Using Unified Famous Objects (UFO) to Automate Alzheimer's Disease Diagnostics

Michael Gubanov
Department of Computer Science & Engineering
University of Washington
Seattle, USA
mgubanov@cs.washington.edu

Linda Shapiro
Department of Computer Science & Engineering
University of Washington
Seattle, USA
shapiro@cs.washington.edu

***Abstract*—In this paper we discuss automatic pre-diagnostics of Alzheimer's Disease using a new object-oriented data integration technology – UFO (Unified Famous Objects). UFO was originally introduced in [3] to simplify access to heterogeneous data in federated data sources.**

Automatic diagnostics; Alzheimer's disease; data integration

I. Introduction

Automating and simplifying patient diagnostics recently started gaining momentum in biomedical informatics research community due to an explosion in the quality of available patient data stored in clinical and medical research environments. Additionally, reference genomic and medical research data, needed to automate diagnostics, are spread among a variety of heterogeneous data sources – Web databases, clinical data stores, research publications, books, and other. Recently, many journals and funding agencies have also started to require research data to be publicly accessible. However the investigators rarely provide their data in a standard format that can be easily accessed by others, which further complicates data access and integration. Thus, data integration technologies like UFO [3] are needed to automatically fuse data from multiple sources in support of advanced biomedical applications like automatic diagnosis.

The UFO technology was introduced in [3] to simplify and automate data integration from federated heterogeneous data sources [1, 2]. It maintains a large collection of *standard data objects* that conceal representation differences in original data sources behind a static standard interface. The user seamlessly gets access to the needed data in distributed data sources by querying one *standard interface*. The system behind the scenes is responsible for location and conversion of needed data sources that happens seamlessly to the user. Here we describe the application of Unified Famous Objects (UFO) to automatic diagnostics of Alzheimer's disease.

II. Simplifying Access to Data Using UFO

UFO was introduced in [3] and evaluated in [1, 2] with the goal to automate and simplify the tedious tasks of data fusion and integration from several data sources. It creates and maintains a large collection of *standard data objects* that accumulate different representations of the same data object and learns to automatically recognize and map to them. Consequently, the user does not have to worry about data representation in the original data sources and can query one standard interface [1]. UFO repository behind the scenes is responsible for location and conversion of needed data sources that happens seamlessly to the user.

Figure 1 illustrates UFO repository architecture. The information is accessed by querying a collection of standard UFOs present in the repository. This way the user can deal with object U-SEQ or U-Mutation as if all the data sources were storing the data in the same database in the same format. New data feeds are imported by having the repository discover and map objects similar to existing UFOs. For instance, by querying the standard object U-Mutation, the user can easily find the needed mutation record without any knowledge which remote data source it comes from and what was the original data format and representation. To keep it up, the repository needs a collection of standard UFOs and algorithms to enrich this collection from external sources. More details about automatic data object extraction, birth and enrichment of UFOs, UFO mapping, querying, and matching accuracy evaluation can be found in [1, 2, 3].

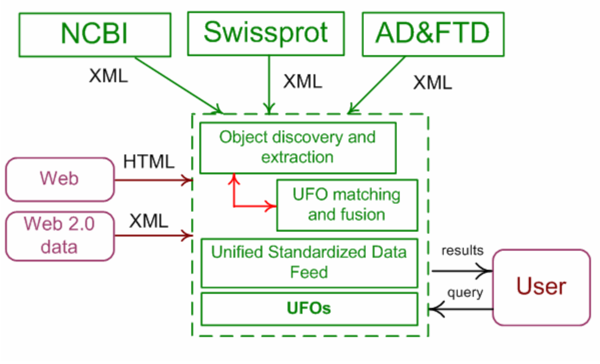

Figure 1. UFO repository architecture

III. Alzheimer's Disease Diagnosis Using UFO

Application of this approach to genetics and system biology is quite prominent, and in this paper will be demonstrated on automatic diagnostics of Alzheimer's disease (AD). AD is the most common form of dementia, characterized by the development of amyloid plaques and neurofibrillary tangles, the loss of connections between nerve cells in the brain, and death of these nerve cells. There are two types of AD - early-onset and late-onset with both of them having genetic links.

Early-onset AD is related to several mutated genes with autosomal dominant inheritance, what means that even if only one of them is inherited from a parent, the person almost always develops the disease. Most autosomal dominant familial AD can be attributed to mutations in one of three genes: amyloid precursor protein (APP) and presenilins 1 and 2. Most mutations in the APP and presenilin genes increase the production of a small protein called Aβ42, which is the main component of senile plaques. Late-onset AD is not directly related to a specific gene, but instead there is increased risk of developing the disease when the apolipoprotein E (APOE) gene is found on chromosome 19.

Figure 3 illustrates automatic pre-diagnostics of an early-onset AD. The system automatically matches the gene or protein sequences from the patient's genetic profile against reference sequences in NCBI (National Center for Biotechnology Information) [6] and Swiss-Prot (Protein knowledgebase) [7], and then predict likely diseases based on the information available in the AD&FTD (Alzheimer Disease & Frontotemporal Dementia) [8] for the detected mutation sites. These semi-structured (XML) data sources are automatically fused and standardized inside UFO repository, so only two queries are actually needed to make a diagnosis. First, the gene name and the type of sequence are acquired from the patience genetic profile. Next, the reference sequence is extracted from the UFO-SEQ populated from NCBI and Swissprot. Finally, the comparison of two sequences yields the mutation sites and numbers that are used to query the UFO-Mutation populated from AD&FTD. The XQueries with the returned results are in Figure 2.

XQueries:

```
    for $y in db2-fn:xmlcolumn('REPOSITORY.
             U-SEQ'),
    where $y/name='PSEN1'
        and $y/seq_type  = 'gene'
        and $y/accession = 'P49768'
    return ($y/sequence);

    for $y in db2-fn:xmlcolumn('REPOSITORY.
             U-MUTATION'),
    where $y/before mutation='E'
        and $y/after mutation='G'
        and $y/site='280'
    return ($y/disease, $y/description,
           $y/frequency, $y/tissue specific);
```

Results:

```
    Defects in PSEN1 are a cause of familial
early-onset Alzheimer disease type 3 (AD3)
[MIM:607822]. AD3 is the most severe form of the
disease, with complete penetrance and an onset
occurring as early as 30 years of age. The second
form is late-onset AD (LOAD), with mean age of
onset greater than 58 years. AD is an autosomal
dominant neurodegenerative disorder characterized
by progressive dementia, parkinsonism, and
deposition of fibrillar amyloid proteins as
intraneuronal neurofibrillary tangles,
extracellular amyloid plaques and vascular amyloid
deposits. The major protein found within these
deposits is a small, insoluble and highly
aggregating polypeptide, beta-amyloid protein
(beta-APP42). Defects in PSEN1 result in an
overproduction of beta-APP42. Variant Pro-166, a
very aggressive mutation that causes onset of AD3
in adolescence, not only induces an exceptionally
high increase of beta-APP42 production, but also
impairs Notch intracellular domain production and
Notch signaling, as well as beta-APP intracellular
domain generation.
```

Figure 2. Querying UFO repository to pre-diagnose Alzheimer's disease based on the patient's genetic profile

This relatively simple example illustrates automatic pre-diagnosis of an early-onset AD using UFO repository. All information about proteins and their mutations from several differently structured data sources was accessed through the same standard interface ("UFO-Seq" and "UFO-Mutation") automatically created by UFO repository based on the original data sources. This example demonstrates that instead of writing 8 queries we can write two queries against standard UFOs. Also, even for the whole sequence of genes with thousand of data-sources involved it still would be just two XQueries against standard UFOs. These two simplifications (standard query interface and fewer queries) are critical either for the study of the late-onset AD or any other genetic disease that would require genomic data from many sources. Currently the AD Genetics Consortium was created as a collaborative effort of AD geneticists to collect more than 10,000 samples for Genome-wide association study (GWAS), the DNA analysis studies needed to identify risk-factor genes.

IV. Related Work

Another important direction of system biology is to study relations between an entire node of 1,000 genes and certain pathologies, with the treatment involving intervention in the entire node. This work goes completely against the traditional "every gene is important, alone" approach and has been already demonstrated viable in areas such as the genetic contributors for type II diabetes [4].

The goal of building better maps of disease requires collecting data from large samples of patients over multiple intervals of time. Traditional studies that were key to the current understanding of diseases have been performed by single institutions, often with the primary goal of taking data to build models that were then communicated as the results and conclusions conveyed by citable scientific articles. Usually data were not shared in the hope of extracting additional information later. In contrast, modern approaches to understanding the complexity of the disease requires genetic data to be paired with additional molecular profile data as well as data that may be used to dissect the underlying regulatory model for the specific cellular context of interest [5].

V. Conclusion

Automatic access and fusion of genetic information from multiple sources becomes critical as personalized medicine, automatic diagnostics are gaining momentum. Use of genomics to provide personalized treatment of cancers is exploding. Based on the fact that each tumor has a diverse, ever-changing, and increasingly actionable genetic profiles, the experts expect that in years ahead clinicians will sequence all tumors and develop customized, personalized treatment plan based on the specifics of mutations.

UFO is an object-oriented data integration approach that simplifies access to federated data sources, thus is crucial to enable advanced applications that depend on distributed heterogeneous data sources [1, 2, 3].

References

[1] M. Gubanov, L. Shapiro, A. Pyayt, “Learning Unified Famous Objects (UFO) to bootstrap information integration”, In Proceedings of the 12th IEEE International Conference on Information Reuse and Integration (IRI), Las Vegas, USA, 2011

[2] M. Gubanov, A. Pyayt, L. Shapiro "ReadFast: Browsing large documents through Unified Famous Objects (UFO)", In Proceedings of the 12th IEEE International Conference on Information Reuse and Integration (IRI), Las Vegas, USA, 2011

[3] M. Gubanov, L. Popa, H. Ho, H. Pirahesh, J. Chang, and S. Chen "IBM UFO Repository", In Proceedings of the 35th Conference on Very Large Databases (VLDB), Lyon, France, 2009.

[4] S. Duffy, "FutureMed Day 2 - Eric Schadt, Esther Dyson, a Tour of Kaiser's Innovation Center, and More", At medGadget.com, 2011

[5] A. Butte, A. Califano, S. Friend, T. Ideker, E. Schadt, "Integrative Network-based Association Studies: Leveraging cell regulatory models in the post-GWAS era," In Nature, 2011

[6] NCBI, http://www.ncbi.nlm.nih.gov/

[7] Swissprot, http://www.ebi.ac.uk/uniprot/

[8] AD&FTD, http://www.molgen.ua.ac.be/admutations/