4. MATERIALS AND METHODS
4.1 NCBI

The National Center for Biotechnology Information (NCBI) is part of the United States National Library of Medicine (NLM), a branch of the National Institutes of Health. It is located in Bethesda, Maryland and was founded in 1988 through legislation sponsored by Senator Claude Pepper. The NCBI houses a series of databases relevant to biotechnology and biomedicine. Major databases include GenBank for DNA sequences and PubMed, a bibliographic database for the biomedical literature. NCBI is directed by David Lipman, one of the original authors of the BLAST sequence alignment program and a widely respected figure in bioinformatics.

The NCBI has had responsibility for making available the GenBank DNA sequence database since 1992. GenBank coordinates with the European Molecular Biology Laboratory (EMBL) and the DNA Data Bank of Japan (DDBJ). Since 1992, NCBI has grown to provide other databases in addition to GenBank. NCBI provides Online Mendelian Inheritance in Man, the Molecular Modeling Database, dbSNP, the Reference Sequence Collection, a map of the human genome, a taxonomy browser, and coordinates with the National Cancer Institute to provide the Cancer Genome Anatomy Project. The NCBI assigns a unique identifier to each species of organism. The NCBI has software tools that are available online. For example, BLAST is a sequence similarity searching program. This online software can do sequence comparisons against the GenBank DNA database in less than 15 seconds. As a national resource for molecular biology information, its mission is to develop new information technologies to aid in the understanding of fundamental molecular and genetic processes that control health and disease. More specifically, the NCBI has been charged with creating automated systems for storing and analysing knowledge about molecular biology, biochemistry, and genetics; facilitating the use of such databases and software by the research and medical community; coordinating efforts to gather biotechnology information both nationally and internationally; and performing research into
advanced methods of computer-based information processing for analysing the structure and function of biologically important molecules. To carry out its diverse responsibilities,

1. It conducts research on fundamental biomedical problems at the molecular level using mathematical and computational methods

2. Maintains collaborations with several NIH institutes, academia, industry, and other governmental agencies

3. Fosters scientific communication by sponsoring meetings, workshops, and lecture series

4. Supports training on basic and applied research in computational biology for postdoctoral fellows through the NIH Intramural Research Program

5. Engages members of the international scientific community in informatics research and training through the Scientific Visitors Program

6. Develops, distributes, supports, and coordinates access to a variety of databases and software for the scientific and medical communities

7. Develops and promotes standards for databases, data deposition and exchange, and biological nomenclature

4.2 GLIMMER
In bioinformatics Gene Locator and Interpolated Markov ModelER, (GLIMMER) was the first system for finding genes that used the interpolated Markov model formalism. It is effective at finding genes in bacteria, archaea, and viruses, typically finding 98–99% of all protein-coding genes. The GLIMMER software is open source. Because of its high accuracy, Glimmer is the system of choice for genome annotation efforts on a wide range of bacteria, archaeal, and viral species. Glimmer was used by the DNA Databank of Japan (DDBJ) to re-annotate all bacterial genomes in the International Nucleotide Sequence Databases. It is also
being used to annotate viruses. Glimmer is part of the bacterial annotation pipeline at the National Center for Biotechnology Information (NCBI), which also maintains a web server for Glimmer, as do sites in Germany, Canada, and elsewhere. Glimmer is a highly cited bioinformatics system in the scientific literature. According to Google Scholar, as of early 2011 the original Glimmer article (Salzberg et al., 1998) has been cited 581 times and the Glimmer 2.0 article (Delcher et al., 1999) has been cited 950 times. Glimmer uses interpolated Markov models (IMMs) to identify the coding regions and distinguish them from noncoding DNA. The IMM approach uses a combination of Markov models from 1st through 8th-order, weighting each model according to its predictive power. Glimmer uses 3-periodic nonhomogenous Markov models in its IMMs. Glimmer was the primary microbial gene finder used at The Institute for Genomic Research (TIGR), where it was first developed, and has been used to annotate the complete genomes of over 100 bacterial species from TIGR and other labs. Glimmer3 predictions are available for all NCBI RefSeq bacterial genomes at their ftp site.

4.3 **G-C % content calculator**
GC-content or guanine-cytosine content is the percentage of nitrogenous bases on a DNA molecule that are either guanine or cytosine. This may refer to a specific fragment of DNA or RNA, or that of the whole genome. When it refers to a fragment of the genetic material, it may denote the GC-content of part of a gene (domain), single gene, group of genes (or gene clusters), or even a non-coding region. Guanine and Cytosine undergo a specific hydrogen bonding, whereas Adenine bonds specifically with Thymine.

The GC pair is bound by three hydrogen bonds, while AT pairs are bound by two hydrogen bonds. DNA with high GC-content is more stable than DNA with low GC-content; however, the hydrogen bonds do not stabilize the DNA significantly, and stabilization is due mainly to stacking interactions. In spite of the higher thermostability conferred to the genetic material, it is envisaged that cells with DNA of high GC-content undergo autolysis. Due to the robustness endowed to
the genetic materials in high GC organisms, it was commonly believed that the GC content played a vital part in adaptation temperatures, a hypothesis that has recently been refuted. However, the same study showed a strong correlation between higher temperatures and the GC content of structured RNAs; GC base pairs are more stable than AU base pairs, due to the fact that GC bonds have 3 hydrogen bonds and AU only has 2 hydrogen bonds, which makes high-GC-content RNA structures more tolerant of high temperatures. More recently, the first large-scale systematic gene-centric association analysis demonstrated the correlation between GC content and temperature for certain genomic regions while not for others.

In PCR experiments, the GC-content of primers are used to predict their annealing temperature to the template DNA. A higher GC-content level indicates a higher melting temperature. The GC-content percentages as well as GC-ratio can be measured by several means, but one of the simplest methods is to measure what is called the melting temperature of the DNA double helix using spectrophotometer. The absorbance of DNA at a wavelength of 260 nm increases fairly sharply when the double-stranded DNA separates into two single strands when sufficiently heated. In alternative manner, if the DNA or RNA molecule under investigation has been sequenced then the GC-content can be accurately calculated by simple arithmetic.

$$GC\% = \frac{G + C}{G + C + A + T} \times 100$$

4.4 ClustalW2
ClustalW is a tool to align three or more sequences together in a computationally efficient manner. Aligning multiple sequences highlights areas of similarity which may be associated with specific features that have been more highly conserved than other regions. These regions in turn can help classify sequences or to inform experiment design. Multiple sequence alignment is also an important step for
phylogenetic analysis, which aims to model the substitutions that have occurred over evolution and derive the evolutionary relationships between sequences.

The most important tool in studying the protein sequences is by multiple alignments. This identifies the conserved sequence regions. This tool is helpful in determining the structure, predicts the function and identifies new proteins. Both global and local alignment is possible. In case of local alignments gaps, which represent deletion or insertion can be avoided whereas in global alignments we need to use these gaps. This computational program is fully automatic one specifically designed for global multiple alignment sequences of protein and DNA. All the parameters can be adjusted in this program. Multiple alignments can also be used to calculate the phylogenetic trees. ClustalW2 attains the multiple alignments of vast sequences biologically. The closest sequence match is computed by the program and orders them by which we can identify differentiate and find similarities between them. Cladograms or Phylograms can be used to determine the evolutionary relationships.

There are a total of seven file formats that the ClustalW2 software automatically accepts, these are NBRF-PIR, ENBL-SWISSPROT, Pearson (Fasta), Clustal (*.aln), GCG9-RSF, GCG-MSF (Pileup) and GDE flat file. There were few limitations of ClustalW2 software that when the program is run, command prompt allowed to just interpret the last few rows and the most significant drawback is that the output cannot be saved. To overcome this kind of problem a code was developed, which is given below:-
There is a choice between two alignments that is the Slow accurate method which is a very slow process. The parameters used in this alignment do not affect the speed of the alignment and give out the percent identity scores. These scores are displayed on the screen. The second alignment is the Fast-Approximate alignment in which four parameters control are used to control it. The first parameter in the Window size in which to increase the sensitivity we can decrease the speed, the second parameter in the Gap penalty which has very less affect on the sensitivity or speed apart from the extreme values, the third one is the Top diagonals in this only the best matches are used in the alignment and the fourth parameter is the K-Tuple size, in which the exact size is used with respect to the matching fragment.

4.5 GENEWIZ
The GeneWiz browser is an interactive web application for visualizing genomic data of sequenced prokaryotic chromosomes. It allows users to carry out various analyses such as mapping alignments of homologous genes to other genomes, mapping of short sequencing reads to a reference chromosome, and calculating DNA properties such as curvature or stacking energy along the chromosome. It produces an interactive graphic that enables zooming from a global scale down to single nucleotides, without changing the size of the plot. Its ability to disproportionally zoom provides optimal readability and increased functionality compared to other browsers. The tool allows the user to select the display of various genomic features, color setting and data ranges. Custom numerical data
can be added to the plot allowing, for example, visualization of gene expression and regulation data. Further, standard atlases are pre-generated for all prokaryotic genomes available in GenBank, providing a fast overview of all available genomes, including recently deposited genome sequences.

**4.6 Microsoft Excel**

Microsoft Excel is a spreadsheet application developed by Microsoft for Microsoft Windows and Mac OS X. It has been a very widely applied spreadsheet for these platforms, especially since version 5 in 1993, and it has replaced Lotus 1-2-3 as the industry standard for spreadsheets. Excel forms part of Microsoft Office.

Microsoft Excel has the basic features of all spreadsheets, using a grid of cells arranged in numbered rows and letter-named columns to organize data manipulations like arithmetic operations. It has a battery of supplied functions to answer statistical, engineering and financial needs. In addition, it can display data as line graphs, histograms and charts, and with a very limited three-dimensional graphical display. It allows sectioning of data to view its dependencies on various factors for different perspectives. It has a programming aspect, Visual Basic for Applications, allowing the user to employ a wide variety of numerical methods and then reporting the results back to the spreadsheet. It also has a variety of interactive features allowing user interfaces that can completely hide the spreadsheet from the user, so the spreadsheet presents itself as a so-called application, or decision support system (DSS), via a custom-designed user interface. In a more elaborate realization, an Excel application can automatically poll external databases and measuring instruments using an update schedule, analyse the results, make a Word report or Power Point slide show, and e-mail these presentations on a regular basis to a list of participants.
4.7 GENE CLUSTER
An increasing body of literature shows that genomes of eukaryotes can contain clusters of functionally related genes. Most approaches to identify gene clusters utilize microarray data or metabolic pathway databases to find groups of genes on chromosomes that are linked by common attributes. A generalized method that can find gene clusters regardless of the mechanism of origin would provide researchers with an unbiased method for finding clusters and studying the evolutionary forces that give rise to them. We present an algorithm to identify gene clusters in eukaryotic genomes that utilizes functional categories defined in graph-based vocabularies such as the Gene Ontology (GO). Clusters identified in this manner need only have a common function and are not constrained by gene expression or other properties. We tested the algorithm by analyzing genomes of a representative set of species. We identified species-specific variation in percentage of clustered genes as well as in properties of gene clusters including size distribution and functional annotation. These properties may be diagnostic of the evolutionary forces that lead to the formation of gene clusters. Computational approaches to identify gene clusters are usually aimed at identifying specific cluster types, such as those that correspond to metabolic pathways or those that represent sets of co-expressed genes. A generalized approach that can identify all clusters in a genome would be of great value for the study of eukaryotic genome organization and evolution. In addition, identification of gene clusters may help to identify functional relationships among genes, and aid in the discovery of metabolic pathways and protein interactions.

4.8 GENEOMIC ISLANDS
Genomic Islands (GIs) are genomic regions that are originally from other organisms, through a process known as Horizontal Gene Transfer (HGT). Detection of GIs plays a significant role in biomedical research since such align genomic regions usually contain important features, such as pathogenic genes. This is a use friendly graphic user interface, Genomic Island Suite of Tools (GIST), which is a platform for scientific users to predict GIs. This software
package includes five commonly used tools, AlienHunter, IslandPath, Colombo SIGI-HMM, INDeGenIUS and Pai-Ida. It also includes an optimization program EGID that ensembles the result of existing tools for more accurate prediction. The tools in GIST can be used either separately or sequentially. GIST also includes a downloadable feature that facilitates collecting the input genomes automatically from the FTP server of the National Center for Biotechnology Information (NCBI). GIST was implemented in Java, and was compiled and executed on Linux/Unix operating systems. Some prokaryotic genomes contain genomic sequences with different patterns than the remaining parts of the host genomes. Such differences may include GC content bias, codon usage bias, k-mer nucleotide frequency bias, and the existence of mobile genes such as integrase genes and transposes genes. In some other cases, such regions are also bordered by transfer RNAs (t-RNA). The abnormal regions that contain such types of characteristics are known as Genomic Islands (GIs). Research on identifying genomic islands has become more important as the scientific community can be significantly benefitted from such findings. Biomedical researchers and microbiologists can use the results to explain the pathogenicity of organisms, or discover industrial important metabolic components from GIs. Based on such findings, pharmacists can use them to design corresponding vaccines and antibiotics, and eventually promote pharmaceutical companies to produce medicines at a large scale.

4.9 GIST

Genomic Islands (GIs) are genomic regions that are originally from other organisms, through a process known as Horizontal Gene Transfer (HGT). Detection of GIs plays a significant role in biomedical research since such align genomic regions usually contain important features, such as pathogenic genes. We have developed a user-friendly graphical user interface, Genomic Island Suite of Tools (GIST), which is a platform for scientific users to predict GIs. This software package includes five commonly used tools, AlienHunter, IslandPath, Colombo SIGI-HMM, INDeGenIUS and Pai-Ida. It also includes an optimization
program EGID that ensembles the result of existing tools for more accurate prediction. The tools in GIST can be used either separately or sequentially. GIST also includes a downloadable feature that facilitates collecting the input genomes automatically from the FTP server of the National Center for Biotechnology Information (NCBI). GIST was implemented in Java, and was compiled and executed on Linux/Unix operating systems.

Some prokaryotic genomes contain genomic sequences with different patterns than the remaining parts of the host genomes. Such differences may include GC content bias [1], codon usage bias [2, 3], k-mer nucleotide frequency bias [4], and the existence of mobile genes such as integrase genes and transposes genes [5]. In some other cases, such regions are also bordered by transfer RNAs (t-RNA) [6]. The abnormal regions that contain such types of characteristics are known as Genomic Islands (GIs). Research on identifying genomic islands has become more important as the scientific community can be significantly benefitted from such findings. Biomedical researchers and microbiologists can use the results to explain the pathogenicity of organisms, or discover industrial important metabolic components from GIs. Based on such findings, pharmacists can use them to design corresponding vaccines and antibiotics, and eventually promote pharmaceutical companies to produce medicines at a large scale. As it is generally believed that each genome contains unique genomic sequence signature, some computation tools based on sequence signature have been developed. Such sequence composition based tools include AlienHunter [7], COLOMBO SIGI-HMM [8], GiDetector [9], IslandPath [10], INDeGenIUS [11], and PAI-IDA [12]. Recent studies have shown that none of these tools can predict GIs accurately in all genomes [13]. Hence it necessary to develop a computational framework that produces a better prediction results by combining the results of existing programs [14]. We have recently developed a tool, EGID, which has shown to optimize the results of individual tools, and produce a better prediction result for all genomes [15].
4.10 Conserved Domain Database for Function Prediction

Domains can be thought of as distinct functional and/or structural units of a protein. These two classifications coincide rather often, as a matter of fact, and what is found as an independently folding unit of a polypeptide chain also carries specific function. Domains are often identified as recurring (sequence or structure) units, which may exist in various contexts. The image below illustrates four "domains" identified as structural units in the MMDB-entry 1IGR, chain A, as segments colored in magenta, blue, brown, and green.

In molecular evolution such domains may have been utilized as building blocks, and may have been recombined in different arrangements to modulate protein function. We define conserved domains as recurring units in molecular evolution, the extents of which can be determined by sequence and structure analysis.

Conserved domains contain conserved sequence patterns or motifs, which allow for their detection in polypeptide sequences. The distinction between domains and motifs is not sharp, however, especially in the case of short repetitive units. Functional motifs are also present outside the scope of structurally conserved domains. The CD database is not meant to systematically collect such motifs.

Multiple sequence alignments provide basis for conserved domain models

The two types of domains shown in the 1IGR illustration above -- 3D domains and conserved domains (or "domain families") -- often coincide with each other. However, because they represent two distinct types of data -- 3D structures and protein sequences, respectively -- they reside in two distinct databases: the Entrez Structure (Molecular Modeling Database, MMDB) and the Conserved Domain Database (CDD). The former includes the spatial (X,Y,Z) coordinates of each atom in a structure (where 3D domains are identified algorithmically), while the latter shows the span and composition of a conserved protein sequence region.
Specifically, conserved domain models are based on **multiple sequence alignments of related proteins spanning a variety of organisms** to reveal sequence regions containing the same, or similar, patterns of amino acids. The illustration below provides an example, showing the multiple sequence alignment for the Furin-like domain, which is present in the Type 1 Insulin-like Growth Factor Receptor (1IGR) protein. Click anywhere on the image to open the complete, interactive CDD record for that domain model, cd00064. A separate section of this help document provides additional information about multiple sequence alignment display options.

In the CDD database, protein sequences from three-dimensional structures are included in domain models whenever possible, as one goal of the NCBI conserved domain curation effort is to make multiple sequence alignments agree with what we can infer from three-dimensional structure and three-dimensional structure superposition, in order to understand sequence/structure/function relationships. The sequence-based domain models and corresponding 3D structures are also cross-referenced to each other through Entrez "Links" between CDD and structure records.