Identification of *Mycobacterium* Species Using Curated Custom Databases

Dan Kuyper[1], Hesham H. Ali[1], Amr M. Mohamed[2] and Steven H. Hinrichs[2]

[1]*Department of Computer Science*
University of Nebraska at Omaha
Omaha, NE 68182-0116
dkuyper@mail.unomaha.edu
hesham@unomaha.edu

[2]*Department of Pathology and Microbiology*
University of Nebraska Medical Center
Omaha, NE 68198-6495
amohamed@unmc.edu
shinrichs@unmc.edu

Abstract

Advances in molecular biology have resulted in the development of diagnostic tests for infectious diseases based on genetic profiles. While probe based assays dominate the field today, sequence based assays hold great promise for the future. However, the variability in quality of sequence information currently present in public databases limits the potential growth and use of sequence based analysis. To address this problem a standardized method for DNA sequence validation and building of custom databases was developed using *Mycobacterium* as a development model. With this model, a computational approach to identification of infectious diseases was developed and evaluated. The web-based application, termed BioDatabase, accomplished genetic sequence identification via the creation of curated databases containing a relatively small set of genetic data specific to a species or group. The process for creation of the custom database included multiple steps beginning with identification of highly conserved start and end sequences and intervening sequence validation parameters. The process eliminated the need for multiple sequence alignment with GenBank sequences, whose information is valuable, yet difficult to properly utilize due to its size and quality. The custom database approach maximized application performance with minimal impact on analysis response time, allowing investigation of optimal sequences for identification of all *Mycobacterium* to the species level. In comparison to the 16S and ITS genetic regions, a curated ITS based approach proved most effective for identification of *Mycobacterium* isolates.

1. Introduction

While the development of molecular assays utilizing specific probes has revolutionized the diagnosis of infectious diseases, even greater potential exists for approaches that determine the specific DNA sequence of microbiologic agents. To achieve this goal, a common region must be identified that contains a smaller signature sequence that is then matched with well characterized reference organisms. This approach is most cost effective when applied to infectious organisms that are difficult to grow in culture or grow at a very slow rate, such as members of the *Mycobacterium* genus that includes the species *M. tuberculosis*.

Tuberculosis is the most common infectious disease of humans world-wide [1]. Successful treatment is based on the accurate identification of *Mycobacterium* to the species level that directs the selection of appropriate antibiotic therapy. While partial sequences for a great number of microbiologic organisms that must be distinguished from *M. tuberculosis* have been placed in public databases, including GenBank, many of the sequences contain errors or the sequences were generated from non-reference isolates, therefore limiting their value. In addition, the number of simultaneous searches of public databases has increased dramatically within the past two years, resulting in progressively longer delays in obtaining search results. The creation of private databases has been proposed as an alternative approach. However, a standardized or consistent approach has not been adopted for review or validation of sequences within the custom or private database. To address theses issues, a process was created for the uniform analysis of sequence data for inclusion in a custom database. A model system was selected using organisms within the *Mycobacterium* genus that are pathogenic for humans. Multiple DNA sequence targets were considered for application of validation and analysis algorithms. These approaches were then incorporated into a web based analytic package to facilitate evaluation by microbiologists and computer scientists.

The BioDatabase package addresses these issues by:

1. Allowing researchers to create custom sets of genetic data suited to their specific needs. These data sets can incorporate information input by the

researcher, as well as information obtained from GenBank through an automated process.
2. Employing the techniques of optimal alignment algorithms. This technique allows researchers to identify sequences using proven algorithmic methods instead of heuristics.
3. Providing researchers the ability to specify fine-tuning parameters such as genetic region start patterns, end patterns, and other characteristics. This criterion is used to ensure integrity of data input to the custom database. It is also used to help validate unknown genetic sequences.
4. Giving researchers the ability to formulate sequence identification concepts and test their ideas against a validated database.

Our concluding case study highlights this capability. The case study involves *Mycobacterium* identification. In the case study, researchers were able to correctly identify 72 of 78 *Mycobacterium* isolates through new sequence identification techniques using the BioDatabase package. These results proved to provide better identification of *Mycobacterium* in most cases than existing techniques.

2. Problem Definition

Identification of unknown genetic sequences is one of the key problems facing biological researchers. This problem is complicated by the sheer size of data available and the tools available to analyze it. NCBI GenBank contains all known nucleotide and protein sequences with supporting bibliographic and biological information [2]. The data provided by GenBank is valuable, but not without pitfalls. For one, its sheer size makes certain operations, such as running optimal alignment algorithms over it, impossible due to time constraints. Therefore, heuristics such as BLAST and FASTA must be employed. A second pitfall is the quality of GenBank data. Although it attempts to control quality through certain mechanisms, it is impossible to ensure good or complete data due to sequencing errors in submitted information, improperly or ambiguously named sequences, or contamination due to sequences intentionally or accidentally inserted during cloning or recombination [3].

The most common tool used in genetic database searches today is BLAST. BLAST is a heuristic, finding the highest scoring local alignments between a query and sequence in a database [4]. Although BLAST is very fast, and is useful in many cases, some drawbacks exist. The biggest is the potential to generate biologically unimportant information. Since it is only a heuristic, researchers still must determine whether sequences constitute a true hit, making BLAST a good starting point, but not an end point in the sequence identification process.

Collaboration with medical researchers revealed the need for small, custom sets of data in biological and medical research work. Our goal was to allow researchers to create custom databases, searchable by algorithms rather than heuristics, useful for work in a number of different areas. These custom databases would allow for identification of sequences in day-to-day operations. They could also be used to formulate new techniques for identifying organisms. Solutions could then be developed to provide more rapid and accurate identification of pathogens for purposes in medical treatment and public health analysis.

3. The Proposed Application

The idea behind the BioDatabase application consists of creating small, high-quality sets of data researchers can work with for genetic sequence identification using optimal alignment algorithms. These custom databases could be specific to a species or subset of a species.

The project was inspired by work with medical researchers, who needed to analyze a relatively small set of sequences. Each of these sequences contained specific regions with highly conserved start and end patterns. These regions also shared other identification characteristics. Given the difficulty of sifting through the massive amounts of genetic information available at the GenBank, primarily due to the inaccuracy and the low quality of many of its sequences, BioDatabase allows sets of more accurate and specific data to be created by researchers to suit their specific sequence identification needs. These small sets of data can then be passed through optimal alignment algorithms, allowing for accurate identification of unknown sequences. Before searches take place, a researcher must create a custom database. This process consists of the following steps:

1. Creating and naming a database container.
2. Defining sequence regions. Each region must have a highly conserved start and end pattern.
3. Assigning characteristics to each region. Possible characteristics for a region include the following list, which could be expanded as needed:
    a. Threshold for wildcards (due to sequencing errors) that are allowed when adding or updating a custom database sequence.
    b. Threshold for wildcards (due to sequencing errors) that are allowed in an unknown sequence during the search process. Providing two thresholds for wildcards allows data residing in the custom database to remain of high quality while allowing unknown

sequences searched against to be of lower quality. BioDatabase does not count wildcards as a scoring or gap penalty in its Similarity and Align algorithms.

c. Characters constituting wildcards. Nucleotides with this character, such as 'N', are counted as a wildcard.

d. Limit of character runs. This helps prevent patterns such as 'AAAAAAA', which are most likely sequencing errors, from making it into the custom database.

4. Adding sequences to the custom database, either manually and/or through automated GenBank retrieval. We describe the process of retrieving information from GenBank in more detail below. Each candidate sequence for the custom database must first pass all validation conditions before it can be added.

3.1 Retrieving Information from GenBank

BioDatabase attempts to mitigate the problems associated with GenBank while still allowing its data to be incorporated for searches. This is done as follow:

1. Selecting taxonomic classifications from the Entrez Taxon database.
2. Retrieving GenBank sequences for selected taxonomic classifications.
3. Validating retrieved sequences against region criteria for the custom database.

NCBI Entrez integrates genetic database information into a central search and retrieval system. These databases include PubMed, Nucleotide and Protein Sequences, Protein Sequences, Protein Structures, Complete Genomes, Taxonomy, OMIM, and many others [5]. Entrez's taxonomy database, Taxon, is a curated set of names and classifications for all organisms in GenBank [5]. Each entry in the Taxon database is assigned a unique identifier, called its tax_id, and a single scientific name. Tax_ids may also have several synonyms (such as common names). Each Taxon entry includes a parent_tax_id, indicating its parent in the phylogenetic tree. Taxon also contains a cross-reference table allowing sequences in GenBank to be referenced by gi_number.

BioDatabase creates a representation of Taxon's phylogenetic tree in a nested-set format, as proposed in [6, 7]. In this model, instead of representing parent-child relationships explicitly, we use two pointers (called left_id and right_id) to provide bounds for a classification. In this representation, each child node's left_id and right_id must be between its parent's left_id and right_id. Using a nested-set allows the BioDatabase application to quickly scan the Taxon tree for parent-child relationships. Using an adjacency list representation, as is the default representation of Taxon, would require more of an intensive approach.

The BioDatabase interface allows researchers to quickly scan Taxon's phylogenetic tree, selecting classifications of interest to them. These classifications are then associated with this custom database. An automated process then uses Taxon's cross-reference table to gather gi_numbers associated with this custom database based on the tax_id(s) selected. Each gi_number represents a candidate database member. BioDatabase then accesses GenBank, retrieving sequence information for each gi_number. The automated process parses the GenBank information, retrieving sequence data. The sequence is then passed through the custom database's validation conditions. If the validation is successful, the sequence is added to the custom database. If the validation conditions fail, the sequence is discarded.

3.2 Searching the Custom Database

After the custom database has been constructed, searches may take place against it. Searching a custom database involves the following steps:

1. Entering unknown sequence information.
2. Selecting custom database sequence regions. This allows researchers to search against one, many, or all custom database regions, depending on the nature of their input sequence.
3. Validating input sequence against custom database conditions. The validation process only occurs for the regions selected in step 2.
4. Returning an error message if the input sequence fails validation conditions. Errors in input sequences must be rectified before the search process can be complete.
5. Computing similarity scores for each selected region against regions for each active sequence in the custom database if the input sequence is valid.
6. Sorting similarity scores from highest to lowest.
7. Outputting results and allowing researcher to view region alignments.

3.3 Algorithms

The BioDatabase application employs a modified version of the similarity algorithm explained in [8] to calculate how alike an unknown sequence's region is with a custom database sequence's region. The modified version of the similarity algorithm takes into account the possibility of wildcards or ambiguous nucleotides in either sequence. Wildcards are not counted as penalties in the scoring process.

The BioDatabase application utilizes a modified version of the Align algorithm described in [8] to show where dissimilarities occurred between an unknown sequence's region and a custom database sequence's region. The Align algorithm returns a color-coded string to display the differences and takes into account wildcard characters in either the input string or the canonical database string. Also, spaces are not inserted where mismatches occur at wildcard characters.

3.3.1 Validation Algorithm

The BioDatabase application employs a Validation algorithm to ensure each sequence in a custom database meets specified criteria. The Validation algorithm works as follows:

Input: Genetic sequence s, Database id d, List of regions R, List of region preferences P

Output: Message M indicating success or failure of the validation

Variables, Methods and Parameters:

- GetRegions (d) gets the list of regions for the database
- GetRegionPreferences (region) gets preferences for a region
- FindStartOfRegion(s, region) finds index for start sequence of region
- FindEndOfRegion(s, region) finds index for end sequence of region
- FindCharacterRuns(s, region) finds index of long character runs for region
- CheckWildcardCount(s, region) finds if input string is over limit in wildcards
- Start index for region Si
- End index for region Ei
- Flag (Y/N) indicating sequence has character runs CR
- Flag (Y/N) indicating sequence exceeds wildcard limit WL

```
Validation
{
  R ← GetRegions (d)
  For each region in R
    P ← GetRegionPreferences (region)
    Si = FindStartOfRegion(s, region)
    If Si == -1 then  M ← "Missing start region"
    Ei = FindEndOfRegion(s, region)
    If Ei == -1 then  M ← "Missing end region"
    If Ei >= Si then  M ← "End region proceeds start region"
    If characterRunLimit in P
      CR ← FindCharacterRuns(s, region)
      If CR == "Y" then  M ← "Character runs exist"
    If wildcardCharacterLimit in P
      WL ← CheckWildcardCount(s, region)
      If WL == "Y" then M ← "Too many wildcards"
    M ← "Sequence Okay"
}
```

4. Case Study: *Mycobacterium* Database

This case study utilizes the BioDatabase application to create an automated computational approach for *Mycobacterium* identification using the 16S and ITS (Internal Transcribed Spacer) sequence regions. Researchers wanted to test the theory that the ITS region could provide a better identification marker for *Mycobacterium* identification than the 16S region. Achieving this goal would lead to better identification of *Mycobacterium* in a more timely and cost effective manner, subsequently providing better outcome for those infected with a pathogenic species. In addition, epidemiological researchers would be better able to identify and track public health outbreaks.

Genus *Mycobacterium* comprises more than 70 species of acid fast bacilli, of which 30 different species have been associated with a wide variety of human and animal diseases [9]. During the last two decades, problems arising from infection with tuberculosis, as well as non-tuberculosis *Mycobacterium*, have become increasingly important [10]. Diseases caused by *Mycobacterium* are major contributors to morbidity and mortality throughout the world. Their impact has increased with the rise of HIV infections, primarily due to *M.tuberculosis* complex and *M.avium* complex [1]. In 1995, the World Health Organization (WHO) estimated that 3.3 million people died from *M.tuberculosis* (Tuberculosis or TB) infection, making it the leading cause of death among adults by a single infectious agent. The WHO estimates that in the next twenty years, over a billion people will become infected with *Mycobacterium*, specifically *M.tuberculosis*. Of this, 200 million will develop symptoms and 35 million will die, mostly in developing countries.

In humans, three main groups of *Mycobacterium* are responsible for the majority of diseases. The first group is *Mycobacterium tuberculosis* complex (mainly *M.tuberculosis* and *M.bovis*): *M.tuberculosis* and *M.bovis* species cause classical tuberculosis in humans. *M.bovis* primarily infects cattle, but can also infect humans through ingestion of milk (mainly among children) or airborne inhalation. The second group is *Mycobacterium avium* complex (MAC): *M.avium* infection is a tuberculosis-like disease that is common among AIDS patients, especially those with advanced disease. The Third one is Non-tuberculosis *Mycobacterium* (NTM); Non-tuberculosis *Mycobacterium* cause severe diseases among

immunosuppressed persons, manifested in the form of skin lesions, pulmonary diseases, and soft tissue (internal organ) lesions [11]. Identification of *Mycobacterium* to the species level is of clinical significance since not all species are of equal clinical importance. Besides, certain drugs are effective only against specific species [12].

Identification of *Mycobacterium* using conventional methods is a slow and tedious laboratory procedure, frequently requiring several weeks for adequate growth and identification. In addition, accurate identification is not always possible by conventional methods. Ambiguous results or errors from conventional methods arise from factors such as lack of adequate growth, contamination, and phenotypic variability [12].

Sequencing of specific genetic elements in *Mycobacterium* provides an alternative to conventional laboratory tests, allowing for rapid and accurate identification of *Mycobacterium* to the genus level. At least three different genes have been reported as useful targets for sequencing to distinguish *Mycobacterium*. They are the 16S rRNA gene, the hsp65 gene, and the recA gene [13]. Most approaches used to identify *Mycobacterium* and to establish phylogenetic relationships have focused on the sequence of the 16S rRNA gene [14]. The fact this gene serves a vital function in *Mycobacterium* makes the extent of its permissible mutations inherently limited [13]. As a result, many species have an identical or highly homogenous 16S rRNA sequence. The ITS (between the 16S and 23S of the ribosomal gene) region has recently been reported as a possible genetic element that can provide for *Mycobacterium* identification [15]. Analysis of the ITS sequence has shown the presence of highly variable regions within constant sequences that provided for optimal usage as a target for phylogenetic analysis and differential identification of *Mycobacterium* species [16]. In the current study, we use the hyper-variable region of ITS sequence to create a custom database for the purpose of rapid and accurate diagnosis and identification of members of genus *Mycobacterium*. For comparison purposes, we include the previously described 16S rRNA sequence (commercially available for *Mycobacterium* identification) in our custom database to prove the superiority of the chosen ITS region over the 16S region in terms of *Mycobacterium* identification.

4.1 BioDatabase for Mycobacterium Species Identification

The custom database for *Mycobacterium* includes two regions, one for 16S and one for ITS. The 16S region is defined by the start sequence "GTCGAACGG" and the ending sequence "GGCCAACTACGT". The ITS region is defined by the start sequence "CACCTCCTTTCT" and the ending sequence "GGGGTGTGG". Both regions contain identical preferences. The wildcard for both regions is 'N'. The character-run limit is set to 6. Next, sequences for the custom database were entered. Searches over both the ITS and 16S regions were then performed with a sample set of 78 specimens, previously identified using laboratory techniques. Figure 1 shows the flow control of the BioDatabase application for this case study.

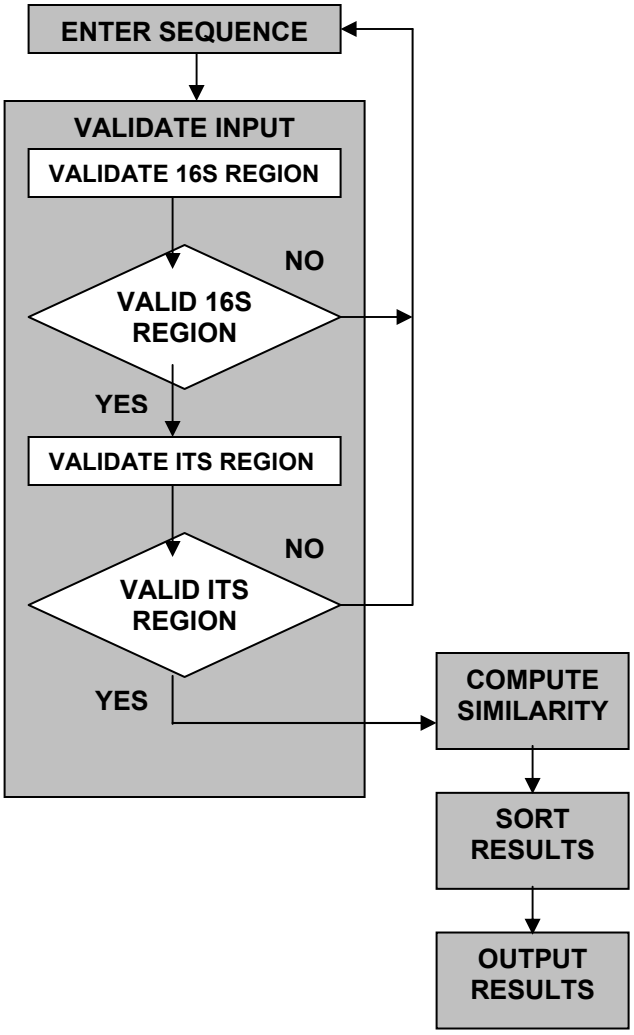

Figure 1: Flow Control Diagram for *Mycobacterium* Identification

4.2 Results

Results showed the ability of the BioDatabase application, using the ITS target sequence, to accurately identify members of genus *Mycobacterium* not only to the species level, but also to the strain level. Out of 78 previously identified isolates, 72 were correctly identified using

BioDatabase. The remaining 6 sequences failed to match with any of the sequences of our database. This could be attributed to the hypervariable nature of the ITS sequence that can differentiate not only species of genus *Mycobacterium*, but also provides for the differentiation between different strains of the same species. Accordingly, these 6 unmatched sequences could potentially be new strains. Further clinical testing will be required to confirm this possibility. The results also proved the superiority of the selected genetic target (ITS sequence) over the widely used target (16S rRNA). This was achieved by the ability of the BioDatabase system to accurately differentiate between certain *Mycobacterium* species using the ITS genetic target database, but not the 16S rRNA database. For example, species like *M. chelonae, M. abscessus,* and *M. furth,* as well as *M. gastri* and *M. kansasii,* were found to have similar sequences at the level of 16S rRNA target. On the other hand, at the level of ITS target, the sequences of these species were different and allowed for their identification and differentiation (data not shown).

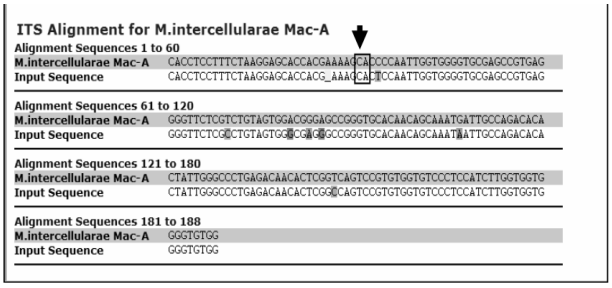

Figure 2: Results from BioDatabase Application

These results allowed researchers to confirm the validity of using the ITS region as a potential identification marker for *Mycobacterium* species. In addition, our results, as shown in Figures 2 and 3, demonstrate the ability of the BioDatabase application to accurately identify sequences as opposed to the Genbank dependent BLAST search. Using the BioDatabase application, as shown in Figure 2, the closest match to the unknown sequence, was identified as strain MAC-A (which was consistent with the conventional biochemical tests). On the other hand, when using BLAST against GenBank as shown in Figure 3, the unknown sequence was identified as *M.malmoense*. The second highest match from BLAST was MAC-A. This was the result of the presence of inaccurate and low quality sequences in the GenBank. In our case, the presence of two ambiguous bases (h,n) in the GenBank sequence, but not in our accurate custom database sequence (highlighted in Figures 2 and 3), provided for the lower score of matching sequence rather than first match to the unknown sequence. This example not only illustrates the inherent problems with the amount and quality of data in Genbank, but also the pitfalls of heuristics such as BLAST.

The overall results in our case study highlight the ability of the BioDatabase application to be customized for sequence identification techniques. In our case study, a database utilizing the 16S and ITS regions of *Mycobacterium* was capable of generating more accurate results than use of existing procedures. The principles employed in our case study could be very easily applied to other species or sequence regions of interest.

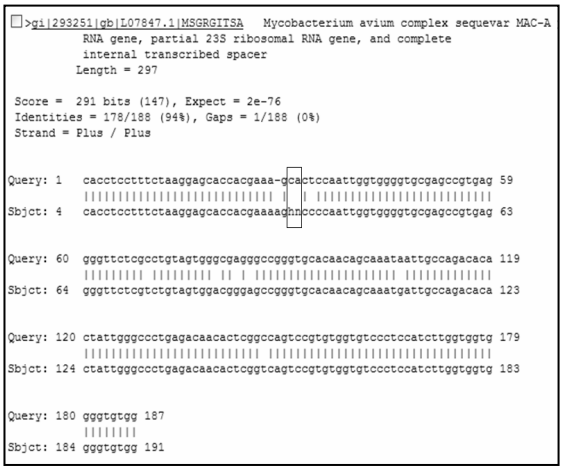

Figure 3: Results from GenBank using BLAST search

The case study illustrates the main advantages of the BioDatabase application:

1. Sequences can be rapidly and accurately identified by specifying regions and fine-tuning parameters.
2. The quality of the output is improved by use of a validated database.
3. Optimal alignment algorithms, rather than heuristics, improved the accuracy of the search.

5. Conclusions

BioDatabase provides a means to perform comprehensive and reliable searches of sequences using a relatively small set of genetic data. BioDatabase has the advantage of enabling researchers to control the quality of their sequence data by defining regions, and by setting specific restrictions for those regions. In comparison with existing approaches such as BLAST, BioDatabase provides more accurate results. This is due to the fact it employs optimal alignment algorithms rather than the heuristics used in

BLAST. Furthermore, the quality of data in custom databases is more reliable than what is available in GenBank, but it can still incorporate GenBank data. BioDatabase also has the advantage of providing completely customizable restriction parameters, helping to ensure an accurate dataset. In addition, output results from the analysis are generated in a simple, easy to interpret format.

6. References

1. B. B. Plikaytis, B. D. Plikaytis, M. A. Yakrus, W. R. Butler, C. L. Woodley, and V. A. Silcox. 1992. Differentiation of Slowly Growing *Mycobacterium* Species, Including *Mycobacterium* tuberculosis, by Gene Amplification and Restriction Fragment Length Polymorphism Analysis. Journal of Clinical Microbiology. 30:1815-1822.
2. D. A. Benson, I. Karsch-Mizachi, D. J. Lipman, J. Ostell, B. A. Rapp, and D. L. Wheeler (2000) GenBank Nucleic Acids Research. 28 (1) :15-18.
3. P. Bork, and A. Bairoch (1996) Go Hunting in Sequence Databases But Watch Out for the Traps Trends Genet 12(10):425-427.
4. S. F. Altschul, Gish, W., Miller, W., Myers, E.W. and D.J. Lippman (1990) Basic Local Alignment Search Tool. J. Mol. Biol. 215: 403-410.
5. The NCBI Handbook, NCBI, 2002.
6. A. Mackey. 2002. Relational Modeling of Biological Data: Trees and Graphs. O'Reilly Bioinformatics Technology Conference. November 27th , 2002.
7. J. Celko, SQL For Smarties: Advanced SQL Programming. 2000 Morgan Kaufman Publishers.
8. J. Setubal, and J. Meidanis, "Introduction to Computational Molecular Biology". PWS Pub, 1997.
9. T. M. Shinnick, and R. C. Good. 1994. Mycobacterial taxonomy. Eur J Clin Microbiol Infect Dis 13:884-901.
10. E. C. Bottger. 1994. *Mycobacterium* genavense: an emerging pathogen. Eur J Clin Microbiol Infect Dis 13:932-936.
11. A. M. Mohamed. 2002. Tuberculosis: A molecular approach to diagnosis and drug resistance. Presentation at University of Nebraska Medical Center. May, 2002.
12. M. Mondragon-Barrator, Vazquez-Chacon CA, Barron-Rivero C, Acosta-Blanco P, Jost KC, Balandrano S, Olivera-Diaz H. 2000. Comparison among three methods for mycobacteria identification. Salud Publica de Mexico 2000: 42:484-489.
13. T. Rogall, J. Wolters, T. Flohr, and E. C. Bottger. 1990. Towards a phylogeny and definition of species at the molecular level within the genus *Mycobacterium*. Int J Syst Bacteriol 40:323-330.
14. J. L. Cloud, H. Neal, R. Rosenberry, C. Y. Turenne, M. Jama, D. R. Hillyard, and K. C. Carroll. 2002. Identification of *Mycobacterium* spp. by using a commercial 16S ribosomal DNA sequencing kit and additional sequencing libraries. J Clin Microbiol 40:400-406.
15. R. Frothingham, and K. H. Wilson. 1994. Molecular phylogeny of the *Mycobacterium* avium complex demonstrates clinically meaningful divisions. J Infect Dis 169:305-312.
16. K. A. De Smet, I. N. Brown, M. Yates, and J. Ivanyi. 1995. Ribosomal internal transcribed spacer sequences are identical among *Mycobacterium* avium-intracellulare complex isolates from AIDS patients, but vary among isolates from elderly pulmonary disease patients. Microbiology 141 (Pt 10):2739-2747.