GENOME: A Networked Database Environment for Human Genome Data

Andreas M. Kogelnik ,2[1] **Shamkant B. Navathe** [2] **Douglas C. Wallace** [1]
kogelnik@emory.edu sham@cc.gatech.edu dwallace@gmm.gen.emory.edu

[1] Center for Molecular Medicine, Emory University School of Medicine
Atlanta, Georgia 30322, U.S.A.

[2] Bioengineering Program, College of Computing, Georgia Institute of Technology
Altanta, Georgia, U.S.A.

Abstract

We have developed the Georgia Tech Emory Networked Object Management Environment (GENOME). GENOME is a prototype database management system (DBMS)/user interface system designed to manage complex biological data, allowing users to more fully analyze and understand relationships in human genome data. The system is designed to allow the establishment of a network of searchable data sources. The DBMS portion of the environment is a hybrid object-relational system which interprets its data structures on-the-fly, resulting in an extremely flexible DBMS. Such a DBMS provides an environment for interrelating distributed data items, allowing users to further explore computational questions in biomedical science in addition to other fields by maximizing access to data. In developing GENOME, we used MITOMAP, a human mitochondrial genome database, as a model genomic database. MITOMAP encompasses one of the most complete collections of genomic data available for a specific locus or chromosome, including functional, population variation, disease mutation, and gene-gene interaction data, as well as complete sequence data for the human mitochondrial chromosome, and thus serves as an excellent model system. An effective DBMS is required for handling the plethora of Human Genome Project data to handle the various locus-specific databases and ultimately to unify all human genetic and biomedical information through the complete human genome sequence. Developing such a DBMS is our goal. We expect that GENOME will be generally applicable to other biological and non biological paradigms as well.

1 Introduction

The Human Genome Project proposes to sequence the entire human genome and to identify DNA sequence variants that are relevant to human diversity and disease. While this project is generating huge quantities of information of relevance to clinicians and human biologists, its diverse nature and enormous extent make it virtually inaccessible to practicing physicians, research scientists, or interested citizens. What is necessary is a system which can integrate the broad spectrum of human genetic data to make it useful and accessible for a variety of individual applications. In this paper, we describe such a comprehensive information storage and retrieval system. The system provides a computing environment which allows 1) the integration of the great diversity of human genetic information through the medium of the DNA sequence and 2) the sharing of such data across the Internet, enabling a spectrum of users with various degrees of familiarity to use it effectively. In the process we have created a prototype database management system (DBMS) capable of managing the volume and complexity of the data that is being generated by the international Human Genome Project and one which will complement existing biomedical literature (MEDLINE) and biological (GenBank) databases.

While the accumulation of information on human biology and health is a primary motivation for the Human Genome Initiative, little attention has been paid as to how such information can be organized or

functionally related to the human genomic sequence which is being generated. There are two primary reasons for this. First, genetic data is complex and multi-layered. Medically relevant genetic information ranges from structural (nucleotide, nucleotide position, mutation position, gene organization, chromosomal assignment), to functional (type of gene, gene expression, tissue expression, mutational studies, biochemical studies), to population (frequencies of mutations in patient cohorts versus control groups, frequencies of mutations in different racial/ethnic groups, linkage disequilibrium), to clinical (patient information, family studies, genotype/phenotype correlations, multifactorial/polygenic predispositions, genetic therapy) data. Thus, given the present state of technology, a successful system requires extremely flexible information systems and experts who understand both the information to be captured and the system on which it will be implemented. Second, no portion of the human nuclear genome yet sequenced has been studied extensively enough to generate the magnitude and variety of data that would necessitate the development of an integrated genetic information system. However, the human mitochondrial DNA (mtDNA) now offers such a data set, with the associated informatics challenge and opportunity. Thus, only recently has the magnitude of such data pushed the limits of information management and integration systems.

We have developed GENOME to model, manage, manipulate, and share the data encompassed by MITOMAP, a human mitochondrial genome database, [4] as well as other data sources. GENOME encompasses a novel method for modeling and sharing data, in particular for highly complex data sets such as genomic data sets associated with DNA sequence data.

1.1 Data Modeling

Scientific, in particular, biomedical databases represent a gross departure from the types of data seen in most large-scale commercial database implementations. The complexity, incompleteness, and extreme range of variability of data and queries have made biological data unsuitable for traditional data models and database implementations. For instance, the primary problem with respect to the relational model and biological data is that relational tables are far too rigid to accommodate the complexity and variability found in biomedical data. A relatively small but complex data set, if kept in a normal form, rapidly must be broken down into a large, unwieldy set of relational tables. Object-oriented (OO) models, while better able to manage the complexity, are less successful at modeling the variability, inconsistencies, and incompleteness of biomedical data as methods must be written to cover all possibilities and changing a model often requires complete reprogramming of database objects. As with other biological databases we have chosen ASN.1 as our internal data transfer and data definition language. ASN.1 provides a rigorous, structured syntax which is platform-independent and thus ensures data integrity as well as data fidelity across platforms as well as time. Thus, in order to fully capture the information available in biological data, we have developed a hybrid data model together with a hybrid database implementation.

Our model retains the concept of data objects with unique identifiers, but it does not include object methods encapsulated within that object. By freeing the object from the requirement of methods, the system gains a number of advantages. There is more flexibility with regard to schema changes and schema inconsistencies when compared with relational and OO models. It greatly simplifies the modeling task so that the end-user of the data can be brought closer to the design process of the data object, thereby improving the fidelity of the data model in addition to the utility of the data being modeled.

Thus, the data can be modeled by creating objects to represent real world objects (just as in OO modeling). However, since in biological systems behaviors are often poorly defined or completely unknown, methods and behaviors are separated from the data modeling process. Data objects can consist of a single attribute of a given data type, or multiple different attribute types as part of a list or set.

By providing rudimentary manipulation functions for core data types and removing the require-

ment of methods for each data object, the system provides a means for non-technical users to easily implement and retain comprehension of their own databases. Users can create new data types by using the core types as building blocks or they can utilize the types defined by others. These non-technical users can now simply describe their data rather than attempting to fit it to rigid fields and the system will construct an ASN.1 schema for the data objects as the user adds data types. Any schema or data object can be referred to by other schemas or data object, allowing the creation of network of schemas and data objects. Schemas and data objects allow references to other schema types directly via nesting and indirectly as prototype templates, allowing for arbitrarily complex networks of objects to be constructed. Prototyping and nesting objects provide an alternative to traditional object-oriented classes and class inheritance which the GENOME data model does not currently support. The use of pre-defined built-in functions give users a means to enter, retrieve, share, and manipulate their data without having to delve into any programming issues. Sophisticated transformations and functions do require some programming expertise and can be added separately. Thus, this flexibility allows arbitrarily complex, arbitrarily large (within the limits of a system's storage capacity) and arbitrarily variable data structures to be defined simply, while retaining a rigorous data schema.

1.2 A Data Modeling Example - MITOMAP and the mitochondrial genome

In modeling our data, we began with an extremely heterogeneous set of data consisting of all of the published information regarding the human mitochondrial genome (MITOMAP) [3, 4, 7]. The human mitochondrial genome is a small but important portion of the human genome. It consists of a closed circular loop of DNA with 16569 positions, and large but incomplete data on 1) the organization and function of genes on this loop; 2) clinical diseases and symptoms associated with particular changes in the DNA; 3) the variation of the DNA between individuals - in particular across specific racial and ethnic populations; and 4) the interaction of two or more genes or gene products. Most importantly, the human mitochondrial chromosome is the only human chromosome to have its complete DNA sequence known. Thus we could use the known sequence as a unique key reference for interrelating the different data areas. By developing a set of schemas for each area and having them refer to one another we have created a data structure which when populated with data allows new relationships between the data to be exposed through querying across these structures (using the sequences as the key unifying element.

Due in part to the extremely high mutation rate of the mtDNA, this relatively small component of the human genome has yielded a disproportionate amount of information concerning human origins, migrations, and diseases. The human mtDNA has accumulated extensive sequence variation over the course of human evolution. To date, our database indicates 884 presumptively neutral point mutations, representing over 5of the exclusive maternal inheritance and a lack of homologous recombination of the mtDNA, neutral polymorphisms have accrued with time on radiating female lineages as women migrated out of Africa and into the different continents to found the modern races. These lineages are demarcated by continent- and population-specific mtDNA polymorphisms which, upon phylogenetic analysis of mtDNA genotypes, define clusters of related haplotypes called haplogroups. To date, one large haplogroup has been found to represent 75sub-Saharan African mtDNAs, seven haplogroups account for 92European-derived mtDNAs, seven haplogroups represent 75mtDNAs, and four haplogroups constitute essentially 100American mtDNAs. In all populations, there exists an impressive amount of genetic substructure for the mtDNA for which GENOME is an ideal tool for furthering investigations.

MtDNA mutations also play a major role in inherited and sporadic human disease and possibly also aging. Greater than 60 mtDNA point mutations and over 100 mtDNA rearrangements have been identified as etiological factors in human disease [2]. Diseases resulting from mtDNA mutations produce a wide spectrum of pathological states ranging from lethal pediatric syndromes to late-onset neurodegenerative disease. Hallmarks of mtDNA diseases include progressive neurological and/or

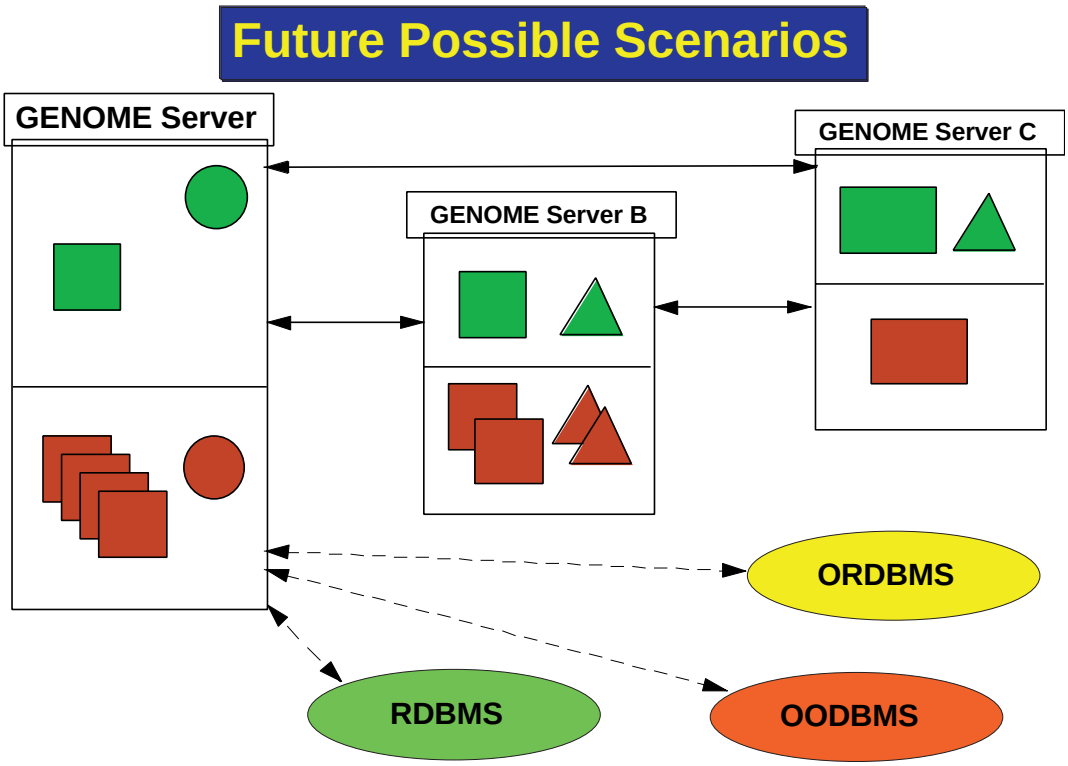

Figure 1: MITOMAP data organization with sample query

muscle pathology, maternal or sporadic inheritance, and phenotypic variability within and between families with the same molecularly defined mtDNA disease [2]. Although heteroplasmy (a mixture of normal and mutant mtDNAs) accounts for a great deal of the variable penetrance and expressivity in these diseases, the natural decline in ATP-generating capacity with increasing age also plays a role, since with age, organ systems ultimately fall below the minimum energetic threshold level necessary to maintain function. The MITOMAP GENOME database allows us to further explore the relationships between disease information, DNA sequence and other genetic data.

Complex genetic and environmental interactions are expected to contribute heavily to the etiology of most common diseases. This area of medical genetics provides perhaps the greatest challenge to bioinformatics, and, until now, has been only superficially examined in nuclear genes due to a lack of data. The assessment of global mtDNA variation has important applications for elucidating the role of mtDNA mutations in disease predisposition and therefore provides a vehicle for the study of complex genetics using bioinformatics. For example, mtDNA haplotype background has been found to influence the relative pathogenicity of the milder of the LHON mutations. LHON is caused by four primary mutations. These are nps 14459, 11778, 3460, and 14484, in decreasing order of severity [2, 1]. To determine if the expressivity of the mutations was affected by background mtDNA haplotype, we determined the complete haplotypes of 47 independent Caucasian LHON pedigrees, encompassing the different primary mutations [1]. Most of the families harboring the np 11778 or np 3460 mutations were found to have a different haplotypes. Hence, their mutations can cause LHON on virtually any mtDNA background. By contrast 75LHON patients and a similar percentage of worldwide Caucasian 14484 patients were associated with a single European haplogroup, J, while about 10Hence, the association between haplogroup J and the 14484 LHON mutation is highly statistically significant [1]. As the 14484 mutation has occurred multiple independent times upon the J lineage, is absent in most normal haplogroup J-positive individuals, and can be found in a heteroplasmic (new mutation) state, it follows

that the 14484 mutation occurs repeatedly, but is only expressed as LHON on the 14484 background. Interestingly, a milder association with haplogroup J mtDNAs has also been observed for the np 11778 mutation. Therefore some factor in the haplogroup J mtDNA must increase the pathogenicity of the 14484 mutation and to a lesser extent the 11778 mutations. It appears, therefore, that some mtDNA lineages in the normal population may promote disease expression through a complex, multifactorial or polygenic mechanism. We are investigating such a mechanism for hypertension and diabetes in African-Americans, in which case the admixture and directional gene flow estimates allowed by mtDNA haplotype data will be essential.

Thus, these studies not only demonstrate the value of the detailed characterization of the global mtDNA variation for identifying mtDNA mutations which confer susceptibility to common chronic diseases, but also emphasizes the need for a comprehensive genetic DBMS to store, combine, and integrate disparate data sets for the study of socially and economically important and genetically complex human diseases. Such a database allows us to easily ask and answer questions which cross-reference data in each of these previously isolated data areas. Important questions such as "What are the evolutionary constraints which occur at specific regions of mitochondrial proteins or mtDNA sequence?", "What are the allele frequencies of specific nucleotide substitutions in various populations?", and "What are the various nucleotide substitutions that have been associated with a specific nucleotide substitution?" can be answered accurately in short order. Hence, the well-studied human mitochondrial genome is unique in that it encompasses the complete spectrum of human genetic data. This makes it an ideal model system for developing a sequence-based computer informatics system for the integration and analysis of genetic and clinical data. We are now in the process of implementing GENOME servers for other genomic loci for which complete sequence data is available.

2 Data Sharing

As a combination of a DBMS, a browser, GENOME has a number of unique characteristics. With GENOME, the traditional distinction between database clients and servers becomes blurred. A GENOME browser can appear to be both a network-browsing client and local database server. However, GENOME provides a method for exchanging structured data objects as opposed to the unstructured documents passed on the WWW. However, a GENOME browser is normally tightly associated with one particular GENOME server, through which it performs many of its actions. In this regard it can be thought of as a window onto a data object server. This server is referred to as its "local" or "home" server, and in most cases the user will establish an identity (or account) on that server so that the user can obtain above average read and write privileges on that server. A browser used without a "home" server will allow viewing of data objects but will be unable to permanently store them (since the user has not identified him/herself to the server). Thus, the server will only provide limited access to information. In order to be able to view and manipulate objects more than transiently, a browser must define a particular server as home by identifying itself (or its user) to obtain a given level of privileges. Browsers can navigate both GENOME databases as well as HTML documents.

GENOME servers are servers which maintain data schemas and data objects and provide these objects to other systems on the network. Data objects stored by a server must have an associated schema on the same server. Schemas are of two types "original" and "reference". An original schema is a data type as defined by a user. A reference schema is a complete duplicate of an original schema with a marker indicating its origin as well as update information. A server with an original schema, in addition to the data structure information, maintains lists of reference schemas on other servers which point to each original schema it houses. In the event of an addition or change to the original schema, the original server notifies the referencing servers of the alteration. Indexes can then be adjusted appropriately on each server. Some users will choose to completely copy (and perhaps modify) an original schema from another server to their own, thereby creating a new original schema on their server. Note again that with biological databases additions to the schema will be much more common

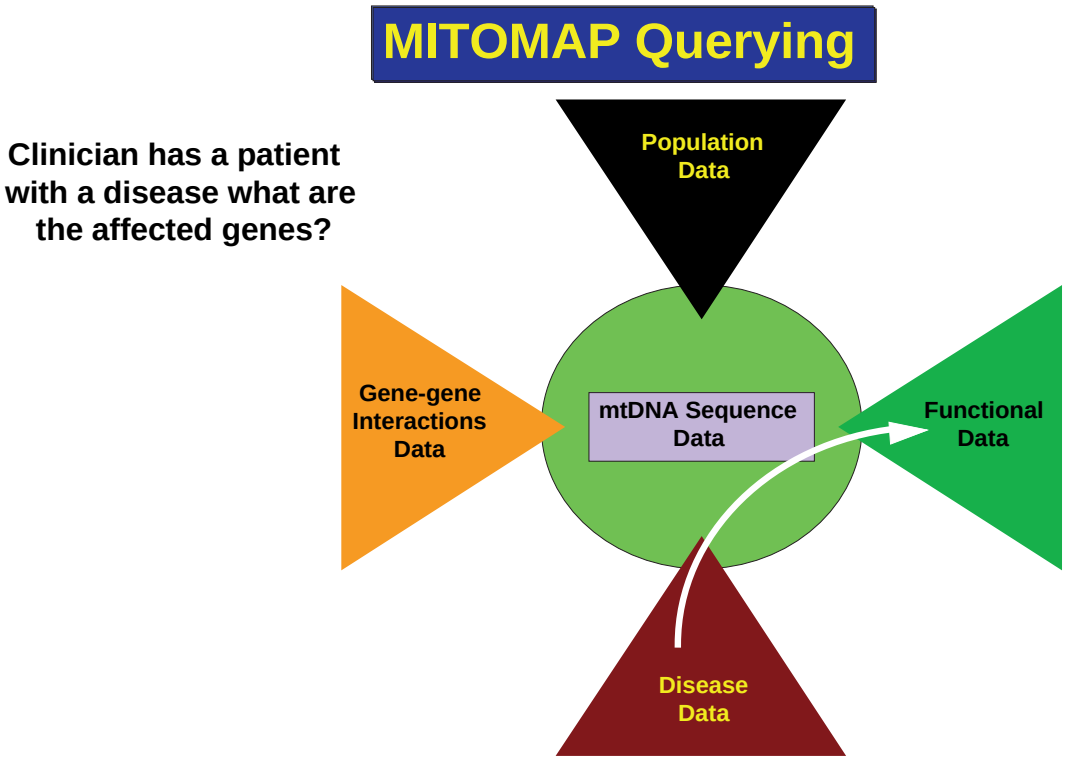

Figure 2: Current and future GENOME network architecture

than other changes - such changes will not significantly affect indexing. When a browser encounters an object with a schema which the browser cannot find on the local server, the schema is requested from the appropriate server and a reference schema is stored on the local server. In the future we hope to extend GENOME by linking GENOME schema objects to scripts which will query other types of database systems (see Fig. 2).

GENOME provides a method for the development and continual updating of our genomic data schema. This data schema, developed specifically for MITOMAP, extends beyond the schema used by GenBank since it must allow for a much wider variety of data types - e.g. individual disease and population data. However, GENOME permits the aligning of schemas such that data and queries can be still be transferred between systems in a meaningful way. The flexibility of the system will allow it to be generalizable to many other applications, however, the remainder of the system implementation gears it to those applications which mimic the biological databases in that data items in general are written once to the database and then infrequently modified but frequently accessed as a reference. Note that this does not imply that the database as a whole is not updated - i.e. new items can be (and are) frequently added, but older items do not change often. Again the biological paradigm diverges from business and accounting applications where the data schemas are relatively simple, but the number of changes and the number of users making changes to the database is large. In contrast biological databases generally have a small group of users who write to the database - i.e. curators who must validate the data before it becomes "public domain", and a large number of users who are interested strictly in accessing and performing manipulations on the data. Note that GENOME has not yet been tested on a large scale and it is unclear what effect it would have on network traffic.

3 System Architecture

The GENOME prototype is set up as a network of data object servers which can request objects from any other server based on the object identifier in a similar fashion as current WWW servers. However, these data object servers will share data objects rather than HTML documents. Each object server or browser has the ability to request objects from any other object server, based on the object identifier (Fig. labelfig1). For example, we have establish a MITOMAP GENOME server which will maintain all of our mitochondrial data objects, and other servers could be established and linked for other locus-specific databases. Clearly this allows an enormous flexibility for data storage and curation. For example, large centralized servers can exist with each individual user having a smaller server (effectively clients) manipulating references to data objects which are actually stored on the central server. Or at the other end of the spectrum, each user could maintain his or her data objects on their individual small server and the larger GENOME servers could request objects from each another. In either paradigm, once an object is read, it can be cached as a reference object on the local server, and an update-notification lock set on the server actually storing the data. This architecture is manageable with biological databases since the number of write operations should remain low when compared with the number of read operations. This allows for more effective and accurate data sharing by allowing the distribution of the responsibility for data curation across many servers (and therefore many individuals). It will take a large number of individuals to curate the enormous amount of data produced by the HGP. In addition this arrangement provides a flexible facility for the scaling of the database over many parallel systems.

4 Querying

Users will be able to formulate queries for their own server, a set of known servers, or unleash a query which will attempt to propagate to all servers which maintain a given type of data object. A query can be formed based on the values of the attributes of an object - in this sense SQL can be used for generating queries. A dynamic interface constructed based on each schema will assist users in formulating queries on specific objects, in addition the interface will allow the anding of multiple queries. Query results can be collection-based (e.g. select a set of object with attribute "X" in the range of 40-100) or attribute specific (e.g. similar to relational operators such as select and join).

In the GENOME paradigm a query can be initiated to span over a single server or all servers on a network which maintain a schema of the same type. This is possible since a server maintaining an original schema keeps track of what other servers might maintain a copy of the schema. Three possible schema-based query types will be supported: a single server query, a query to a specified list of servers, and an "all servers" query. Browsing can also be considered to be a form of non-schema-based querying and focuses on one schema at a time. The simplest scenario is a single server query. Here, a server performs the query only on its own objects and displays the results to the user.

The second scenario is when a user has designated a specified list of servers to be queried. In this case it is not necessary that the user know whether or not a given server maintains a schema used in the query. The querying server sends the request to each server in the list and awaits each result, compiling and displaying the results as they are returned, and notifies the user when all results have been displayed.

An "all servers" query scenario might proceed as follows. The querying server sends the query to all servers which it knows maintain the schema(s) involved in the query. Upon receipt of the message, each server sends a message to any servers which it knows maintain the schema(s) being queried as well as sending a message to the query-originating server naming the servers from which it can expect to receive results. After sending the messages each server executes the query on its set of original objects matching the schema(s) in question and returns the result to the originating server. The originating server meanwhile keeps track of which servers have been sent the query, which have responded, and

which it is still awaiting. As results come in, the server displays them, allowing a user to terminate the search before all results are received if desired. Again the user is notified upon completion of all queries. For schemas which are frequently queried, it is our intention to implement servers which automatically share schema and object storage information so that queries might propagate through a network of servers more quickly. A number of issues regarding network bandwidth and traffic remain to be explored.

Object browsing can be thought of as requesting all schemas from a given server. With this is user can select which types of schema objects are to be queried further. This sort of non-focused query is intended only to provide the user with a path for random explorations of the data or a platform to launch further focused queries.

5 Summary and Conclusions

This work is supported in part by a National Library of Medicine Fellowship in Applied Informatics to AMK and by NIH grants (GM46915, NS21328, HL30164, NS30164, AG10130, DK45215) to DCW. This work is being accomplished using the C programming language utilizing numerous libraries including one obtained from the National Center for Biotechnology Information (National Library of Medicine, USA) and is being integrated with existing database and WWW facilities.

References

[1] Brown, M.D., Torroni, A., Record, C.L., Wallace, D.C., "Phylogenetic analysis of Leber's hereditary optic neuropathy mitochondrial DNAs indicates multiple independent occurrences of the common mutations," *Human Mutation* 6:311–325, 1995.

[2] Brown M.D., Wallace, D.C., "Molecular basis of mitochondrial DNA disease," *Journal Bioenergetics and Biomembranes* 26:273–289, 1994.

[3] Kogelnik, A.M, Lott, M.T., Brown, M.D, Navathe SB, Wallace, D.C., "MITOMAP: a human mitochondrial genome database," *Nucleic Acids Research* 24(1):177–179, 1996.

[4] Kogelnik, A.M, Lott, M.T, Brown, M.D, Navathe SB, Wallace, D.C., "MITOMAP: an update on the human mitochondrial genome database," *Nucleic Acids Research* 25(1), 196–199, 1997.

[5] Wallace, D.C., "1994 William Allan Award Address: Mitochondrial DNA Variation in Human Evolution, Degenerative Disease, and Aging," *American Journal of Human Genetics* 57:201–223, 1995.

[6] Wallace, D.C., Brown, M.D., Lott, M.T., "Mitochondrial Genetics," in *Principles and Practice of Medical Genetics*, (ed. D. Rimoin, J.M. Connor, R.E. Pyeritz) 277–332, 1997.

[7] Wallace, D.C., Lott, M.T., Brown, M.D., "Report of the committee on human mitochondrial DNA," In Cuttichia, J. (ed) *Human Gene Mapping 1995: a compendium*, The Johns Hopkins University Press, Baltimore, pp. 910–954, 1995.