Patrick McGarvey
Final Project
Genomics and Bioinformatics
15 December 2000

# **Issues Surrounding Genomic Database Interoperability**

Over the past five years there has been an explosion in the amount of data that now comprises ever more bioinformatics databases. While this information has proliferated and advanced the study of genomes and genes, a major barrier to furthering the study is the lack of cohesion between the databases. The problem is multi-faceted; the first problem is the many different methods for referring to the same section of DNA and its products, and on the second is the disparate structure of the databases. The methods for organizing genomic data into relational databases have been identified and addressed in many unique ways, yet no one system has emerged as the de-facto method of annotation. Additionally, as a field bioinformatics has, thus far, failed to implement any standardization for the storage and presentation of the data derived from its study. "Integration of molecular biology databases remains limited in practice despite its practical importance and considerable research effort" (Cheung, et al[1]). There are two ways of addressing the issues of database interoperation; either scientists can deal with the databases as they are and create software solutions for database integration, or the discipline can adopt a set of standards for annotating genes.

The first major problem preventing the smooth interoperation of databases is the many unique ways to which a gene and its products are referred. A casual browse through the ExPASy website, one of the well-funded sites, is instructional. Looking at a SWISSPROT entry for the Yeast gene, ARGD_YEAST indicates that there are 31 hypertext links to outside sources using 12 different methods of identification. While links are an essential part of the web and bioinformatics information sources, they are by no means an effective method of seamless data integration. "To realize the full potential of biological databases (DBs) requires more than the interactive, hypertext flavor of database interoperation that is now so popular in the bioinformatics community" (Karp, 1995[2]). There are so many different ways of referring to the same section of DNA and its products. As innovation continues without a set of rules which allow more in-depth integration of databases, deeper insights will continue to elude scientists.

The second problem with bioinformatics databases is the different methods of setting them up. As investigators look to use their own data, they organize their data to suit their own needs. Consequently, almost every research group uses their own method of setting up their data. Due to the independent nature of academic research, most people only consider their own needs when coming up with methods for storing data. Even large coordinated projects like the Human Genome Project (HGP) is encountering this problem. "One practical problem in achieving these goals [of nonredundant data sets] is that the data repositories of concern to the HGP are maintained independently and are not coordinated with each other in terms of structure and/or content."[3] If a standard for organizing databases had been established, then greater integration would be possible.

Part of the problem lies in the nature of bioinformatics research, which is fragmented, high-paced, and dominated by the pressure to get ahead and be the first to innovate. In the rush to research and innovate, conventions are frequently forgotten, ignored, or disregarded as obsolete and archaic resulting in disparate data sets. "These databases are often isolated and are characterized by various degrees of heterogeneity: they usually represent different views (schemas) of the scientific domain and are implemented using different data management systems (Markowitz, et al. 1995[4]). But alas, there are solutions to this problem and all is not lost.

There are two viable solutions for dealing with the two similar problems; either software solutions can be developed or a standardized system for annotation and data structure can be implemented. Several groups are involved with research into database integration and they have come up with software solutions for integrating databases by merging the desired information into appropriate groupings in a new database. Macauley et al. used a model system to examine the intricacies of database integration, and found this to be an effective method of addressing the problems[5]. Cheung, et al. developed a system based upon the structurally simple entity-attribute-value (EAV) model[6]. This is a software system for bringing together databases with different structures. While the EAV model that they use is robust and solves many of the problems of integration, it is an end-use solution that has the potential to simply increase the noise in genomics studies. Perhaps the EAV system could be used to pull the disparate data together into one coherent database, but software solutions may just add another level to the confusion and become another form of database interaction without solving the deeper problems.

The second solution is to implement some sort of standardization on the discipline. While databases like the NCBI, SWISSPROT, and the PDB have established themselves as leaders in the field even they have incompatibilities. Importantly this idea of standardization is not novel; other disciplines have boards that are responsible for establishing standards and holding the discipline to those established norms. Chemistry has the IUPAC which acts to impose nomenclature requirements upon the world of Chemistry research. Prior to the foundation of this unilaterally acting board, chemistry nomenclature and standardization was nonexistent. A danger of introducing a brand new standard for annotation and database structure now is the problems it presents to older databases. A significant portion of the research already performed in the field was compiled in databases years ago, and these databases are stilled used at times for research purposes. The time may have already passed when a standard could be imposed without leaving behind the majority of the work already performed.

In high-paced fields a time arises where reason needs to reign in the field and it appears as though this time may have already passed for the field of Bioinformatics. Database structure and genomic annotation schemes are two serious problems impeding the full realization of genomics research. It seems likely that software solutions will reign supreme as the field grows ever more complex and varied. While serious headway has been made in the field of biochemistry, until the data from these discoveries can be harnessed effectively the discipline will suffer.

[1] Cheung, KH, Nadkarni, PM, and Shin DG. (1998) A metadata approach to query interoperation between molecular biology databases. *Bioinformatics*, 14, 486-497.
[2] Karp PD. (1995) A strategy for database interoperation. *J Comput Biol* 2(4), 573-86.
[3] Cheung, et al.
[4] Markowitz VM, Ritter O. (1995) Characterizing heterogeneous molecular biology database systems. *Comput Biol*. 2(4), 547-556.
[5] Macauley J, Wang H, and Goodman N. (1998) A model system for studying the integration of molecular biology databases. *Bioinformatics*. 14, 575-582.
[6] Cheung, et al.