

## Chapter 12

# **BIOINFORMATICS APPLICATIONS OF WEB SERVICES, WEB PROCESSES AND ROLE OF SEMANTICS**

Satya Sanket Sahoo and Amit Sheth

*Large Scale Distributed Information Systems (LSDIS) Lab, Department of Computer Science, University of Georgia, GA, USA. – {sahoo,amit}@cs.uga.edu*

## **1. INTRODUCTION**

The Human Genome Project (HGP) started in 1990 and ended in 2003, with the aim of discovering the 20,000-25,000 human genes (Barbara R. Jasny et. al. 2003), was the progenitor of the discipline of bioinformatics (David S. Roos 2001). The use of computational tools to store the large amount of data generated by the HGP, to retrieve data and critically to share the data for further study led to development of web based tools and a nascent data management framework.

A biological experimental process consists of multiple stages from 'culture' (involving the growing or collection of sample that contains material of interest) to analysis of the output of a software application. As we see in Figure 12-1, data is generated at all the stages of the experimental lifecycle, in various formats, with different context of use and in extremely large volume. This experimental lifecycle (with various modifications in terms of implementation) with rapid increase in automation at each step is increasingly characterizing biology, from Genomics to Proteomics to Glycomics. This approach is also called 'High-Throughput Experiment' and

is being aggressively adopted by the biological community to deal with the inherent complexity of biology.

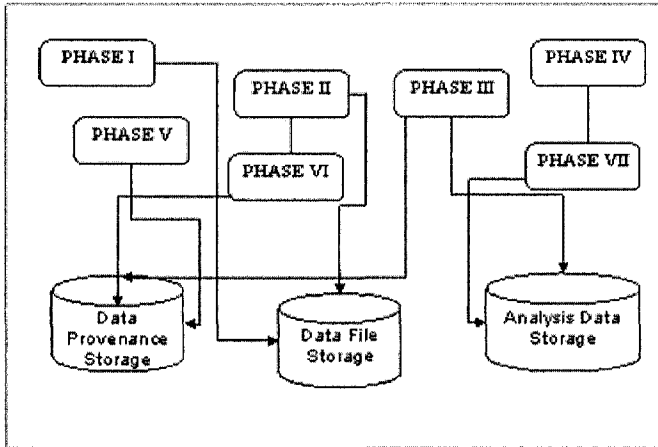


Figure 12-1. A generic biological experimental lifecycle

One of the early data management policy decisions in the HGP was making available the generated data to the community-wide research teams for further study. The World Wide Web played an important role in sharing of this data and making tools, using this data, available for use to biologists. Currently, a large number of applications like BLAST – for homology based search, GenScan – for *ab initio* gene prediction, CUBIC – for binding site prediction, microarray data analysis or ProDom – for protein domain partition are web based tools that use web accessible databases like PDB, KEGG, nr or SwissProt to provide a wide range of computational tools to biologists.

Web Service, with its attributes of platform – independence, web-based access is an ideal framework for ensuring the worldwide use of these bioinformatics resources. Hence, Web Services have been rapidly adopted by the community to enhance the accessibility and usability of their tools. Many bioinformatics tasks involve complex, multi-step processes. If the intermediate steps are implemented as Web Services, their integration to form a Web Process is a logical next step.

## 2. SEMANTIC WEB SERVICES IN LIFE SCIENCE

The chapter focuses on a wide spectrum of disciplines in biological sciences and the application of Semantic Web Services, but there are

multitudes of other fields in bioinformatics that are not covered. Hence, the readers are encouraged to use this chapter as a learning ground to understand the uses of Semantic Web Services and apply it in the context of other areas of biological sciences and related bioinformatics.

There are now more a thousand Web Services (Stevens's et. al. 2006) offering access to biological resources including, public sequence databases, sequence alignment tools and, format converters. Most of these resources are standalone computational tools with minimal interoperability amongst themselves. Often, the output of one Web Service has to be manually ported from one service to another by the user. For example, a BLAST (Altschul SF et. al. 1997) Web Service may require the input data to be in a standard format (like FASTA), and the users have their data in a local format. But, there is another Web Service that takes in data, in any comma separated format, and converts it into FASTA format. Thus the user has to physically move the output of the converter service to the BLAST Web Service as input.

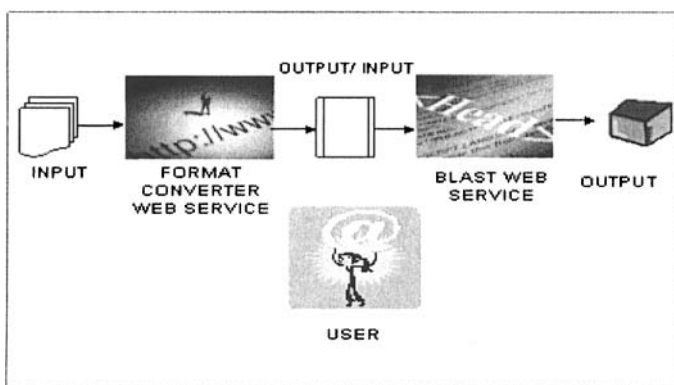


Figure 12-2. Current Web Services often require manual intervention

This form of manual intervention is not feasible in high throughput experimental framework that involves largely automated generation of extremely large amount of data. Hence, composition of Web Services into Web Processes is increasingly becoming a prerequisite in bioinformatics.

Search and discovery of relevant Web Services by researchers can be optimized by use of Semantic Web technology. Using semantic annotation of Web Services, using frameworks like WSDL-S (R. Akkiraju et. al. 2005), will enable semi-automated or automated discovery of Web Services. Moreover, semi-automated or automated composition of candidate Web Services into Web Processes, involving complex processes, mandates the

use of Semantic Web technology to match input, output and data formats of constituent Web Services and their seamless integration.

### 3. BIOINFORMATICS WEB SERVICES AND PROCESSES

In the following sections, we describe various fields of biological research and the application of Web Services and Web Processes in these areas. In the section, ‘Case Study’ we discuss in-depth the role of semantics in the search, discovery and integration of Web Services into Web Processes, with specific example in glycoproteomics. The three broad areas of life sciences research, we describe, are Computational Genomics, Computational Proteomics and Structural Bioinformatics.

#### 3.1 Computational Genomics

The use of computational tools to analyze and interpret genomic data is a broad definition of computational genomics. We cover two specific sections of this vast and rapidly developing field namely, ‘genomic sequence comparison’ and ‘finding potential genes’ in a sequences organism.

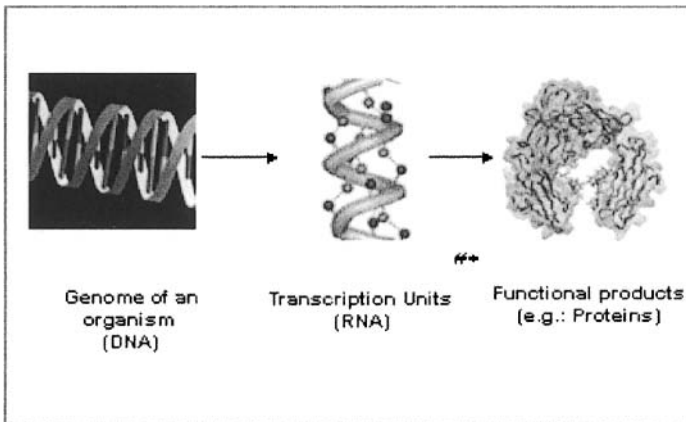


Figure 12-3. The central dogma of biology<sup>1</sup>

<sup>1</sup> # RNA image source: <<http://www.fhi-berlin.mpg.de>>

\* Protein image source: <[http://glycam.ccruc.uga.edu/glycam\\_research.html](http://glycam.ccruc.uga.edu/glycam_research.html)>

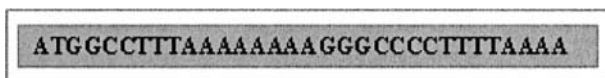
The genome of an organism (constituted of the Deoxyribose Nucleic Acid i.e. DNA) contains the genetic information that is needed by an organism to manufacture needed biological substances to survive. Parts of this genome is *transcribed* into a biological substance called Ribose Nucleic Acid i.e. RNA. This RNA is, in turn, *translated* by other cellular units (ribosomes) that manufacture the corresponding protein or other needed substances. This is also called as the ‘central dogma’ in biology.

Using computational tools, in addition to traditional experimental approaches, the computational genomics field involves gene finding and sequence comparison among other steps. The use of Semantic Web Processes that integrate heterogeneous computational resources, implemented as semantic Web Services, will increasingly play a critical role in aiding genomic researchers.

### 3.1.1 Bio-sequence comparison

#### Background

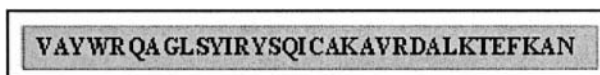
The DNA and RNA biological entities in an organism are made up of linear sequences of biochemical substance called nucleotides. These nucleotides are represented by four ‘bases’ namely Adenine (A), Guanine (G), Cytosine (C) and Thymine (T) (which is replaced by Uracil (U) in RNA).



```
ATGGCCTTTAAAAAAGGGCCCCTTTAAAA
```

*Figure 12-4.* An example of sequence of nucleotides

In case of proteins, the sequences are made of amino acids. There are 20 known amino acids and their combination (along with other biological entities like sugars) decides their biological functions. Each of the 20 amino acids is represented using a specific character, similar to the nucleotide sequences).



```
VAYWRQAGLSYIRYSQICAKAVRDALKTEFKAN
```

*Figure 12-5.* An example of sequence of amino acids

In this section, we focus on the comparison of two or more nucleotide (DNA) or amino acid (protein) sequences. The main aim of aligning sequences is to understand or discover functional, structural and evolutionary similarities. The comparison is done; for example, between a newly sequenced genome of an organism against existing genomes to discover their functionality or identify gene sequences (contain the code for a given protein or other biological entity). The degree of similarity between sequences is a pointer to the gene functionality or identification of the unknown sequence. To compare these linear sequences, they are aligned using algorithmic approaches (that may also use various heuristics to reduce the search space). There are various types of alignments:

- a) **Global vs. local alignment:** In case of global alignment, the sequences are compared in their entirety and gaps in the sequences are inserted, where needed, to make the compared sequences of same length. But, in case of local alignment, a particular portion of the sequence is compared against a portion of another sequence. The aim of local alignment is to look for the optimal alignment between the sub-regions.
- b) **Gapped vs. ungapped alignment:** The alignment algorithm introduces gaps in the sequences to optimize the match, in case of gapped alignment. In case of ungapped alignment, gaps are not introduced in the sequences.
- c) **Pairwise vs. multiple alignments:** Alignment involving two sequences is called pairwise alignment and that involving multiple sequences is called multiple alignment.

There may any permutation of the above types of alignment, for example, local pairwise ungapped alignment or global multiple ungapped alignment.

### **Role of Semantic Web Services**

There are many web-based algorithms for alignment of sequences, with the Basic Linear Alignment Search Tool (BLAST) as the most popular tool. There are two variants of BLAST tool:

- a) **NCBI BLAST:** <http://ncbi.nlm.nih.gov/BLAST>
- b) **WUBLAST:** <http://blast.wustl.edu>

BLAST utility is available in form of Web Services. The Web Services have been developed by many research groups namely, European Bioinformatics Institute (EBI, [www.ebi.ac.uk/Tools/webservices/WSWUblast.html](http://www.ebi.ac.uk/Tools/webservices/WSWUblast.html)), IBM alphaWorks (<http://www.alphaworks.ibm.com/tech/ws4LS>) and are also available as parallel or distributed implementations. For example, the WSWUblast, at EBI, is used to compare a novel sequence with those in a protein or nucleotide database (<http://www.ebi.ac.uk/Tools/webservices/services.html>).

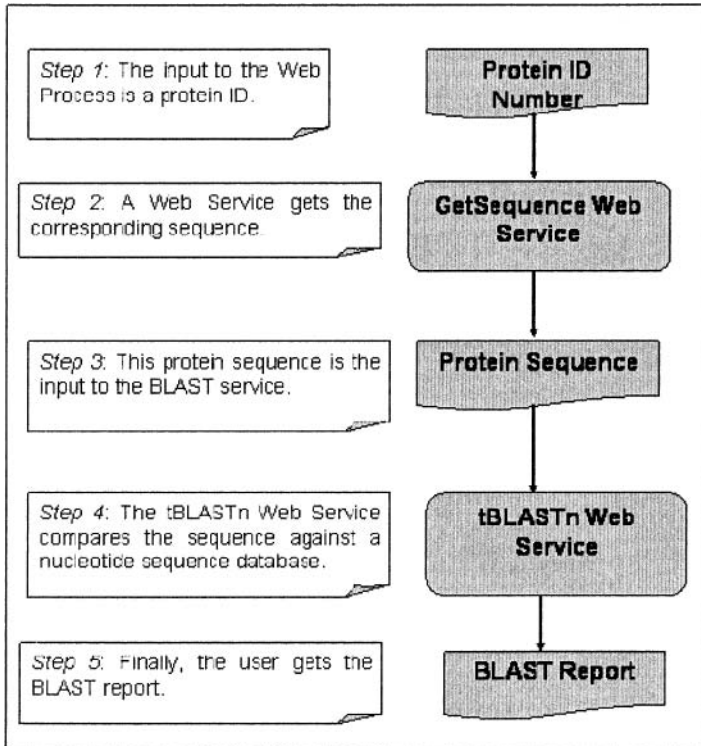


Figure 12-6. A bio-sequence comparator Web Process involving multiple Web Services in sequence

There has been a lot of progress on integrating these Web Services into Web Processes. Many of these initiatives use semantics in the composition of Web Processes using a combination of generic Web Service description and domain ontologies. The generic Web Service description ontologies such as WSDL-S, OWL-S (David Martinet. al. 2004) specify common Web Service concepts. The domain ontology specifies concepts that relate the Web Service to a domain, such as type of service. Workflow engines, namely Taverna (Tom Oinn et. al. 2004) and Pegasus (Sohrab P Shah et. al. 2004), are initialized with available BLAST related Web Services that can be configured and enacted as a workflow.

### 3.1.2 Computational gene finding

A gene in an organism's genome codes for a protein or other biological substances. Computational gene finding involves the identification of sections in the genome of an organism that encodes for relevant bioentity.

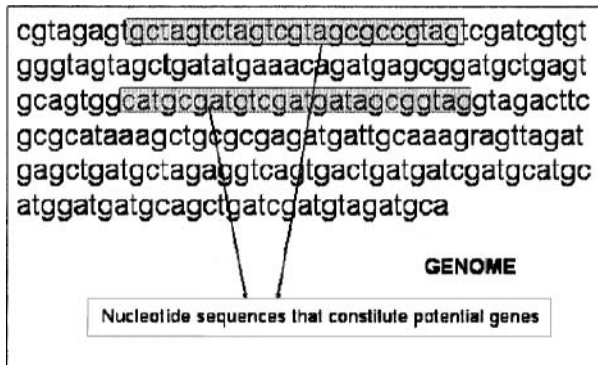


Figure 12-7. Genes in a genome

There are two approaches for gene finding:

- a) **Homology-based methods:** A newly sequenced genome, with unknown genes locations, is compared to homologs in sequence databases. By finding similar sequences, with known genes, to the newly sequenced genome, genes in the newly sequenced genome are predicted.
- b) **Ab initio methods:** This method involves the prediction of genes in a genome using common distinguishing characteristics of known genes.

Some of the common distinguishing characteristics of genes are coding regions and boundaries of coding region.

#### Role of Semantic Web Services:

There are many computational gene prediction tools (using *ab initio* method) that use different algorithmic approaches using multiple modeling techniques. The main drawbacks of homology based technique are the required availability of homologous genomes to the newly sequenced genome (else, homology based prediction is not possible) and the, often, inaccurate prediction of gene boundaries. The following are some of the popular available tools using *ab initio* techniques:

- a) **GRAIL:** (<http://compbio.ornl.gov/Grail-1.3/>) This is a gene finding program for eukaryotic genome, including human and mouse



b) **GeneScan:** (<http://genes.mit.edu/GENSCAN.html>) This tool is based on generalized hidden markov model (GHMM) which models both strands of the DNA. It is mainly used for eukaryotic genomes.

c) **Glimmer:** (<http://cbcb.umd.edu/software/glimmer/>) This tool is generally used for gene prediction in prokaryotic genomes.

Only GeneScan, out of the above listed tools, is also available as a Web Service. A scenario for the use of a Web Process would be for the comparison of results from similar tools (implemented as Web Services) to arrive at a common predicted gene list. This combined approach to *ab initio* gene prediction, using different algorithm and representational model, may be of interest to bioinformaticians.

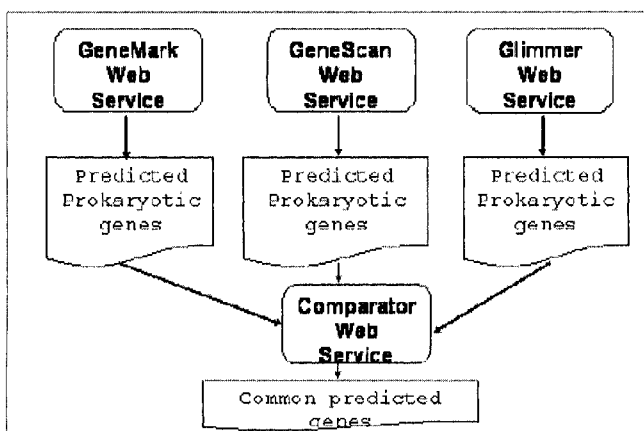


Figure 12-8. This Web Process involves parallel and sequential execution of multiple Web Services to predict genes in given genome.

### 3.2 Computational Proteomics

Proteomics is the study of complete set of proteins produced by a species. The main goal of proteomics is to identify and quantitate the proteins that are present in an organism, cell type, tissue or other cellular parts. We cover one sub area in the proteomics i.e. the prediction of the function of a protein. Other areas of computational proteomics involve the use of similar suite of computational tools, used independently or in combination, to study and analyze proteins.

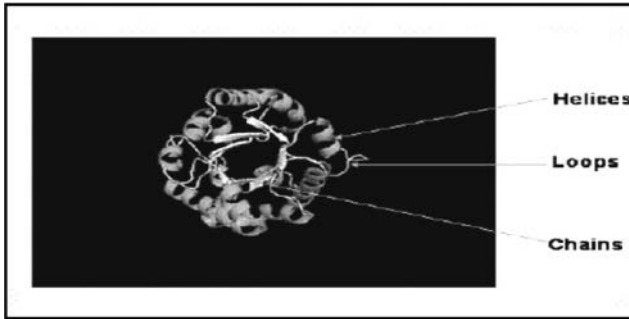


Figure 12-9. Secondary structure of Triose Phosphate Isomerase protein (1Chain)\*

### 3.2.1 Functional prediction of proteins

Proteins may be classified according to their structure or their functions. These classification parameters are not mutually exclusive, but are interdependent. Proteins function is determined by many factors including its constituent sequence, its structure as well as other attached biological entities like sugars. The structure of a protein is also determined by its function, evolved over a period of time.

Protein function may be predicted at multiple levels of specificity:

- a) **Generic function:** For example, a given protein is an enzyme
- b) **Specific function:** The given protein is an enzyme involved in digesting other proteins.

#### **Role of Semantic Web Services:**

There are many different approaches to predict the function of a protein, including:

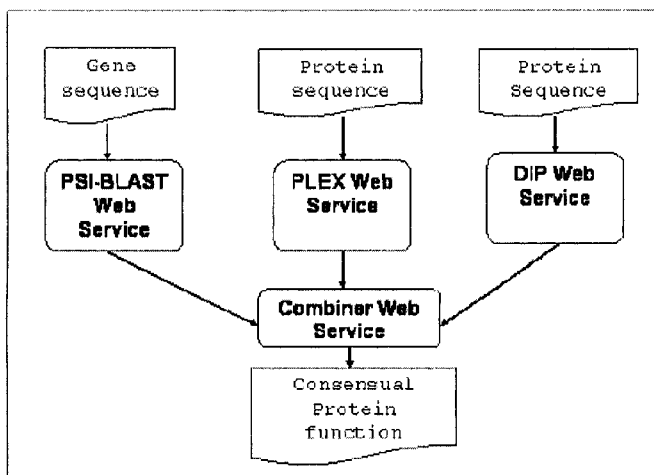
- a) **Sequence comparison:** The new sequence is aligned to known genes in a sequence database and function of the new gene is derived from the known genes. One of the BLAST tools, PSI-BLAST, is used for sequence comparison.
- b) **Phylogenetic profile analysis:** The phylogenetic profile of a protein is a string that encodes the presence or absence of protein in a sequenced

\*protein image source: RCSB PDB (<http://pdbeta.rcsb.org/pdb>) using PyMOL (<http://pymol.sourceforge.net/>) application

genome. The phylogenetic profile of proteins that participate in common functions are often 'similar'. An online tool that does phylogenetic profile analysis is Protein Link Explorer (PLEX) at <http://bioinformatics.icmb.utexas.edu/plex/plex-new.html>.

c) **Protein – protein interaction:** The interaction between two proteins is a useful way to predict the function of new protein. There are many public, web-based protein interaction databases like Protein Interaction Database (DIP) at <http://dip.doe-mbi.ucla.edu/>.

Similar to gene prediction method, these multiple approaches to function prediction in protein may be combined to arrive at a consensual result. This would involve the implementation of the above listed resources as Web Services. These Web Services may be composed, with a number of permutations, into a Web Process.



*Figure 12-10. A Web Process combining multiple Web Services to output a consensual protein function*

### 3.3 Structural Bioinformatics

The determination of structure of biological entities including proteins, RNA, and simulation of interactions between proteins are computationally intensive areas of research in bioinformatics. The structure of a biomolecule plays a critical role in determining its characteristics and functionality.

### 3.3.1 Molecular Dynamic simulation of proteins and interactions

The constituents of biological entities i.e. molecules are perpetually in motion, except at absolute zero temperature. As relevant biological activity do not take place at absolute zero temperature, the motion of the constituent molecules in biological substances determine their conformation. In flexible molecules, such as RNA, proteins or sugars, a single structure cannot describe their structure. Hence, the structure of such biological entities is a suite of individual conformations.

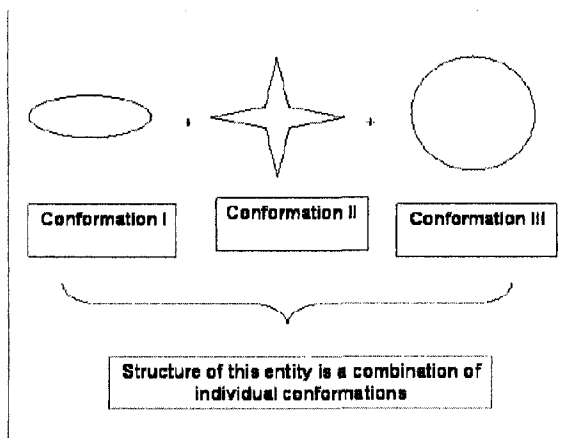


Figure 12-11. The suite of conformations, varying over a particular parameter (E.g. time)

#### Role of Semantic Web Services

The simulation of these individual conformations is calculated using the multitude of forces that act on an entity. There are many algorithmic approaches that take into consideration the various factors acting on an entity to determine the different conformations that fit.

The implementations of these algorithms are extremely expensive in terms of computational resources. There are multiple approaches to optimize the performance of these applications, including dedicated clusters and grid computing.

Grid based Web Services are an exciting area of current bioinformatics research. The notion behind this approach is to distribute the computation of a Web Service across a grid, perhaps transparently to the user, to enhance the time based performance parameters. <sup>my</sup>Grid (Carole Goble 2005) is a project involved in the use of grid based services (mostly Web Services) for data and application resource integration.

Web processes, composed of 'grid-aware' Web Services would be ideal to carry out molecular dynamics simulation computations. A potential Web Process may be a process involving the multiple services that simulate the conformation of a biomolecule under multiple conditions, namely temperature, pressure or time.

#### **4. CASE STUDY**

The common thread in all the above discussed fields of bioinformatics is the implementation of available resources as Web Services and their integration into Web Processes to carry out complex, multi-step biologically relevant function. The discovery of candidate Web Services and their integration into Web Processes is possible only within a semantic framework. In this section, using a case study, we will expand on the application of semantics in the implementation of Web Services and composition of Web Processes in glycoproteomics.

##### **Background**

Proteins, the biological workhorse in an organism, have many modifications after their translation (refer to figure on 'The central dogma in biology') called post-translational modifications. These post-translational modifications play an important role in deciding the function of a protein. One of the post-translational modifications involves the attachment of glycans (modifications of sugars), this process is called glycosylation. Glycoproteomics involves the study of interactions between proteins and glycans. One of the main objectives of glycoproteomics is to identify glycoproteins and quantify their presence.

As part of the Integrated Technology Resource for biomedical glycomics, established by National Center for Research Resources, a team of biologists, biochemists at the Complex Carbohydrate Research Center (CCRC) and computer scientists at the Large Scale Distributed Information Systems (LSDIS) lab at the University of Georgia are working towards the standardization of experimental protocols for high-throughput glycoproteomics research. The different phases of a workflow involved in the glycoproteomics experiment are detailed in Figure 12-12.

The workflow involves both wet-lab experiments (involving experiments conducted by biologists) that cannot be completely automated using computational applications. But, there are many steps that can be automated and exposed as Web Services.

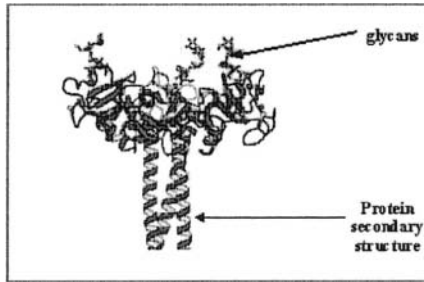


Figure 12-12. The result of a post-translational modification (glycosylation) in proteins (Glycoprotein image source: <http://www.functionalglycomics.org/static/consortium/>)

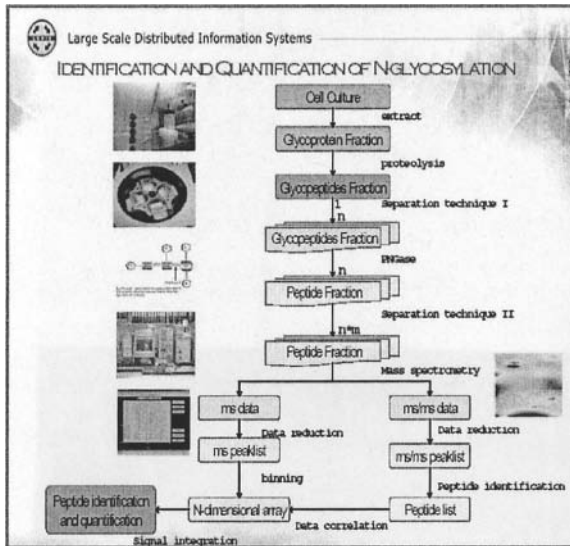


Figure 12-13. The workflow being developed as part of the biomedical glycomics project at the Complex Carbohydrate Research Center (CCRC) and the Large Scale Distributed Information Systems (LSDIS) Lab

```

<?xml version="1.0" encoding="UTF-8"?>
<wsdl:definitions targetNamespace="urn:ngp"
.....
xmlns:
wssem="http://www.ibm.com/xmlns/WeBServices/WSSem
antics"
xmlns:
ProPreO="http://l3dis.cs.uga.edu/ontologies/ProPre
e0.owl" >

<wsdl:types>
  <schema targetNamespace="urn:ngp"
    xmlns="http://www.w3.org/2001/XMLSchema">
.....
  </complexType>
  </schema>
</wsdl:types>
  <wsdl:message name="replaceCharacterRequest"
wssem:modelReference="ProPreO#peptide_sequence">
    <wsdl:part name="in0"
type="soapenc:string"/>
    <wsdl:part name="in1"
type="soapenc:string"/>
    <wsdl:part name="in2"
type="soapenc:string"/>
  </wsdl:message>

```

Figure 12-14. An example WSDL-S of a Web Service used in the glycomics workflow

Semantic annotation of these Web Services with concepts from domain ontology enables their search, discovery and integration using semantic techniques. The domain ontology, ProPreO (S. S. Sahoo et. al. 2005), is used in the semantic annotation of Web Services used in the glycomics workflow. ProPreO is an ontology to model the complete glycoproteomics experiment.

The semantic annotation of these Web Services is at two levels:

- a) **Service level:** This annotation describes the Web Service as a monolithic entity. Hence, a user searching for a Web Service that can parse a protein FASTA file and output a list of protein sequences may search using keywords that describe the task implemented by the Web Service.
- b) **Operation level:** The specific operations in a Web Service may also be annotated using relevant concepts from ProPreO. The annotation includes the description of the input and output of an operation.

Service level semantic annotation help in the search and discovery of individual Web Services, whereas, operation level semantic annotation enable the use of multiple Web Services (or their operations) to be integrated into a Web Process.

## **5. CONCLUSION**

The use of Web Services is increasing at a rapid rate in bioinformatics. Web Services offer the ability of providing web-based access, platform-independent development and deployment. Web Processes, constituted of Web Services, enable automation of complex multi-step processes. The use of Web Services technology enables biologists to process and analyze data at equal pace with high-throughput experimental data generation. But, with increasing number of available Web Services, it is almost impossible to search for a suitable Web Service with specific input and output, by a researcher. Further, the composing of a Web Process using these candidate Web Services is a daunting task for any user.

Hence, use of semantics namely, ontology-based keywords to annotate Web Services enable application to search, discover and integrate Web Services seamlessly. We describe the use of WSDL-S as a method to semantically annotate Web Services. As the field of bioinformatics grows, with an attendant increase in number of available Web Services, the use of semantics is assuming a critical role in enabling their usage by biologists as part of their standard suite of research tools.

## **6. ACKNOWLEDGMENT**

This work is part of the Integrated Technology Resource for Biomedical Glycomics (5 P41 RR18502-02), funded by the National Institutes of Health National Center for Research Resources.

The background content in section 3 is based on the contents of course BCMB 8210, offered by the Institute of Bioinformatics, University of Georgia. All involved teaching faculty (Dr. Ying Xu, Dr. Jessica Kissinger, Dr. PhuongAn Dam and Dr. Rob Woods) are acknowledged.



## **7. QUESTIONS FOR DISCUSSION**

Beginner:

1. Which project is widely believed to be the progenitor of the field of bioinformatics?
2. Why Web Services form an ideal framework for the development and deployment of bioinformatics computing resources?

Intermediate:

1. What are the two types of annotation of Web Services used in the biomedical glycomics project?
2. Name the different types of BLAST search listed at the NCBI BLAST website.

Advanced:

1. Make a list of bioinformatics Web Services registries. Also, list the approach implemented to search and discover Web Services in the Web Services registry.
2. Identify a Web Services, from the three areas of bioinformatics research areas (except structural bioinformatics), which may be implemented over a grid to optimize performance.
3. What are the two different types of ontologies used in the annotation of Web Services?
4. What are the advantages of using ontology based keywords in annotation of Web Services against the use of words from a simple controlled vocabulary?
5. A number of biological domain ontologies are listed at Open Biological Ontologies (OBO) at <http://obo.sourceforge.net/>. List all relevant Web Services for annotating a Web Service that compares gene sequences.

## **8. SUGGESTED ADDITIONAL READING**

- “Current topics in computational molecular biology”, T Jiang, Y Xu and MQ Zhang, MIT Press, 2002

## **9. REFERENCES**

Barbara R. Jasny and Leslie Roberts, Building on the DNA Revolution, Science Apr 11 2003:  
277

- David S. Roos, Bioinformatics--Trying to swim in a sea of data, *Science* Feb 16 2001: 1260-1261
- Roberts Stevens, Olivier Bodenreider, and Yves A. Lussier, *Semantic Webs for Life Science*, PSB 2006, January 3-7, 2006, Grand Wailea, Wailea, Maui
- Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.* 1997 Sep 1;25(17):3389-402.
- R. Akkiraju, J. Farrell, J. Miller, M. Nagarajan, M. Schmidt, A. Sheth, K. Verma, "Web Service Semantics - WSDL-S (Position Paper for the W3C Workshop on Frameworks for Semantics in Web Services)
- David Martin, Massimo Paolucci, Sheila McIlraith, Mark Burstein, Drew McDermott, Deborah McGuinness, Bijan Parsia, Terry Payne, Marta Sabou, Monika Solanki, Naveen Srinivasan, Katia Sycara, "Bringing Semantics to Web Services: The OWL-S Approach", *Proceedings of the First International Workshop on Semantic Web Services and Web Process Composition (SWSWPC 2004)*, July 6-9, 2004, San Diego, California, USA
- Tom Oinn, Matthew Addis, Justin Ferris, Darren Marvin, Martin Senger, Mark Greenwood, Tim Carver, Kevin Glover, Matthew R. Pocock, Anil Wipat and Peter Li. Taverna: A tool for the composition and enactment of bioinformatics workflows *Bioinformatics Journal* 20(17) pp 3045-3054, 2004, doi:10.1093/bioinformatics/bth361
- Sohrab P Shah, David YM He, Jessica N Sawkins, Jeffrey C Druce, Gerald Quon, Drew Lett, Grace XY Zheng, Tao Xu, BF Francis Ouellette. Pegasys: software for executing and integrating analyses of biological sequences. *BMC Bioinformatics* 2004, 5:40
- Scott Doubet and Peter Albersheim, CarBANK. *Glycobiology*, 2, 1992, 505
- Sahoo, S. S.; Sheth, A. P.; York, W. S.; Miller, J. A. "Semantic Web Services for N-Glycosylation Process", *International Symposium on Web Services for Computational Biology and Bioinformatics*, VBI, Blacksburg, VA, May 26-27, 2005.
- Jun Zhao, Carole Goble and Robert Stevens *Semantic Web Applications to E-Science in silico Experiments In Thirteenth International World Wide Web Conference (WWW2004)* pp. 284-285, New York, May 2004
- Carole Goble Putting Semantics into e-Science and Grids in *Proc E-Science 2005, 1st IEEE Intl Conf on e-Science and Grid Technologies*, Melbourne, Australia, 5-8 December 2005