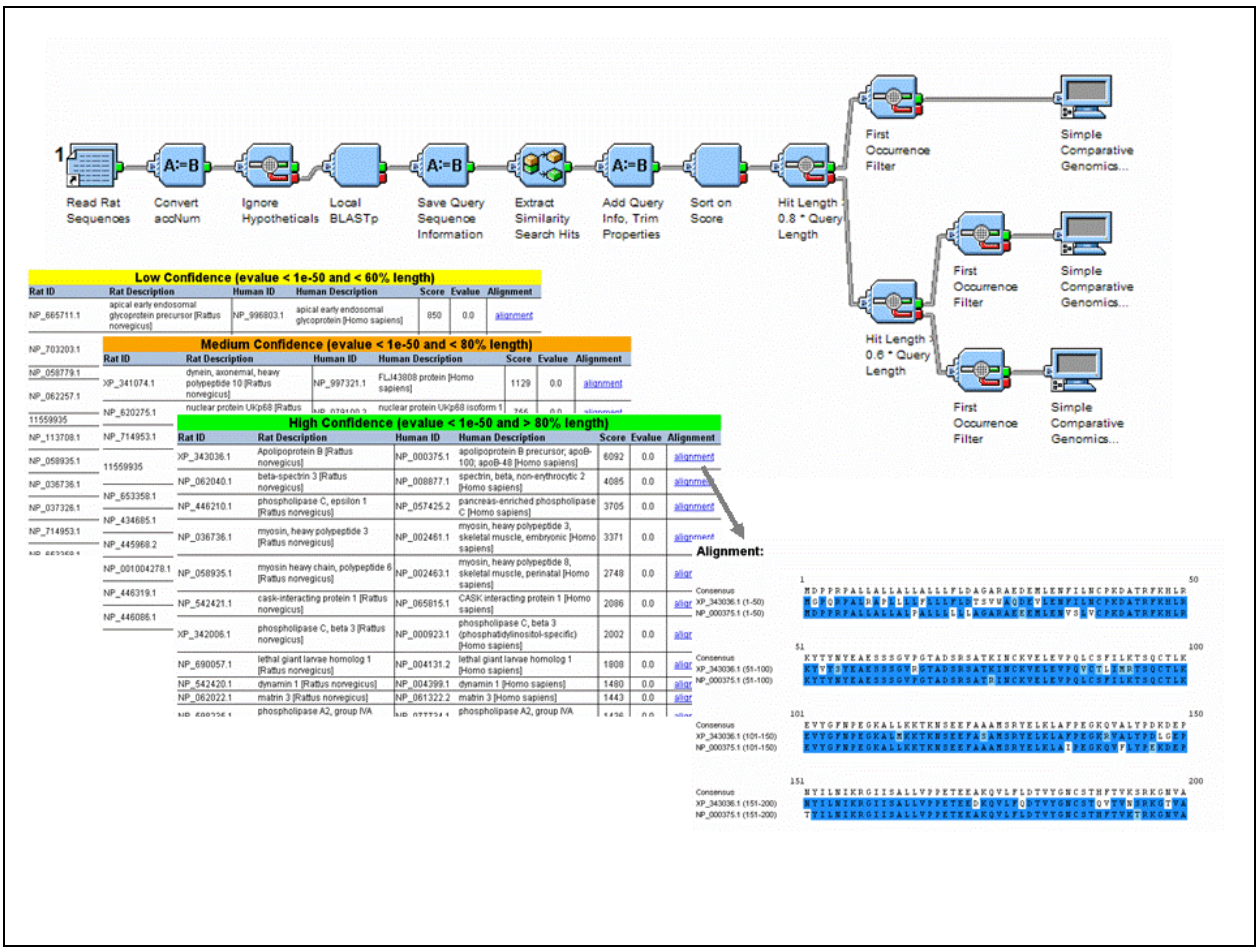

Figure 4: Performing comparative genomics

Finding siRNA Off-Target Sites

One of the newer methodologies used in target validation is called RNA interference (RNAi). One type of RNAi approach uses the effects of small interfering RNAs (siRNAs) to silence a gene. Because there are many programs that predict what siRNAs and associated target sites are good candidates for RNAi experiments, Pipeline Pilot is ideal for automating these tasks.

The example below shows a protocol that uses the *Find siRNA Target Sites* component (a program from EMBOSS called *sirna*) to create a list of potential siRNA target sites. Within the same pipeline, it runs a sequence similarity search using BLASTn to find all potential locations where a predicted siRNA molecule could bind with an organism's genome, thus identifying unwanted off-target sites (in red) and preferred on-target sites (in green).

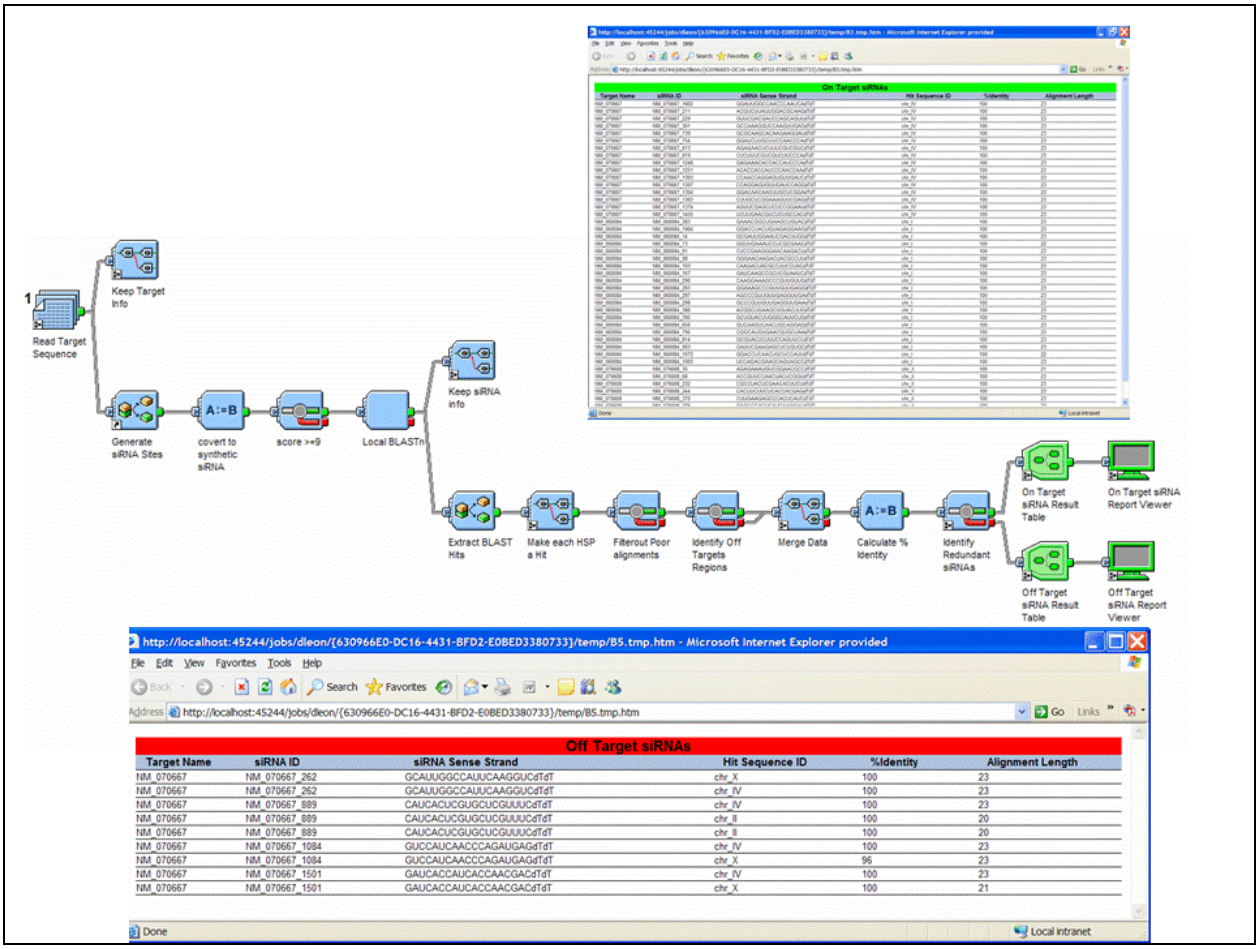

Figure 5: Identifying siRNA off-target sites

Connecting Genes with Relevant Compounds through Chemogenomics

You can use Pipeline Pilot with Kyoto Encyclopedia of Genes and Genomes (KEGG) to connect genes with relevant compounds (or vice versa). You can even integrate a Web service using Simple Object Access Protocol (SOAP). KEGG comprises current knowledge on molecular interaction networks: metabolic pathways, regulatory pathways, and molecular complexes.

In the example below, the protocol starts with an Enzyme Commission (E.C.) number, accesses KEGG, and finds all of the pathways that have that enzyme. Next, compound identifiers are extracted from the pathways and are queried against compound databases to get their structure information. After compound structure manipulations, the information is merged into a final table. The table shows one E.C. number with multiple pathways and associated compounds. You can use these compound structures in additional cheminformatics tasks, such as clustering and substructure searching.

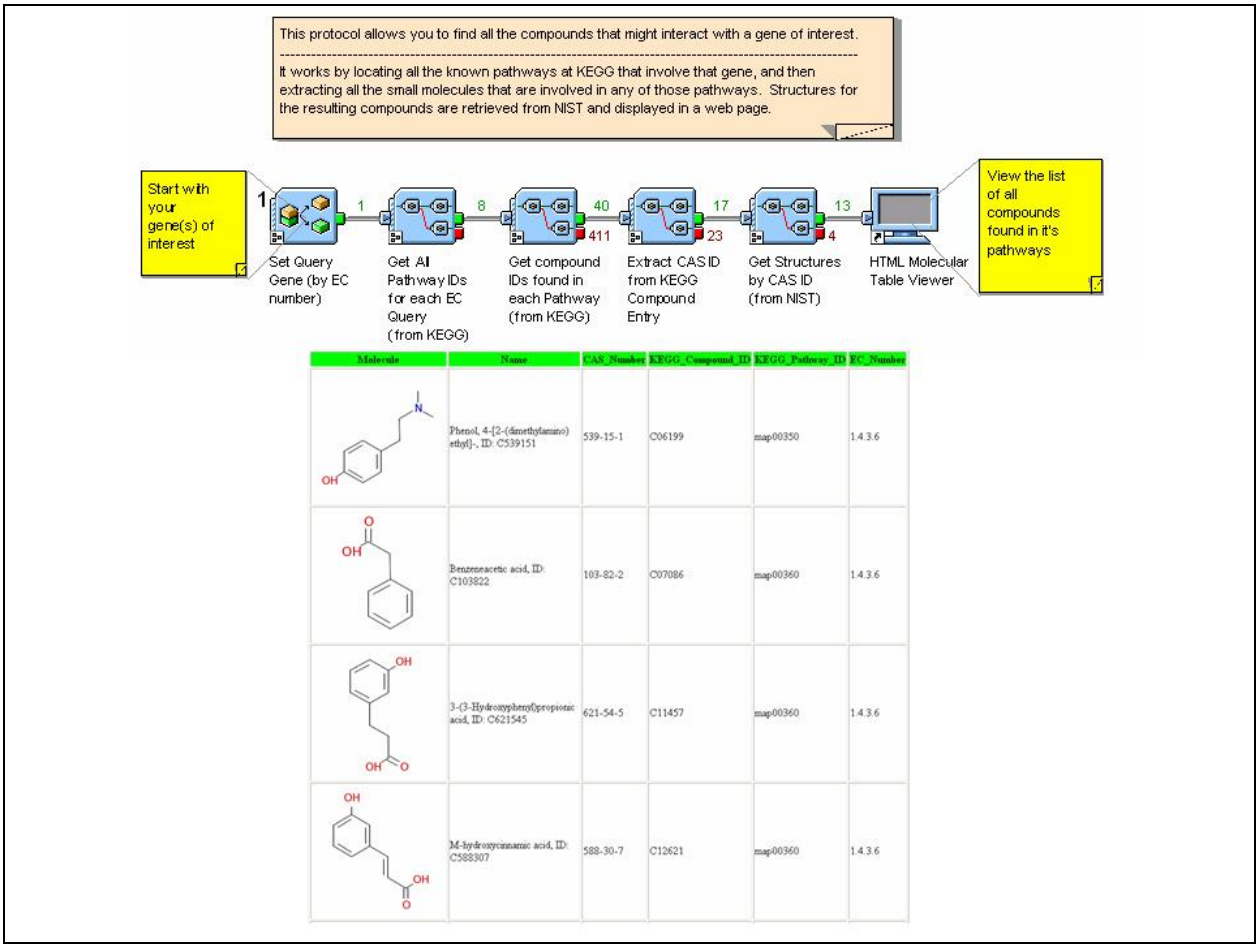

Figure 6: Exploring enzymes, pathways, and compounds

Combining Bioinformatics, Cheminformatics, and Text Analytics into a Protocol

Pipeline Pilot provides an ideal platform to incorporate a variety of *in silico* approaches for finding interesting associations between biological and chemical entities. To illustrate this type of approach, the example below demonstrates how to mine data from Broad Institute's chemical genetics initiative (called ChemBank) and merge the associated data (using the Chemistry Collection). The merged data allows for the clustering of the molecules and their associated protein targets.

The protocol then uses these protein target names (using the Sequence Analysis Collection) as search terms against GenBank to find related sequences. In a parallel pipeline, the same protocol uses the protein target names as query terms against PubMed (using the Text Analytics Collection). The final result is an HTML view showing a cluster of compounds with their associated targets, and an assortment of HTML reports to show scientific articles associated with the each target name.

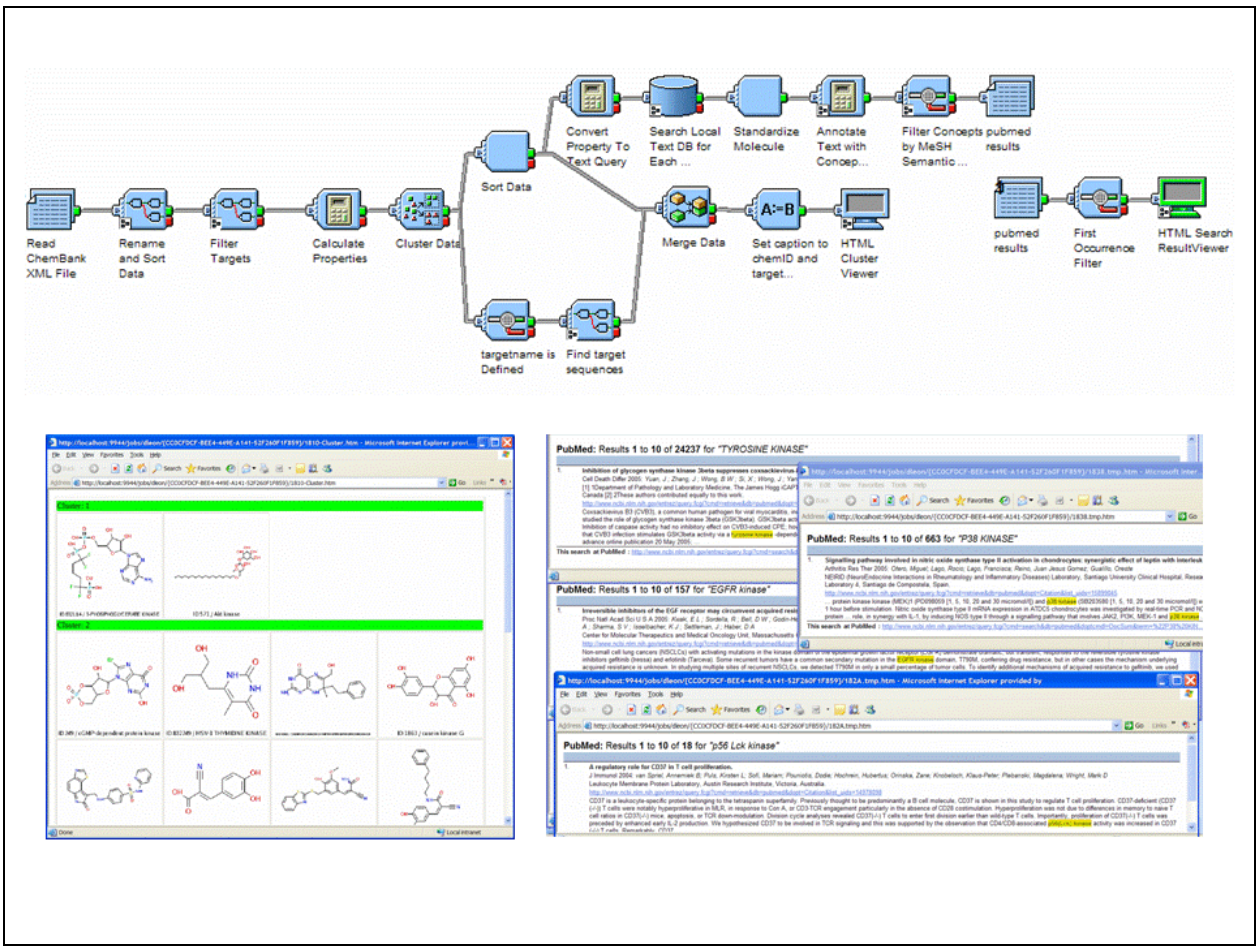

Figure 7: Combining various Pipeline Pilot collections

Summary

As research informatics has matured, there is an increased acceptance of how biological data should be processed. These accepted processes require the automation of many types of workflows and pipelines, and Pipeline Pilot is an excellent tool to facilitate such tasks.

References

Clamp, M., Cuff, J., Searle, S. M. and Barton, G. J. (2004), "The Jalview Java Alignment Editor." Bioinformatics 12, 426-7.

Eddy, S.R. (1998) "Profile hidden Markov model." Bioinformatics 14, 755-763.

Markel, S and León, D. (2003) Sequence Analysis In A Nutshell. O'Reilly, Sebastopol, CA.

Rice P., Longden I. , and Bleasby A. (2000) "EMBOSS: The European Molecular Biology Open Software Suite." Trends in Genetics 16 (6) 276-277.

Rozen, S. and Skaletsky, H. (2000) "Primer3 on the WWW for general users and for biologist programmers." In S. Krawetz and S. Misener, eds. Bioinformatics Methods and Protocols in the series Methods in Molecular Biology. Humana Press, Totowa, NJ, 365-386.

Rutherford K., Parkhill J., Crook J., Horsnell T., Rice P., Rajandream M-A. and Barrell, B. (2000) "Artemis: sequence visualization and annotation." Bioinformatics 16, (10) 944-945.