Design and Development of a Database for the Classification of *Corynebacterium glutamicum* Genes, Proteins, Mutants and Experimental Protocols

Ashfaq Muhammad

Master's dissertation

University of Skövde

31 October 2006

Design and Development of a Database for the Classification of *Corynebacterium glutamicum* Genes, Proteins, Mutants and Experimental Protocols

Ashfaq Muhammad

Submitted by Ashfaq Muhammad to the University of Skövde as dissertation towards the degree of Master by examination and dissertation in the School of Humanities and Informatics.

31 October 2006

I certify that all material in this thesis which is not my own work has been identified and that no material is included for which a degree has previously been conferred on me.

____

Ashfaq Muhammad

Design and Development of a Database for the Classification of *Corynebacterium glutamicum* Genes, Proteins, Mutants and Experimental Protocols

Ashfaq Muhammad

University of Skövde,
S-541 28 Skövde, Sweden
a03ashmo@student.his.se

Abstract. Coryneform bacteria are largely distributed in nature and are rod like, aerobic soil bacteria capable of growing on a variety of sugars and organic acids. *Corynebacterium glutamicum* is a non-pathogenic species of Coryneform bacteria used for industrial production of amino acids. There are three main publicly available genome annotations, Cg, Cgl and NCgl for *C. glutamicum*. All these three annotations have different numbers of protein coding genes and varying numbers of overlaps of similar genes. The original data is only available in text files. In this format of genome data, it was not easy to search and compare the data among different annotations and it was impossible to make an extensive multidimensional customized formal search against different protein parameters. Comparison of all genome annotations for construction deletion, over-expression mutants, graphical representation of genome information, such as gene locations, neighboring genes, orientation (direct or complementary strand), overlapping genes, gene lengths, graphical output for structure function relation by comparison of predicted trans-membrane domains (TMD) and functional protein domains protein motifs was not possible when data is inconsistent and redundant on various publicly available biological database servers. There was therefore a need for a system of managing the data for mutants and experimental setups. In spite of the fact that the genome sequence is known, until now no databank providing such a complete set of information has been available. We solved these problems by developing a standalone relational database software application covering data processing, protein-DNA sequence extraction and management of lab data. The result of the study is an application named, CORYNEBASE, which is a software that meets our aims and objectives.

1- Background

1.1 Biology of *Corynebacterium glutamicum*

As this is a bioinformatics project involving development of a database software application addressing a problem in biology, a very brief background on the bacteria, databases and software is given in this chapter, so as to enhance the understanding of the reader.

The whole genomes of more than 185 micro-organisms has been sequenced and has become important for the detailed understanding of complex cellular mechanisms [1, 19]. One of these organisms, *Corynebacterium glutamicum* bacteria, is frequently used for commercially producing the amino acids, nucleic acids and organic acids [2, 17, 23]. It belongs to a supra-generic group of gram-positive bacteria named *Corynebacterianeae*, which includes mycobacterium, nocardia and other phylogenetically related bacteria [30].

Coryneform bacteria are largely distributed in nature and are rod like, aerobic soil bacteria capable of growing on a variety of sugars and organic acids. *Corynebacterium glutamicum* is a non pathogenic species of coryneform becteria. In industry, this bacterial species is used for producing amino acids like Lysine [3], L-Glutamate [4, 5, 25], L-threonine [6], L-ornithine [7] and L-valine [28]. L-Lysine can only be produced by a mutant of *C. glutamicum* named MH20-22B [26]. These amino acids are used by the food industry and animal feed industry for enhancing the flavor of many sorts of food stuffs and supplementing the animal feed, respectively. Protein in general is an important nutrient in animal feed, too. Animal feeds are supplemented with essential amino acids because animal feed ingredients such as wheat, soya, corn, and fish meal are deficient in lysine, threonine, methionine and tryptophan. All these amino acids are produced by industrial fermentation, except methionine which is synthesized chemically. Lysine, threonine and tryptophan are produced by direct fermentation using mutants of *C. glutamicum* and recombinant strains of *E.coli* [14].

C. glutamicum was first known as Micrococcus glutamicus alongside with many other names like *Brevebacterium lactofermentum, B. flavum, B. divaricatum and C. lilium,* but finally this was resolved by the latest taxonomy introduced by Liebl et al [29] where *C. glutamicum* was primitively categorized as L-glutamic acid secreting bacterium. *C. glutamicum* excretes glutamaic acid after treatment with penicillin [8], induction of limited biotin [9] and adding of fatty acid ester surfactants [10]. As these treatments alter the structure of the cell surface, until 1980's it was considered that glutamic acid comes out passively from the cell membrane. This was termed leakage model in the literature, but many research results that do not agree with this leakage model are cited in [10-13]. Today, the exact mechanism of glutamate secretion in *C. glutamicum* is to be elucidated [20].

Production of amino acids with microbes like *Corynebacterium glutamicum* along with the scientific research, started in the mid 1950's [24] when it was discovered in Japan by Kinoshita and co-workers [25]. This bacterium has the characteristic property of excreting a substance named, glutamic acid, and it is therefore known as a glutamate secreting bacterium. *C. glutamicum* is now considered a microorganism of vital importance biotechnologically because this bacterium is used to produce around one million tonnes of amino acids per annum. Among them half of the amount goes for sodium glutamate or mono sodium glutamate (MSG) used as flavor enhancer [18] and more than 0.6 million ton for the synthesis of L-Lysine used as feed additive. The market is increasing rapidly, especially for L-Lysine, and is rising by 10% every year.

Random mutagenesis and selection was used earlier for empirical improvement of bacterial strains used for the production of amino acids, but nowadays the advancement in metabolic and genetic engineering techniques has made it possible to selectively improve the strains using the latest knowledge of metabolic pathways and their regulatory mechanisms. This has opened varied areas of research in biotechnology, for example the presence of particular amino acid export carriers as well as those of cyclic fluxes within the anaplerotic reactions (those forming intermediates of the citric acid cycle), findings that go far beyond *Corynebacterium glutamicum* and amino acid production. The complete genome of *C. glutamicum* and other closely related bacteria has been sequenced [18,21,22] So, further research can be done on all regulatory phenomena and their interactions at the cellular level, genome wide transcription and proteome. This complete Corynebacterial genome may help to describe principal methods of obtaining missing genetic information and thus will help in rational development of new strains which could be more efficient for the industrial production of amino acids. Besides, previous bioinformatics analysis have provided exciting and worthy information for industrial and science applications, because new genome information and novel bioinformatics tools are being generated at a greater pace, therefore a lot of interesting discoveries can be predicted from Corynebacteria in the future. Due to the extra ordinary importance in food companies, feed industry and existing global analysis, *C. glutamicum* has become an ideal for further analysis and modelling to explore, understand and exploit its complete metabolic and regulatory potential. In short, it deserves a whole monograph of its own [15].

Along with the significance of its classic characteristic of glutamate secretion and production of directed mutants, which are biotechnologically important, the beauty of being handled easily in the lab, *Corynebacterium glutamicum*'s importance is evident to the lab scientist. Also, *C. glutamicum* grows rapidly and develops dense cell colonies, especially when compared to ones with *E.Coli.* Further more a broader range of technologies and techniques exist for genetic modifications with *C. glutamicum*. Even more worth to know is that *Corynebacterium glutamicum* is a non-pathogenic microorganism and is classified as 'Generally Regarded As Safe' (GRAS) [27].

C. glutamicum is not only important biotechnologically but also non-biotechnologically. As in taxonomy, this species of coryneform bacteria comes form a suborder *Corynebacterianeae* which contains two bacteria, namely *Mycobecterium tuberculosis*, a very stumbling bacteria to deal with while experimenting in the lab, and *Rhodococcus genus* which is comparatively less explored. The small genome of *Rhodococcus genus* together with genome sequences of relevant *Mycobacterium* and *Corynebacterium* species, as well as above given features of *C. glutamicum,* have made it an ideal micro-organism with the help of which basic characteristics of *Corynebacterianeae* can be further elucidated. For example the functioning and synthesis of the outer membrane, which is otherwise only available in gram-negative bacteria, while *C. glutamicum* is a gram-positive [16]. Although *Corynebaterianeae* are Gram-positive bacteria, nevertheless its cell envelope comprises of plasma membrane and peptideglycan-based cell wall, and they share with Gram negative bacteria the characteristic of making an extra outer cell layer which is different from a cell membrane [31]. The presence of this outer membrane diffusion barrier is reinforced by characterization of cell envelope proteins capable of forming pores [32, 33]. Mutants of *C. glutamicum,* which produce vital, biotechnologically important compounds, are also available for many other amino acids [34]. Understanding of the regulation and mechanism of export carriers was not fast initially because it was not proven that particular export carriers exist for this, therefore, the concepts of the diffusion model [35] and functional inversion of the uptake system [36], efflux mediated by osmotically controlled pores [37] are under discussion, at least for ABC transporters. As glutamate and lysine carry a net charge, the idea of passive diffusion should be rejected [38] and since both are transported against the concentration gradient, the idea of pores fails [39]. Therefore, it is believed nowadays that certain carrier proteins do exist to transport the amino acids from cytoplasm to external medium. Besides, active transport is also reported in *C. glutamicum* [40].

1.2 Databases and software paradigms

Regardless of the field of research and science, significance of data and efficient management of data to enhance its proper usage can never be ignored. By definition a database is no more than a collection of records which are managed by software. Database systems have its origin from file systems in 1960's. File systems do support data storage and retrieval but lack concurrent access, synchronization, querying and manipulating the data. Databases have become important not only to business but also for many scientific disciplines. Data management is done by the investigators of the human genome, by biochemists researching on the medicinal proteins and by other scientists [43]. The main strength of these database applications is due to special software underneath called a 'database management system' (DBMS). A DBMS is a powerful tool for handling huge quantities of data safely and efficiently over a long time span. DBMSs are among the complex kinds of software capable of transaction management, making persistent storage, executing standardized queries and providing programming interfaces [45, 46].

Relational database management systems (RDBMS) emerged in the early 1970's. An RDBMS create views of the data stored in tables termed 'relations' where columns have data against the attributes. Each row in a table is a record, termed 'tuple'. Data is not accessible directly because of a complex data structure underneath which enables the database developers to interact with data using standard languages like 'Structured Query Language' (SQL), 'Data Manipulation Languages' (DML), and Data Definition Languages (DDL) [44]. Among the various types of application architectures, one used in databases is client-server. The client requests services and the server responds to the client's requests. Actually the client and server are two machines on which the client software and server software are configured, respectively. Users interact with the client software having interfaces to send queries to the server [43].

Software engineering is the application of a disciplined and quantifiable approach to the development and maintenance of software. The systematic and disciplined approach encompasses computer science, management and engineering techniques [49]. The term SE became popular after the NATO conference on SE headed by Dr. F. L. Bauer at Germany [47]. There are many definitions of SE and Software. For example "Software engineering is the establishment and sound engineering principles applied to obtain reliable and efficient software in an economical manner". "Software is both a product and a vehicle for delivering a product" [48]. Among various software types the one resembling with our project is 'Application software'. The key factor in developing the quality software is the software process, which gives a framework of project activity management. Depending on the nature of the project, software process varies. A basic process framework broadly includes requirement collection, planning, designing, development and deployment [49].

Software process models are adapted according to requirements and describe tasks, milestones and products. The most traditional paradigm is waterfall fits well when each and every requirement is known before. When requirement are not known fully, prototyping is used to gather requirements. Incremental model suggests building smaller chunks first and then growing in terms of functions and data [50].

From the above review of the background literature is it is evident that *Corynebacterium glutamicum* has been actively investigated in terms of various experimental thoughts of glutamate secretion, transporters and mutants. As this project is aiming to incorporate the information from *Corynebacterium* transporter classes, transporter substrates, cloning experimental setups used to generate mutant strains, so this was necessary to have at least a minimal understanding of the scientific background for which this microorganism has become so important in the food industry and medical science.

2 Introduction

2.1 Problem Definition

There are three main publicly available genome annotations for *C. glutamicum.* One is from Kitasato University, Japan, represented by the genome accession number BA000036 as shown in figure 2.2, where gene id's are suffixed by 'Cgl' [41]. The second is from University of Bielefeld, Germany, denoted by the genome accession number BX927147 in figure 2.2, where gene id's are suffixed by 'Cg' [42]. The third one is from NCBI with genome accession number NC_003450, as shown in figure 2.2, having gene id's suffixed by 'NCgl' [55]. All these three annotation sources have different number of protein coding genes and varying numbers of overlaps of similar genes. Differences in total number of gene entries and overlapping gene entries can be best viewed in figure 2.1. The text in bold in figure 2.1 shows the annotation source along with the total number of protein coding genes in each annotation. The numbers in overlapping regions of the figure 2.1 represent the common genes in respective annotations. For example the number of common genes in the Cg and the NCgl annotations is 2436. The central triangle shaped overlapping region of figure 2.1 represents the number of common genes (1822) in all three annotations. This variation in annotations is because of the differences in methodologies involved during sequencing processing, because the annotations have been done at three different places, namely Japan, Germany and at NCBI. Besides the variation in the number of annotated genes and similarities in the corresponding genes, there is a problem of missing data, meaning that some genes are present in one annotation but not in others. Also, there are conflicting overlaps, i.e. the sequence lengths differ for the same gene in different annotations. As the Cgl annotation has been re-annotated, some public data sources have updated information based on this source, while others have the older one. This has led to inconsistent data. Last but not least, as the data about *Corynebacterium* is stored in different data sources, it is redundant as well.

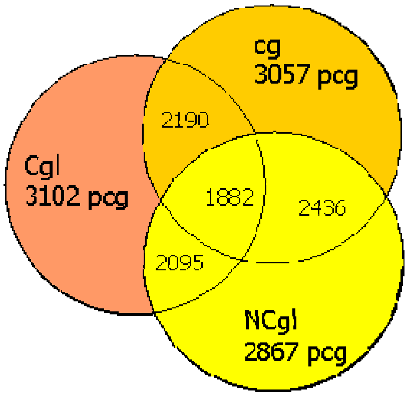
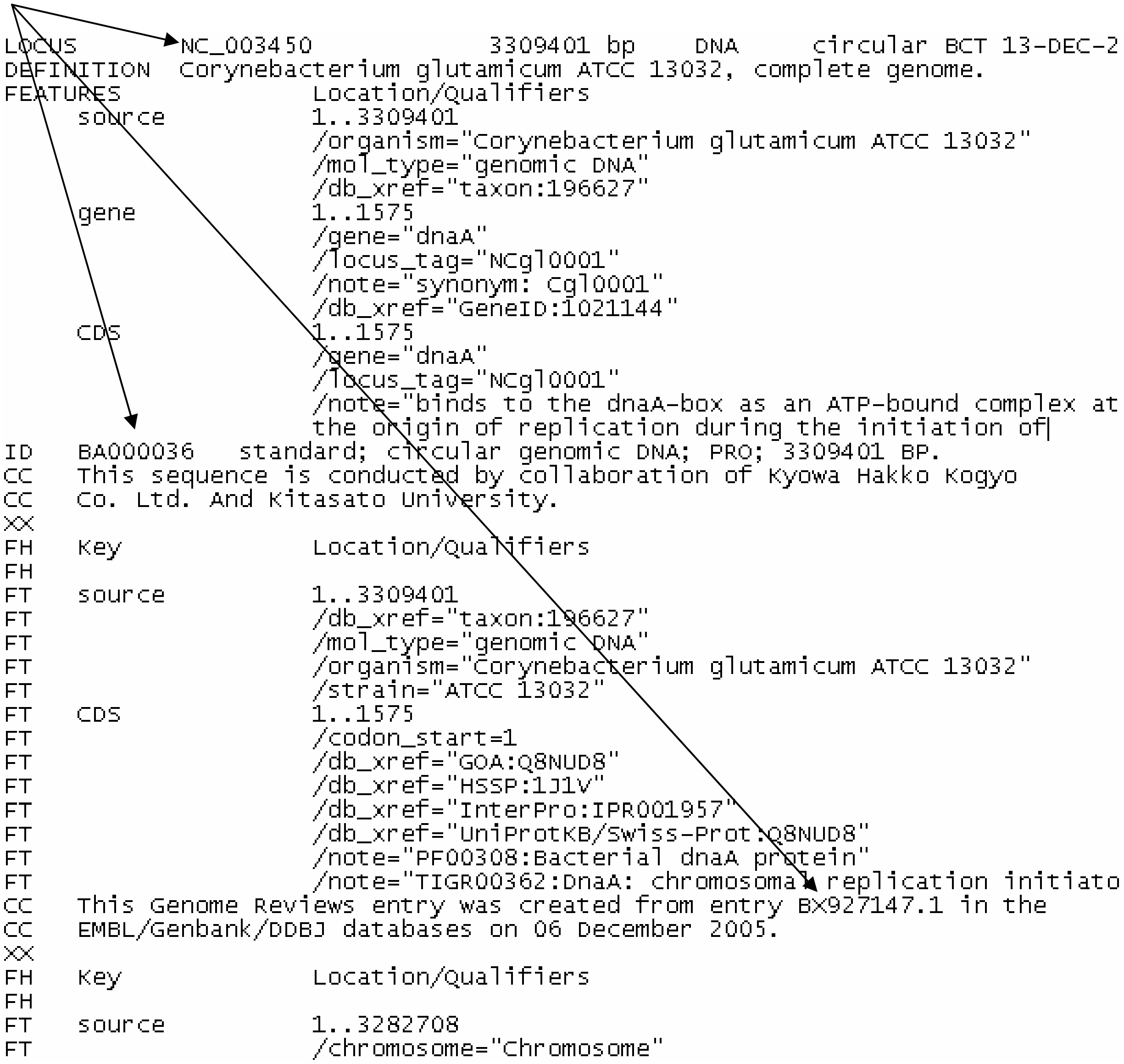


Figure 2.1: Protein coding genes

Despite of all mentioned problems, if a lab scientist is motivated to scan through different data sources manually, the data of *C. glutamicum* genome annotation is available in the text files containing data in single line database entries format. Names of different parameters of the complete proteome are also different in many cases, which is a reason of confusion and ambiguity for one who is scanning these text files manually to make a comparative study for the gene of interest. Each text file contains more than one hundred thousand lines. A combined view of the file formats is shown in figure 2.2. The arrows point to the genome accession numbers and show the start of the corresponding text file in the combined view of figure 2.2. Annotation files showing its complete contents can be viewed by accessing the links[1].

[1] http://www.ebi.ac.uk/cgi-bin/expasyfetch?BA000036, ftp://ftp.ebi.ac.uk/pub/databases/genome_reviews/dat/BX927147_GR.dat



```
LOCUS       NC_003450           3309401 bp    DNA     circular BCT 13-DEC-2
DEFINITION  Corynebacterium glutamicum ATCC 13032, complete genome.
FEATURES             Location/Qualifiers
     source          1..3309401
                     /organism="Corynebacterium glutamicum ATCC 13032"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:196627"
     gene            1..1575
                     /gene="dnaA"
                     /locus_tag="NCgl0001"
                     /note="synonym: Cgl0001"
                     /db_xref="GeneID:1021144"
     CDS             1..1575
                     /gene="dnaA"
                     /locus_tag="NCgl0001"
                     /note="binds to the dnaA-box as an ATP-bound complex at
                     the origin of replication during the initiation of|
ID   BA000036   standard; circular genomic DNA; PRO; 3309401 BP.
CC   This sequence is conducted by collaboration of Kyowa Hakko Kogyo
CC   Co. Ltd. And Kitasato University.
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..3309401
FT                   /db_xref="taxon:196627"
FT                   /mol_type="genomic DNA"
FT                   /organism="Corynebacterium glutamicum ATCC 13032"
FT                   /strain="ATCC 13032"
FT   CDS             1..1575
FT                   /codon_start=1
FT                   /db_xref="GOA:Q8NUD8"
FT                   /db_xref="HSSP:1J1V"
FT                   /db_xref="InterPro:IPR001957"
FT                   /db_xref="UniProtKB/Swiss-Prot:Q8NUD8"
FT                   /note="PF00308:Bacterial dnaA protein"
FT                   /note="TIGR00362:DnaA: chromosomal replication initiato
CC   This Genome Reviews entry was created from entry BX927147.1 in the
CC   EMBL/Genbank/DDBJ databases on 06 December 2005.
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..3282708
FT                   /chromosome="Chromosome"
```

Figure 2.2: Genome data format in text files[2]

Figure 2.2 shows a screen shot of text files from the original annotations of Cg, Cgl and NCgl. In this format of genome data, it is not easy to search and compare the data among different annotations and it is difficult to make an extensive multidimensional and customized search against different protein parameters. Comparison of all genome annotations in terms of construction, deletion, over-expression, mutants and features like gene locations, neighboring genes, orientation (direct or complementary strand), gene overlap, or varying sequence lengths, are not easy to manage either. Also, generating graphical representations of structure-function relations by comparison of predicted trans-membrane domains (TMD) and protein motifs is difficult when data is inconsistent and redundant on various publicly available biological database servers.

[2] http://www.ebi.ac.uk/cgi-bin/expasyfetch?BA000036, ftp://ftp.ebi.ac.uk/pub/databases/genome_reviews/dat/BX927147_GR.dat

2.2 Related work

Similar work has been partially done before on various other organisms and on *Corynebacterium glutamicum* These tools are mostly web-based and are available on different web resources. We plan to develop a desktop application which will incorporate the functionality available in previous tools along with additional features. There are web based databases for numerous micro-organisms, for example 'EchoBase' [51] which is a data source on the web providing details of experiments about genes along with their products in the bacterium *Escherichia coli*. Similarly data banks are available for cyanobacteria genes and mutants [52], the *Bacillus subtilis* Genome Database [53] and Aramemnon, a plant membrane protein database [54] for *Arabidopsis*. Information regarding the *C. glutamicum* complete genome is available in various databases because *C. glutamicum* is an important micro-organism commercially. However, not all of them are open to public access. An example of such a database is owned by 'Degussa AG' (a chemical producing company in Germany). A web resource having the proteins of the *C. glutamicum* is available at [66] and is maintained by the DNA Databank of Japan (DDJB). This web tool is based on the Cgl [41] annotation done in Kitsato University in Japan. Another web-based repository of *C. glutamicum* proteins is at [69] which is also a part of the Encyclopedia of Life and maintained by the University of California. This database is based on the data from the NCgl [55] annotation from NCBI. A web tool displaying the condensed genome of *C. glutamicum* can be found at [67] maintained by the Institute of genomic research TIGR. This tool is based on the Cg [42] annotation done at University of Bielefeld, Germany. A putative peptidase database [68] of *C.glutamicum* is maintained by the Sanger Institute. A program named, tRNAscan-SE for detecting the transfer RNA genes in genomes is cited in [70]. This tool is also available to on the web.

From the above review of the already done related work, we develop new ideas for this project. For example there were separate tools for specific information as mentioned in [66, 67, 69]. We thought of providing more than that information in our software by providing access to all three annotations, so that it could make the comparison in a similar format. There was no graphical display for genomes in these tools, but in our project we displayed the genome information in a graphical way. We developed a similar kind of genome display as we had found in a database for *E. coli* named EchoBase [51], but we extended the same display to three annotations by generating a link among the corresponding genes from the three annotations. In our project we also included the information about the trans-membrane helices, transporter classes, substrate and homologues. All this information could previously be found by various tools but not at one place. As all the mentioned already existing tools were developed and customized according to their own needs, so even using so many tools in various places was not the complete solution. Further integration of the own lab data was also part of this project. Though we did not invent something new, but we did developed a new tool for *C. glutamicum.*

An additional aim that we specified was to develop modules capable of managing the available mutants, experimental setups regarding growth, transport and metabolomics of *C. glutamicum* and other information that might typically be generated in a bio-chemistry lab. To cope with all mentioned problems and requirement specifications, a database software application (named as 'Corynebase') for the classification of *C. glutamicum* genes, proteins, mutants and experimental protocols, was proposed as solution on a single platform fulfilling required features.

2.3 Aims and Objectives

The key features for designing of production strains are the modulation of synthesis pathways for specific products as well as the efficient use of transporters for the import of starting substances and export of products. The intention might be to set up a new screening program for transporters of biotechnologically important compounds. The developed database should cover the categorization of available information. This information is based on the known genome sequence as well as huge amounts of information generated by numerous tools in public databases. The databank should include:

- Information regarding genes and DNA-protein sequences (NCBI, UniProt, SRS)
- Prediction of trans-membrane helices (TMH's) as well as domain structures (TmHMM, Pfam)
- Functional predictions by identification of homologues (Blast)
- Predictions of transporter class and substrates (TransportDB)

- Information about available mutants regarding the phenotype, and the experimental setup used for the characterization (Lab insights, literature)
- Description of known transporters (Lab insights, literature)

Developing the application covering data processing, protein-DNA sequence extraction and management of lab data will be the main objective. Finally the database deployment and setup of the environment, following a short briefing to a group of intended user of the software, will be the ultimate aim of this project. In spite the genome sequence known, until now no databank providing such complete information is available. The aim of the software developed in this project is that it should be possible to use for the *in silico* characterization of proteins, design of mutants and classification of experimental data. As we are mainly interested in membrane fluxes as part of the production process of biotechnologically important compounds, therefore, 'Corynebase' will be useful for the identification of putative transporters for such substrates. To cover all proteins involved in import of precursors as well as the export of products, the database organization will be done at the genome level. Consequently, keeping in view the requirement analysis, we aim our software to cover the following functionality:

Module 1: Parsing three complete genome annotations into database relations along with an update system.
Module 2: Extensive and diverse searches covering main annotations and other important information.
Module 3: Graphical visualization and comparison of all genome annotations data (start, end, neighbours, orientation ...).
Module 4: Graphical visualization of protein structure information (THM, motif prediction ...).
Module 5: Information system about transporters and functional properties (pI, Mol Wt, Length COG, KEGG, Transport DB...).
Module 6: Protein comparison and protein homologues.
Module 7: Available mutant information management.
Module 8: Management of experimental setups, protocols and results.
Module 9: Reporting services for search results including reports for 'Risikobewertung gentechnischer Arbeiten' (Risk evaluation of genetic work, which must be approved under German law before working on a project involving genetic modifications and cloning experiments).

The overall aim of this project is to develop a database application capable of assisting group members working with *Corynebacterium* at the laboratories of Institute of Biochemistry, University of Cologne, Germany. That is why the development of this database application software is based on the requirements gathered from the researchers in the lab. With the development of this application, it is aimed that this will not only be a useful application for all members of the research group but also may be of interest to every one researching on *C. glutamicum* because no such application exists so far. The provision of this kind of application which is not available anywhere, will be a contribution to the scientific community, especially to the research group at the institute for which the software is customized to.

There are some web-based tools [51-54, 66-70] available, which provide very specific information on certain aspects of *Corynebacterium*. The novelty with this software is that it aims to provide complete and comprehensive information from three different genome annotations. It also aims to give a visual representation to users along with a facility of comparing the information from different sources to counter check the redundancies and differences. Furthermore this project aims to provide a private mutant information exchange system to the lab group members. In this way they will be able to share the confidential mutant information or other comments within the group while using this application over the local area network. As the mutants of the *C. glutamicum* could be of high commercial value, therefore this information cannot be made available for public use.

3 Material and Methods

3.1 Method

For implementing this database software application, the first phase was the initial understanding of the problem domain, for which literature was studied on *C. glutamicum* annotations along with investigating thoroughly the existing problems in the annotations. After the initial understanding of the problem, the major task was the collection of data for *C. glutamicum* genome annotation.

In the beginning we relied on the data provided by the sequence retrieval system (SRS) at European Bioinformatics Institute (EBI), European Molecular Biology Labs (EMBL) [56], for our search function in the software application. We preferred this set of data for searching as it was possible to have correspondence between the Cg [42] and Cgl [41] annotation under the parameter name OrderedLocus. Data at EBI was organized under the UniProt [57] id's because UniProtKB/SwissProt, UniProtKB/TrEMBL [58], and the same UniProt id's are also referenced in the Cg, Cgl and NCgl annotations. However there were inconsistencies in UniProt id's in these sources because of updates at EBI which have not been synchronized to the data at other sources. Another limitation with data at EBI was the complete absence of information about tRNA's and rRNA's, therefore we discarded the idea of using the data at EBI as our search data against *C. glutamicum* genes.

```
rs.Open "insert into cgTab values ('" & Cds & "'," & geneStart
& "," & geneEnd & "," & geneLen & ",'" & geneDir & "','" &
Evidence & "','" & Gene & "','" & geneSyn & "','" & locusTag
& "','" & Product & "','" & ecNumber & "','" & Functions &
"','" & bioProcess & "','" & cComponent & "','" & proteinId &
"','" & dbXref & "','" & transTable & "','" & Translation & "')",
dataSource, adOpenDynamic, adLockOptimistic
```

Figure 3.1.1: DML statement inserting a database entry

Finally, we relied on the original genome data from Cg, Cgl and NCgl annotations. As mentioned before, data in these annotations was in single line database entries format, therefore it was not usable directly in the relational database at back end. Then it was decided to write parsers for these three annotation files. The function of the parser was to take these text files as input and parse the whole genome data, including all parameters, protein sequence and DNA sequence, to database relations. An example of data entry by the parser into the database is shown in the DML command shown in figure 3.1.1

After the parser had been written there still remained a lack of correspondence between the three annotations regarding annotation for a particular gene. Only in the NCgl was it possible to get a corresponding gene entry in Cgl, but there was no correspondence for the same gene in the Cg annotation. To overcome this problem, we created our own correspondence among these three annotations. We managed this because the NCgl annotation already contained synonyms from the Cgl annotation and data at EMBL contained a correspondence between Cgl and Cg annotation. All Cg id's were extracted by creating a SQL join on CGL id's. This was possible because of the Cgl correspondence both with NCgl and Cg at two different data sources. We combined this generated correspondence with the data from Cg annotation to use it in our search function.

```
    adoRunTime.RecordSource = "select * from combine
where (ID =" & Val(str) & ") OR " _
    & "(Transporter_class LIKE '%" & str & "%') OR " _
    & "(Cg LIKE '%" & str & "%') OR " _
    & "(Cgl LIKE '%" & str & "%') OR " _
    & "(Ncgl LIKE '%" & str & "%') OR " _
    & "(Transporter_substrate LIKE '%" & str & "%') OR " _
    & "(geneName LIKE '%" & str & "%') OR " _
    & "(geneLen =" & Val(str) & ") OR " _
    & "(Product LIKE '%" & str & "%') OR " _
    & "(ecNumber LIKE '%" & str & "%') OR " _
    & "(Functions LIKE '%" & str & "%') OR " _
    & "(bioProcess LIKE '%" & str & "%') OR " _
    & "(cComponent LIKE '%" & str & "%') OR " _
    & "(proteinId LIKE '%" & str & "%') OR " _
    & "(Cds LIKE '%" & str & "%') OR " _
    & "(geneDir LIKE '%" & str & "%') OR " _
    & "(geneSyn LIKE '%" & str & "%') OR " _
    & "(dbXref LIKE '%" & str & "%') OR " _
    & "(Evidence LIKE '%" & str & "%') OR " _
    & "(locusTag LIKE '%" & str & "%') OR " _
    & "(TopHmm LIKE '%" & str & "%') OR " _
    & "(TopPho LIKE '%" & str & "%') OR " _
    & "(TopSos LIKE '%" & str & "%') OR " _
    & "(Translations LIKE '%" & str & "%')" _
```

Figure 3.1.2: SQL query to search a keyword

The software application was organized into the following components.

A. Parsing Module (Cg, Cgl, NCgl)
B. Search Module
C. Module for genome organization
D. Protein structure
E. Protein functions & homologues
F. Mutants information

G. Experimental setups and protocols
H. Reporting services
I. Self extracting installation setup and distribution package.

```
adoCg.RecordSource = "select * from cgTab where locusTag
LIKE '%" & Trim(txtcgID.Text) & "%' OR"_
& "(geneName LIKE '%" & Trim(txtcgID.Text) & "%') order
by locusTag" _
```

Figure 3.1.3: SQL statement for genome organization

The search module was implemented using data from the generated correspondence with which it was possible to search details of genes in any of the annotations. The SQL statement shown in figure 3.1.2, queries the key words against the parameters in the SQL command.

The module for graphical genome organization was implemented by extracting information from the relevant annotations on the basis of id's searched by the search module as shown in figure 3.1.3. Search navigation implemented as part of search module was linked to the genome organization module.

```
gStart = frmGene.adoCg.Recordset("geneStart")
gEnd = frmGene.adoCg.Recordset("geneEnd")
StartingD = (gStart / 60) + 1
  If gStart <= 60 Then
    StartingD = StartingD - 1
  End If
EndingD = (gEnd / 60) + 1
tempStr = Left(cng(tempStr), InStr(1, Cstr(StartingD), ".") - 1)
tempStr = Left(cng(tempStr), InStr(1, Cstr(EndingD), ".") - 1)
For i = Starting To Ending Step 1
  Adodc1.ConnectionString = dataSource
  Adodc1.CommandType = adCmdText
  Adodc1.RecordSource = "select dnaSequence from cgSeq
where ID = " & i
  strSeq = strSeq & Trim(Adodc1.Recordset(0)
Next
Text1.Text = Text1.Text & vbCrLf & ">" &
Trim(frmGene.txtcgID.Text) & vbCrLf & Mid(strSeq, gStart -
(Starting - 1) * 60, gEnd - gStart)
```

Figure 3.1.4: Code snippet of DNA extraction

Protein and DNA sequences were also extracted from the genome annotation files which our parser component has previously converted to the database relations. A code snippet of DNA extraction is shown in figure 3.1.4. The next module was developed to show the graphical comparative view of the predicted trans-membrane domains form TmHMM [59], Phobius [60] and SOSUI [61] and predicted protein motifs from Pfam [62] and InterPro [63]. The module of proteins functions, transporters and homologues were developed by using data from TransportDB [64] and BLAST [65] results because just the sequence is of no use unless analyzed by comparing against available databases to get clues regarding relatives and function. The two modules (F & G) use lab data regarding available mutants, experimental setups and protocols regarding growth, transport and metabolomics in *C. glutamicum*. A code snippet of display of the protein

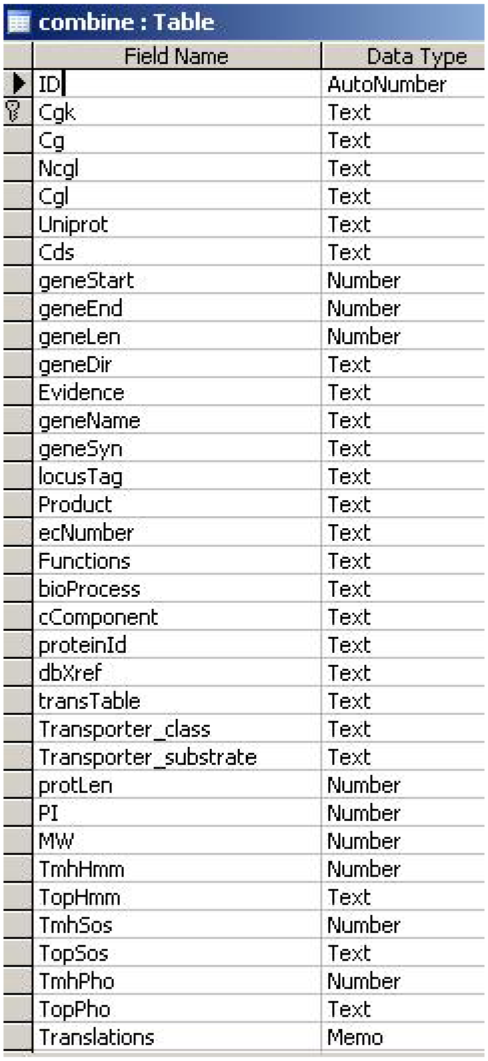
```
  For i = 1 To adoDomain1.Recordset.RecordCount Step 1
  Load lblcgFam1(i)
  lblcgFam1(i).Visible = True
    If Not IsNull(adoDomain1.Recordset("famEnd")) And Not IsNull(adoDomain1.Recordset("famStart")) Then
      Wid = adoDomain1.Recordset("famEnd") - adoDomain1.Recordset("famStart")
      lblcgFam1(i).Width = Wid * residueFactor
        If i = 1 Then
          diff = adoDomain1.Recordset("famStart")
        Else
          diff = adoDomain1.Recordset("famStart") - xEnd
            If diff < 0 Then
              diff = 1
            ElseIf diff > 2000 Then
              diff = 2000
            End If
         End If
          lblcgFam1(i).Left = lblcgFam1(i - 1).Left + lblcgFam1(i - 1).Width + (diff * residueFactor)
          famText = famText & adoDomain1.Recordset("famName") & " " & adoDomain1.Recordset("famName2") & "
              [" & adoDomain1.Recordset("famStart") & "-" & adoDomain1.Recordset("famEnd") & "]"
            If Not IsNull(adoDomain1.Recordset("Description")) Then
              famText = famText & " " & adoDomain1.Recordset("Description")
            End If
        lblcgFam1(i).Caption = famText
        lblcgFam1(i).ToolTipText = famText
        xEnd = adoDomain1.Recordset("famEnd")
        domainCounter = domainCounter + 1
     Else
      lblcgMsgFam1.Visible = True
     End If
  If Not adoDomain1.Recordset.EOF Then
    adoDomain1.Recordset.MoveNext
  End If
  famText = ""
Next
```

Figure 3.1.5: Code snippet for graphical display of protein domains.

motifs is shown in figure 3.1.5

3.2 Data

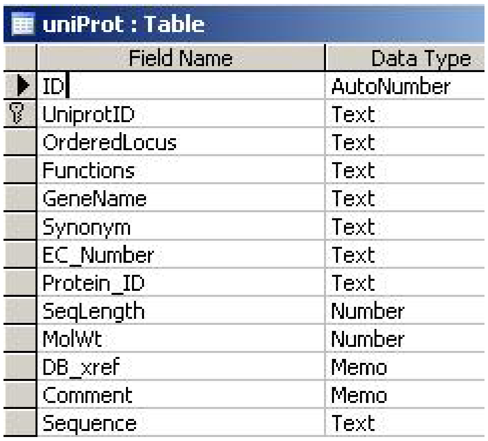
Figures 3.2.1 to 3.2.10 show the database schemas used for CORYNEBASE.



combine : Table

Field Name	Data Type
ID	AutoNumber
Cgk	Text
Cg	Text
Ncgl	Text
Cgl	Text
Uniprot	Text
Cds	Text
geneStart	Number
geneEnd	Number
geneLen	Number
geneDir	Text
Evidence	Text
geneName	Text
geneSyn	Text
locusTag	Text
Product	Text
ecNumber	Text
Functions	Text
bioProcess	Text
cComponent	Text
proteinId	Text
dbXref	Text
transTable	Text
Transporter_class	Text
Transporter_substrate	Text
protLen	Number
PI	Number
MW	Number
TmhHmm	Number
TopHmm	Text
TmhSos	Number
TopSos	Text
TmhPho	Number
TopPho	Text
Translations	Memo

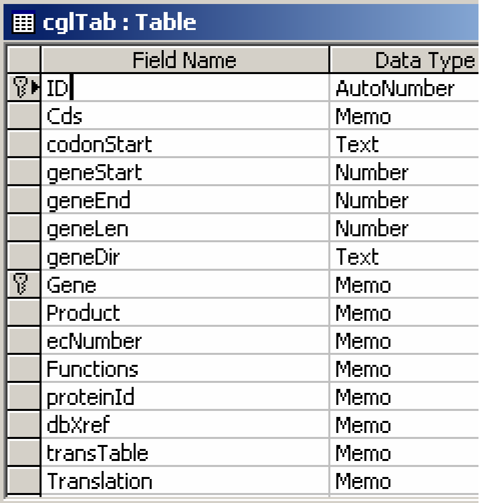
Figure 3.2.1: Combine



uniProt : Table

Field Name	Data Type
ID	AutoNumber
UniprotID	Text
OrderedLocus	Text
Functions	Text
GeneName	Text
Synonym	Text
EC_Number	Text
Protein_ID	Text
SeqLength	Number
MolWt	Number
DB_xref	Memo
Comment	Memo
Sequence	Text

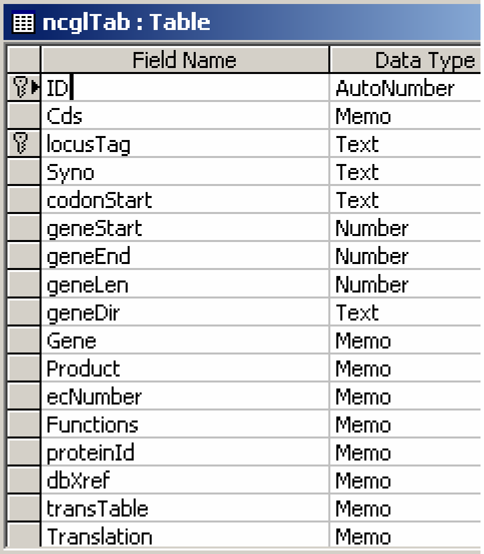
Figure 3.2.2: Uniprot



cglTab : Table

Field Name	Data Type
ID	AutoNumber
Cds	Memo
codonStart	Text
geneStart	Number
geneEnd	Number
geneLen	Number
geneDir	Text
Gene	Memo
Product	Memo
ecNumber	Memo
Functions	Memo
proteinId	Memo
dbXref	Memo
transTable	Memo
Translation	Memo

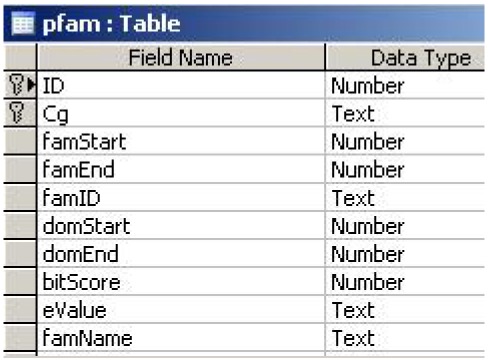
Figure 3.2.3: cglTab



ncglTab : Table

Field Name	Data Type
ID	AutoNumber
Cds	Memo
locusTag	Text
Syno	Text
codonStart	Text
geneStart	Number
geneEnd	Number
geneLen	Number
geneDir	Text
Gene	Memo
Product	Memo
ecNumber	Memo
Functions	Memo
proteinId	Memo
dbXref	Memo
transTable	Memo
Translation	Memo

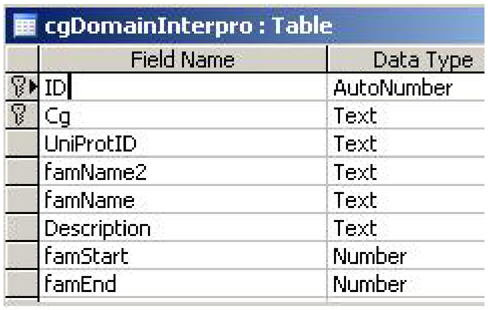
Figure 3.2.4: ncglTab



pfam : Table

Field Name	Data Type
ID	Number
Cg	Text
famStart	Number
famEnd	Number
famID	Text
domStart	Number
domEnd	Number
bitScore	Number
eValue	Text
famName	Text

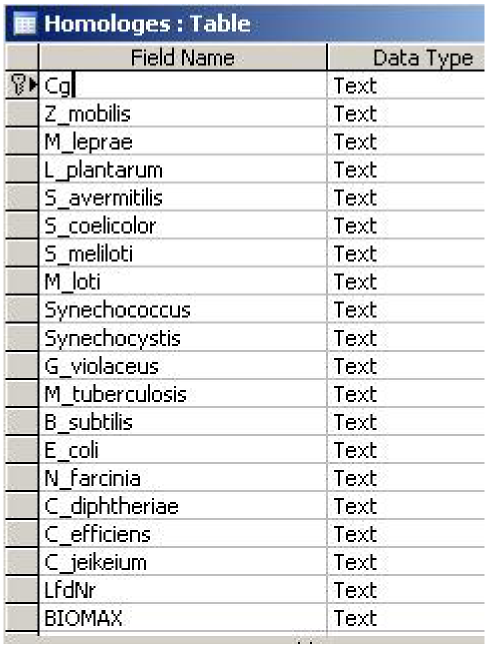
Figure 3.2.5: pfam



cgDomainInterpro : Table

	Field Name	Data Type
🔑▶	ID	AutoNumber
🔑	Cg	Text
	UniProtID	Text
	famName2	Text
	famName	Text
	Description	Text
	famStart	Number
	famEnd	Number

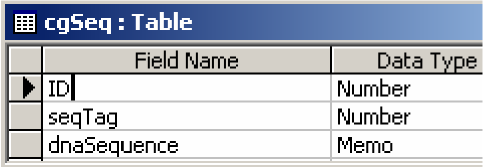
Figure 3.2.6: cgDomainInterpro



Homologes : Table

	Field Name	Data Type
🔑▶	Cg	Text
	Z_mobilis	Text
	M_leprae	Text
	L_plantarum	Text
	S_avermitilis	Text
	S_coelicolor	Text
	S_meliloti	Text
	M_loti	Text
	Synechococcus	Text
	Synechocystis	Text
	G_violaceus	Text
	M_tuberculosis	Text
	B_subtilis	Text
	E_coli	Text
	N_farcinia	Text
	C_diphtheriae	Text
	C_efficiens	Text
	C_jeikeium	Text
	LfdNr	Text
	BIOMAX	Text

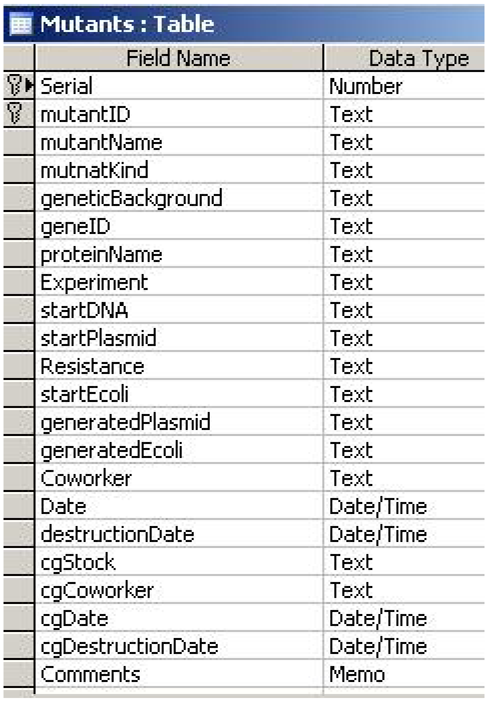
Figure 3.2.7: Homologues



cgSeq : Table

	Field Name	Data Type
▶	ID	Number
	seqTag	Number
	dnaSequence	Memo

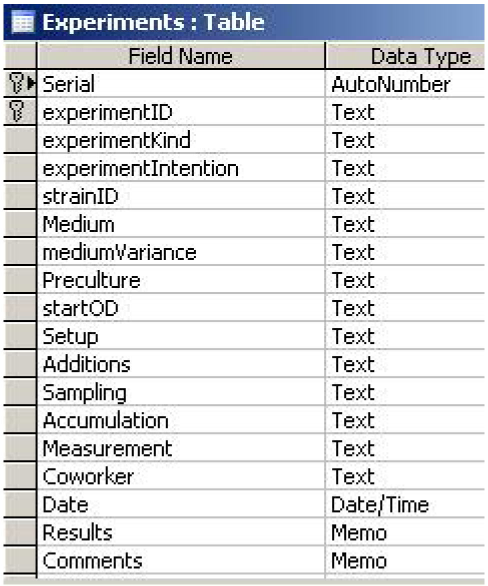
Figure 3.2.8: cgSeq



Mutants : Table

	Field Name	Data Type
🔑▶	Serial	Number
🔑	mutantID	Text
	mutantName	Text
	mutnatKind	Text
	geneticBackground	Text
	geneID	Text
	proteinName	Text
	Experiment	Text
	startDNA	Text
	startPlasmid	Text
	Resistance	Text
	startEcoli	Text
	generatedPlasmid	Text
	generatedEcoli	Text
	Coworker	Text
	Date	Date/Time
	destructionDate	Date/Time
	cgStock	Text
	cgCoworker	Text
	cgDate	Date/Time
	cgDestructionDate	Date/Time
	Comments	Memo

Figure 3.2.9: Mutants

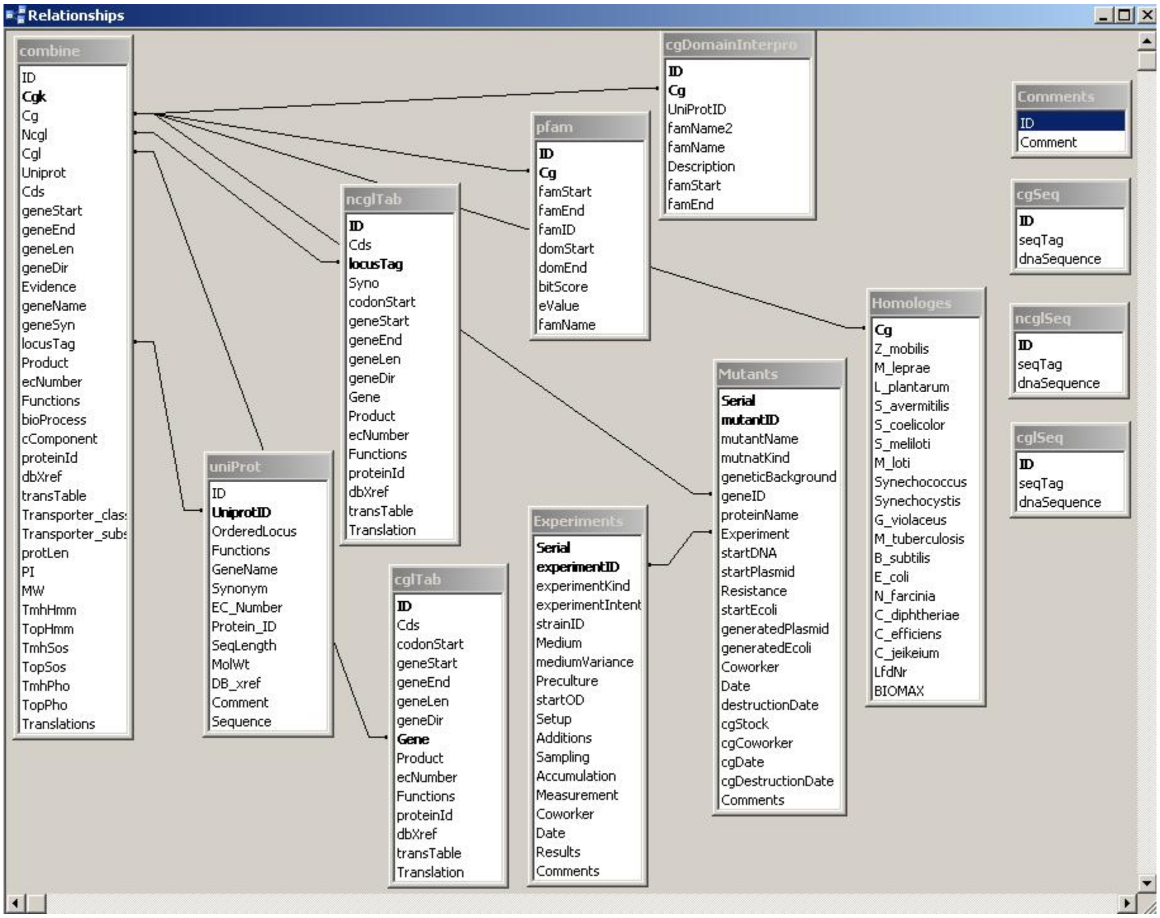


Experiments : Table

	Field Name	Data Type
🔑▶	Serial	AutoNumber
🔑	experimentID	Text
	experimentKind	Text
	experimentIntention	Text
	strainID	Text
	Medium	Text
	mediumVariance	Text
	Preculture	Text
	startOD	Text
	Setup	Text
	Additions	Text
	Sampling	Text
	Accumulation	Text
	Measurement	Text
	Coworker	Text
	Date	Date/Time
	Results	Memo
	Comments	Memo

Figure 3.2.10: Experiments

The schema 'combine' in figure 3.2.1 is used for generating a correspondence between the three main annotations Cg, Cgl and NCgl. A separate identifier Cgk was created to access all these three identifiers. This identifier and serial number for number of records was used as the primary key. The other columns were the parameters extracted from the Cg annotation. Information about the number of trans-membrane helices predicted along with their topologies from three different servers namely TmHMM, Phobius and Sosui. This table also contains data about isoelectric point and molecular weight of the proteins. The schema uniport shown in figure 3.2.2 contains the data available from EBI. This data is taken from other annotations and is accessible by UniprotID which is a primary key in this table. The column orderedLocus contains the correspondence between the Cg and Cgl annotations. The schema cglTab shown in figure 3.2.3 contained data generated by our parser by using the genome annotation data from the Cgl annotation. Column Gene is a Cgl id in this table and is used as the primary key. The schema in figure 3.2.4 contains data generated by our parser by using the genome annotation data form the ncgl annotation. Column locusTag is a ncgl id in this table and is used as the primary key along with the serial ID. The schema in figure 3.2.5 contains the information about the protein motifs predicted by the Pfam server. The schema in figure 3.2.6 contains the information about the protein motifs predicted by the Interpro server. The schema shown in figure 3.2.7 contains the information about the homologues of Corynebact- erianeae as well as *C. glutamicum*. This information also includes best homologues from the *E. Coli, B.subtilis* and *M.tubercluosis.* All are accessible via Cg id's. The schema in figure 3.2.8 contains the DNA sequence from Cg, Cgl and NCgl annotations parsed by the parser. The schema in figure 3.2.9 contains lab data about the mutants along with the names of the persons working with these. The schema in figure 3.2.10 contains lab data about the experimental setup and protocols.



(Figure 3.2.11: Entity Relationship Diagram)

Figure 3.2.11 shows the relationship between all schemas used for CORYNEBASE in the form of an entity relationship diagram (ERD). The entity named 'combine' of ERD in figure 3.2.11 shows three main ids from the three annotations, namely Cg, Cgl and NCgl. All these three ids's in the entity 'combine' have been linked via a third identifier, named 'Cgk'. Data form UniProt also contains id's from the Cg and Cgl annotations in attribute OrderedLocus. The 'LocusTag' attribute of the entity 'combine' has a one-to-one relationship with attribute UniprotID in entity 'uniProt'. Attributes named 'Gene' and 'locusTag' in entities 'cglTab' and 'ncglTab' have a one-to-one relationship in entity 'combine' through attributes 'Cgl' and 'Ncgl' respectively. Attribute 'Cg' in entity 'combine' has one-to-many relationship with entities namely 'cgDomainInterpro' and 'pfam' Entities named 'Experiments' and 'Mutants' have a one-to-many relationship between attributes named, 'experimentID' and 'Experiment' respectively. Entities named, 'combine' and 'Homologes' have one-to-one relationship through attribute 'Cg'. Tables containing the DNA sequence extracted by the parser are stored in the independent schemas namely cgSeq, cglSeq and nclgSeq.

3.3 Process Model

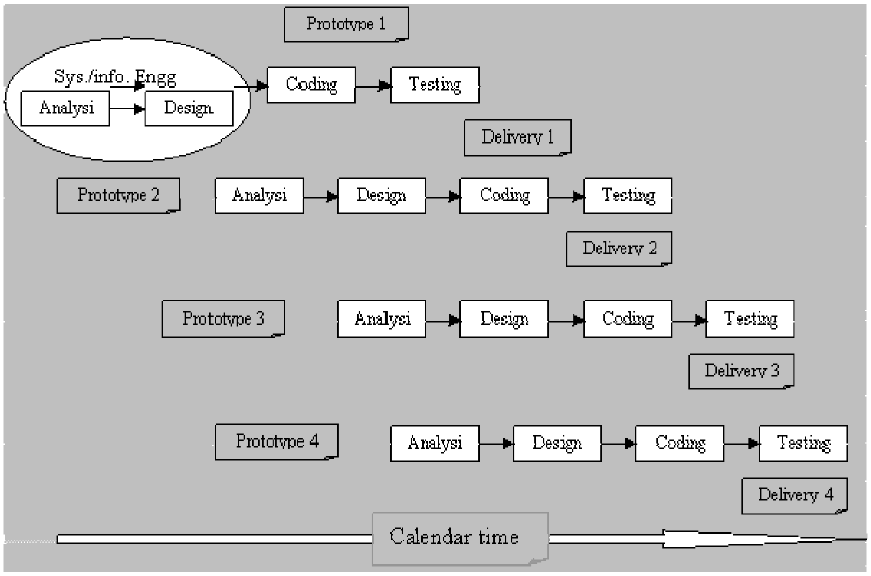


Figure 3.3.1: Prototyping model in an incremental style

The process model followed during the software development life cycle (SDLC) for Corynebase is shown in Figure 3.3.1. As we were facing a lot of problems in the reliable data collection, we managed the software process by using a combination of prototyping paradigms of software engineering in an incremental style. As it was not possible to gather all the software requirements in the beginning, due to problems faced during data collection, we divided the project into modules and started developing prototypes for each module and refined the requirements specifications during the prototype testing. We iterated the development and testing of the modules along side adding the functional and data increments to working modules. The same approach was followed for each module. As we started to implement this software project, we have managed to implement large part of the work, so it was not appropriate to discard this paradigm later, when we came across the idea of prototyping in combination with the incremental paradigm. And this combination worked successfully during the whole SDLC.

3.4 Application Architecture

The architecture of the application is a two tier client server application, where a client is the software application getting services form the database server at the backend. Data from various public servers is parsed by the parser and stored in the database in the form of relations. The data from the Lab is also stored in the data repository. The client application will be installed on the machines of users connected in local area network and their requests will be served by the data stored at the database server. A layout of the application architecture is shown in figure 3.4.1.

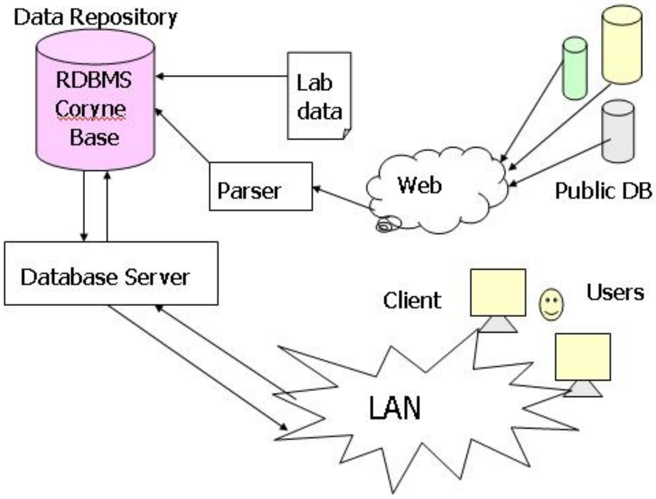


Figure 3.4.1: Application architecture

3.5 Software/Hardware Requirements

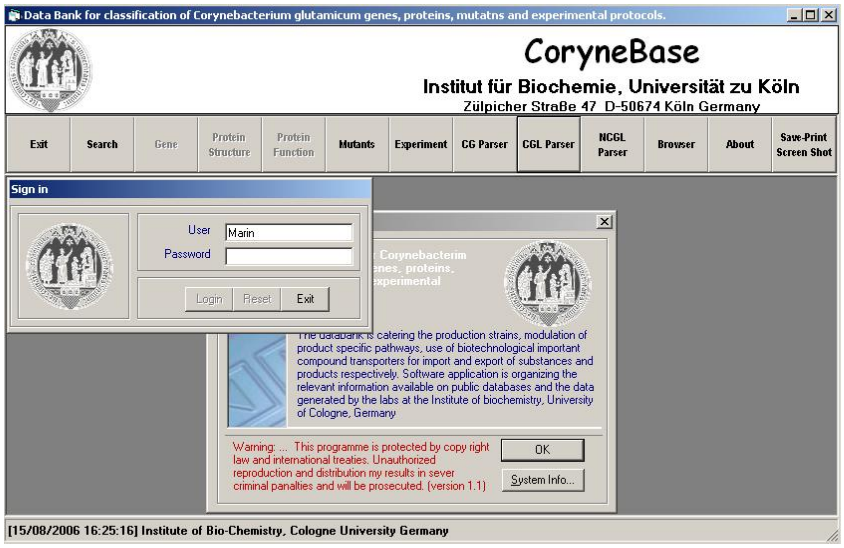
For the accomplishment we selected tools from Microsoft's Windows platform as Windows-based software was specified in the requirement specification. We chose Visual Basic for the front end design and SQL server was used as back end data repository

The following were the minimum software and hardware requirements for the machine to run the software:

- Operating system Windows 2000, XP or later.
- Processor of 500 Mega hertz. Preferably Pentium-III or more advanced processors.
- RAM having a memory of at least 128 MB.
- Hard disk of approximately 10 GB or more.

4 Results & Analysis CORYNEBASE

Figure 4.1 shows the main interface of the Corynebase which is a Multiple Document Interface (MDI), which acts as a container for the remaining windows of the application. A splash window showing the title of the application, version and copy right information is shown when the application is started, before the program has been loaded successfully. A user login system was also implemented in order to restrict unauthorized access. For example the function of parsing the text files is not accessible to every one, but only to the administrator. Because the application provides the facility to enter user comments about a particular gene, mutant or experiments protocol, it was necessary to implement this login system to track the actions performed by different users. An example of such an action is to know which comment has been given by which lab worker. In each window of the user interface it is possible to navigate either by search results or by neighbouring genes.



Data Bank for classification of Corynebacterium glutamicum genes, proteins, mutatns and experimental protocols.

CoryneBase

Institut für Biochemie, Universität zu Köln

Zülpicher Straße 47 D-50674 Köln Germany

Sign in

User: Marin

Password:

... Corynebacterim ...enes, proteins, ...xperimental

The databank is catering the production strains, modulation of product specific pathways, use of biotechnological important compound transporters for import and export of substances and products respectively. Software application is organizing the relevant information available on public databases and the data generated by the labs at the Institute of biochemistry, University of Cologne, Germany

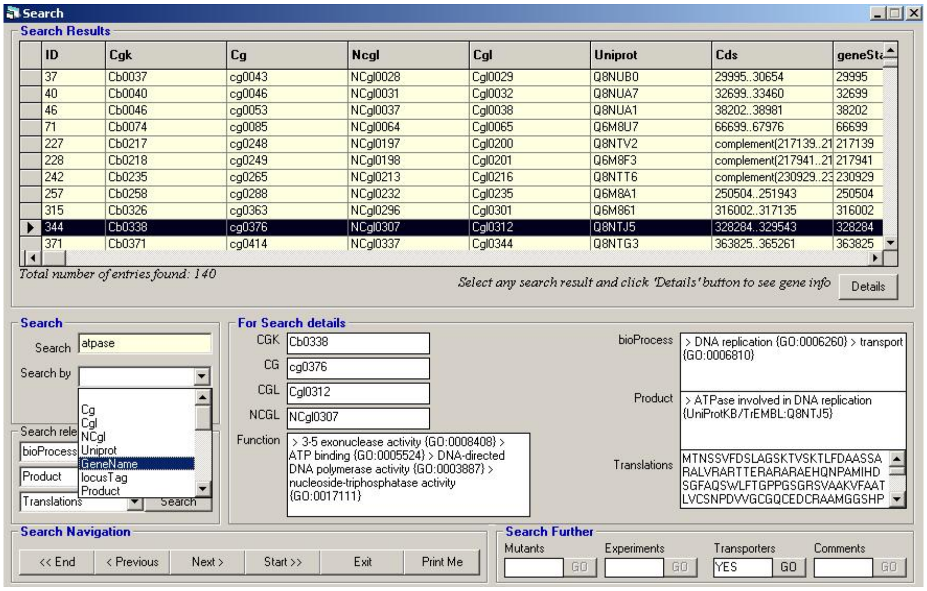
Warning: ... This programme is protected by copy right law and international treaties. Unauthorized reproduction and distribution my results in sever criminal panalties and will be prosecuted. (version 1.1)

[15/08/2006 16:25:16] Institute of Bio-Chemistry, Cologne University Germany

Figure 4.1: Corybebase main window

4.2 Search Function

Figure 4.2 is a screen shot illustrating the search function of Corynebase. With the help of this function one can search any term regarding *C. glutamicum* genes and proteins. In this window genes from all three annotations can be searched and viewed in detail. There is a search function with which a term can be searched in a specific category. The details of the search items are shown in the text boxes placed below the search results. Three other relevant parameters of the user's own choice can be searched by choosing in the combo boxes in the section of 'search relevant'. These parameters are not static; the user can change them during the search. The bounded text boxes will display the information accordingly. There is also the provision of navigating the results in the section of 'search navigation', with the help of 'start', 'end', 'next' and 'previous' buttons.



Search

Search Results

ID	Cgk	Cg	Ncgl	Cgl	Uniprot	Cds	geneSta
37	Cb0037	cg0043	NCgl0028	Cgl0029	Q8NUB0	29995..30654	29995
40	Cb0040	cg0046	NCgl0031	Cgl0032	Q8NUA7	32699..33460	32699
46	Cb0046	cg0053	NCgl0037	Cgl0038	Q8NUA1	38202..38981	38202
71	Cb0074	cg0085	NCgl0064	Cgl0065	Q6M8U7	66699..67976	66699
227	Cb0217	cg0248	NCgl0197	Cgl0200	Q8NTV2	complement(217139..21	217139
228	Cb0218	cg0249	NCgl0198	Cgl0201	Q6M8F3	complement(217941..21	217941
242	Cb0235	cg0265	NCgl0213	Cgl0216	Q8NTT6	complement(230929..23	230929
257	Cb0258	cg0288	NCgl0232	Cgl0235	Q6M8A1	250504..251943	250504
315	Cb0326	cg0363	NCgl0296	Cgl0301	Q6M861	316002..317135	316002
344	Cb0338	cg0376	NCgl0307	Cgl0312	Q8NTJ5	328284..329543	328284
371	Cb0371	cg0414	NCgl0337	Cgl0344	Q8NTG3	363825..365261	363825

Total number of entries found: 140

Select any search result and click 'Details' button to see gene info

Search

Search: atpase

Search by:

Search rele...: bioProcess, Product, Translations

For Search details

CGK: Cb0338

CG: cg0376

CGL: Cgl0312

NCGL: NCgl0307

Function: > 3-5 exonuclease activity {GO:0008408} > ATP binding {GO:0005524} > DNA-directed DNA polymerase activity {GO:0003887} > nucleoside-triphosphatase activity {GO:0017111}

bioProcess: > DNA replication {GO:0006260} > transport {GO:0006810}

Product: > ATPase involved in DNA replication {UniProtKB/TrEMBL:Q8NTJ5}

Translations: MTNSSVFDSLAGSKTVSKTLFDAASSA RALVRARTTERARARAEHQNPAMIHD SGFAQSWLFTGPPGSGRSVAAKVFAAT LVCSNPDVVGCGQCEDCRAAMGGSHP

Search Further

Mutants:

Experiments:

Transporters: YES

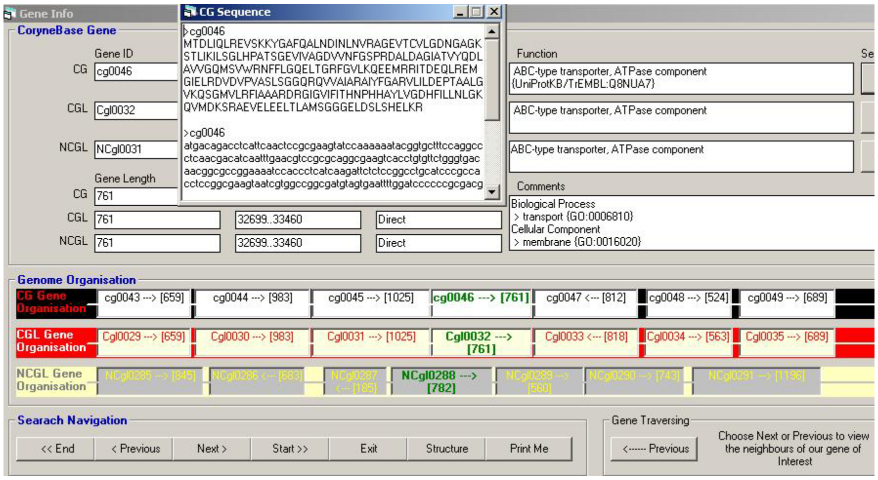
Comments:

Figure 4.2: Search window

The total number of entries found in the search is also show in the lower left corner of the window. The 'Details' button shows the genome organization of the selected gene in a new window.

4.3 Genome Organization

Figure 4.3 is a snap shot from the module of genome organization. In this window the selected gene in the search window is shown along with details regarding the gene ids, gene names, synonyms, gene lengths, starting-ending position, functions, biological processes, cellular components and direction of the genes in all three annotations. The amino acid sequence as well as the DNA sequence can be obtained in a new window by clicking the 'show sequence' button in front of the relevant annotation. The location of genes and their neighbouring genes are compared graphically, so that one can have a comparative view of the differences in the gene of interest, neighbouring genes and inter-gene distances in the three annotations simultaneously. Gene id's written in color point to the gene of interest, while every gene in other annotations has its own color scheme. An indication of the involvement of a particular protein in the mutants and experiments information is also given and can be viewed in detail by the relevant 'show' buttons in the 'further information' section. Search navigation is also possible directly form this screen. In case of this search, the screen is refreshed and shows the results for the next entry.



```
>cg0046
MTDLIQLREVSKKYGAFQALNDINLNVRAGEVTCVLGDNGAGK
STLIKILSGLHPATSGEVIVAGDVVNFGSPRDALDAGIATVYQDL
AVVGQMSVWRNFFLGQELTGRFGVLKQEEMRRITDEQLREM
GIELRDVDVPVASLSGGQRQVVAIARAIYFGARVLILDEPTAALG
VKQSGMVLRFIAAARDRGIGVIFITHNPHHAYLVGDHFILLNLGK
QVMDKSRAEVELEELTLAMSGGGELDSLSHELKR

>cg0046
atgacagacctcattcaactccgcgaagtatccaaaaaatacggtgctttccaggcc
ctcaacgacatcaatttgaacgtccgcgcaggcgaagtcacctgtgttctgggtgac
aacggcgccggaaaatccaccctcatcaagattctctccggcctgcatcccgcca
cctccggcgaagtaatcgtggccggcgatgtagtgaattttggatccccccgcgacg
```

Gene ID: CG cg0046; CGL Cgl0032; NCGL NCgl0031

Function: ABC-type transporter, ATPase component {UniProtKB/TrEMBL:Q8NUA7}; ABC-type transporter, ATPase component; ABC-type transporter, ATPase component

Gene Length: CG 761; CGL 761 | 32699..33460 | Direct; NCGL 761 | 32699..33460 | Direct

Comments: Biological Process > transport {GO:0006810} Cellular Component > membrane {GO:0016020}

CG Gene Organisation: cg0043 ---> [659] | cg0044 ---> [983] | cg0045 ---> [1025] | cg0046 ---> [761] | cg0047 <--- [812] | cg0048 ---> [524] | cg0049 ---> [689]

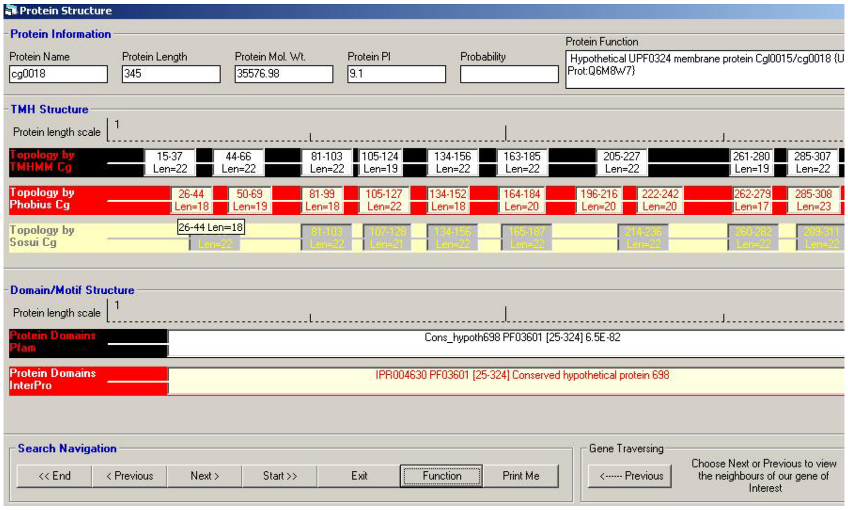
CGL Gene Organisation: Cgl0029 ---> [659] | Cgl0030 ---> [983] | Cgl0031 ---> [1025] | Cgl0032 ---> [761] | Cgl0033 <--- [818] | Cgl0034 ---> [563] | Cgl0035 ---> [689]

NCGL Gene Organisation: NCgl0285 ---> [845] | NCgl0286 <--- [683] | NCgl0287 <--- [185] | NCgl0288 ---> [782] | NCgl0289 ---> [560] | NCgl0290 ---> [743] | NCgl0291 ---> [1196]

Figure 4.3: Genome organization

4.4 Protein Structure

The screen shot in Figure 4.4 shows a part of the module of Protein structure information. Protein name, length, molecular weight (MW), isoelectric point (pI), and the probability of a protein of being trans-membrane are shown in the 'protein information' section. The topology of the trans-membrane proteins predicted by three different prediction servers, namely, TmHMM, Phobius and SOSUI is compared graphically based on the data from the Cg annotation. Trans-membrane helices are shown as boxes according to their lengths as well as the starting and ending positions of residues forming a trans-membrane region. Every protein has been adjusted on a 100% scale of length starting from 1 to the total length of the protein. This was done so that



Protein Name: cg0018; Protein Length: 345; Protein Mol. Wt.: 35576.98; Protein PI: 9.1; Probability:

Protein Function: Hypothetical UPF0324 membrane protein Cgl0015/cg0018 {U Prot:Q6M8W7}

Topology by TMHMM Cg: 15-37 Len=22 | 44-66 Len=22 | 81-103 Len=22 | 105-124 Len=19 | 134-156 Len=22 | 163-185 Len=22 | 205-227 Len=22 | 261-280 Len=19 | 285-307 Len=22

Topology by Phobius Cg: 26-44 Len=18 | 50-69 Len=19 | 81-99 Len=18 | 105-127 Len=22 | 134-152 Len=18 | 164-184 Len=20 | 196-216 Len=20 | 222-242 Len=20 | 262-279 Len=17 | 285-308 Len=23

Topology by Sosui Cg: 26-44 Len=18 | Len=22 | 81-103 Len=22 | 107-128 Len=21 | 134-156 Len=22 | 165-187 Len=22 | 214-236 Len=22 | 260-282 Len=22 | 289-311 Len=22

Protein Domains Pfam: Cons_hypoth698 PF03601 [25-324] 6.5E-82

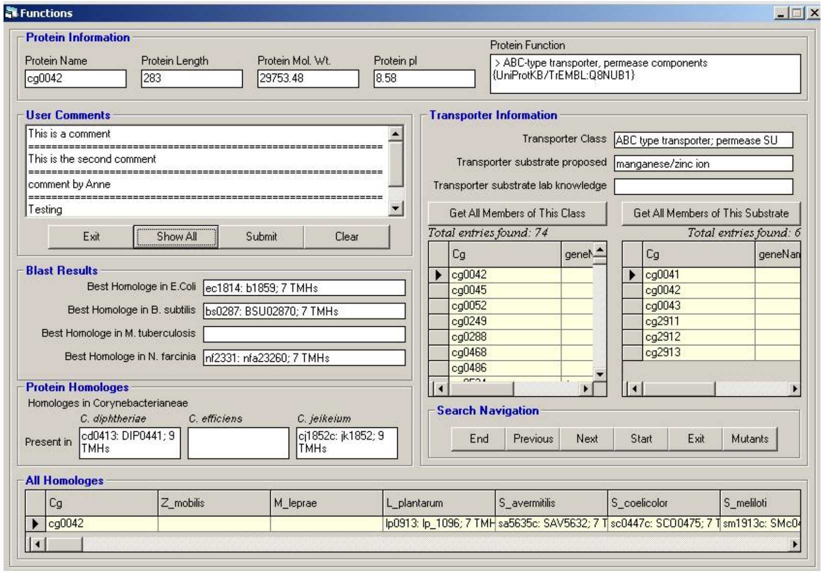
Protein Domains InterPro: IPR004630 PF03601 [25-324] Conserved hypothetical protein 698

Figure 4.4: Protein structure

proteins of varying lengths may fit and so that all the trans-membrane regions can be displayed in a specified area of the screen. The domain/motif structure of the protein in question has also been covered by this module in the section of ‘domain/motif structure’ where protein domains are predicted with Pfam and Interpro. Search navigation is also possible directly form this screen. In the case of this search, the screen is refreshed and shows the results for the next entry.

4.5 Protein Functions, Transporters and Homologues

Figure 4.5 shows the results from the implementation of the module for the information regarding transporters, substrates, and homologues. The protein information is the same as in the previous screen shot. This is a convenience in viewing because the user could come directly from the genome organization to this window without viewing the protein structure information. In the ‘transporter identification’ section of figure 4.4 the transporter and the substrate are shown along with a provision of getting all members of this transporter class and all members of this substrate in a separate grid. Information about the best homologues of other related bacterias is shown in the section of ‘Blast Results’ and homologues in *Corynebacterianeae* are shown in the section of ‘Protein Homologues’. A grid at the bottom of figure 4.4 with section heading, named ‘All Homologues’ shows all other homologues. There is a provision of submitting user comments as well as viewing the comments given by other users of the application.



Protein Name: cg0042
Protein Length: 283
Protein Mol. Wt.: 29753.48
Protein pI: 8.58
Protein Function: > ABC-type transporter, permease components {UniProtKB/TrEMBL:Q8NUB1}

This is a comment
==
This is the second comment
==
comment by Anne
==
Testing

Transporter Class: ABC type transporter; permease SU
Transporter substrate proposed: manganese/zinc ion
Transporter substrate lab knowledge:

Total entries found: 74

	Cg	geneN
▶	cg0042	
	cg0045	
	cg0052	
	cg0249	
	cg0288	
	cg0468	
	cg0486	

Total entries found: 6

	Cg	geneNar
▶	cg0041	
	cg0042	
	cg0043	
	cg2911	
	cg2912	
	cg2913	

Best Homologe in E.Coli: ec1814: b1859; 7 TMHs
Best Homologe in B. subtilis: bs0287: BSU02870; 7 TMHs
Best Homologe in M. tuberculosis:
Best Homologe in N. farcinia: nf2331: nfa23260; 7 TMHs

Homologes in Corynebacterianeae

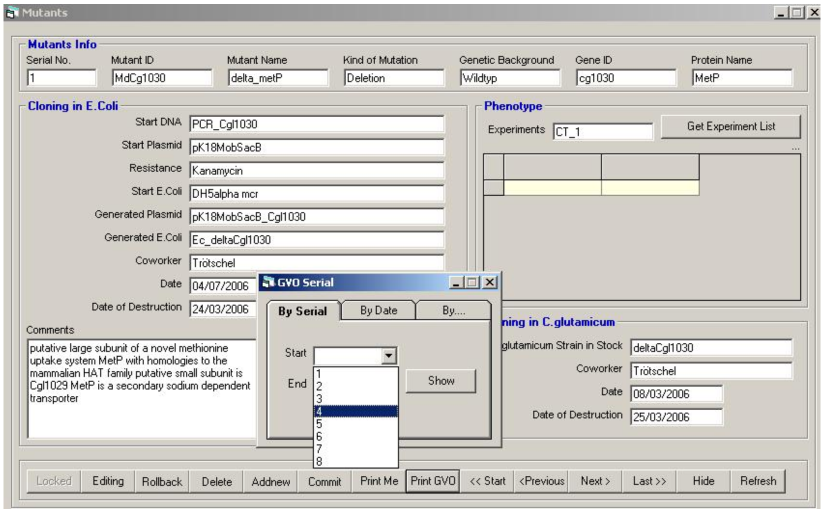
	C. diphtheriae	*C. efficiens*	*C. jeikeium*
Present in	cd0413: DIP0441; 9 TMHs		cj1852c: jk1852; 9 TMHs

	Cg	Z_mobilis	M_leprae	L_plantarum	S_avermitilis	S_coelicolor	S_meliloti
▶	cg0042			lp0913: lp_1096; 7 TMH	sa5635c: SAV5632; 7 T	sc0447c: SCO0475; 7 T	sm1913c: SMc0

Figure 4.5: Protein functions, transporters and homologues

4.6 Mutants

Figure 4.6 is a snap shot from the module named Mutant information management which provides data regarding mutant ids, mutant names, kinds of mutations, genetic backgrounds along with the cloning information in *C. glutamicum* as well as in *E. coli* and relevant experiment ids. The function of adding new records, deleting/editing the existing information and printing are implemented. Figure 4.6[a] is a snap shot of a special report for ‘risk evaluation of genetically modified work’’ used for higher administration of the institute in order to get approval for genetic medications. In this module



Serial No.	Mutant ID	Mutant Name	Kind of Mutation	Genetic Background	Gene ID	Protein Name
1	MdCg1030	delta_metP	Deletion	Wildtyp	cg1030	MetP

Start DNA: PCR_Cgl1030
Start Plasmid: pK18MobSacB
Resistance: Kanamycin
Start E.Coli: DH5alpha mcr
Generated Plasmid: pK18MobSacB_Cgl1030
Generated E.Coli: Ec_deltaCgl1030
Coworker: Trötschel
Date: 04/07/2006
Date of Destruction: 24/03/2006

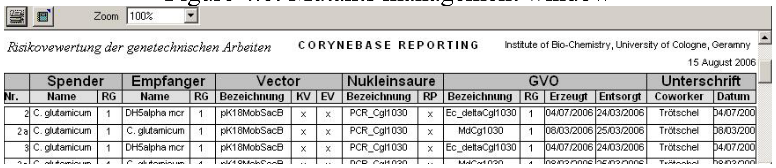
Comments: putative large subunit of a novel methionine uptake system MetP with homologies to the mammalian HAT family putative small subunit is Cgl1029 MetP is a secondary sodium dependent transporter

Experiments: CT_1

Start:
End: 1 2 3 4 5 6 7 8

glutamicum Strain in Stock: deltaCgl1030
Coworker: Trötschel
Date: 08/03/2006
Date of Destruction: 25/03/2006

Figure 4.6: Mutants management window



Risikovewertung der genetechnischen Arbeiten CORYNEBASE REPORTING Institute of Bio-Chemistry, University of Cologne, Geramny

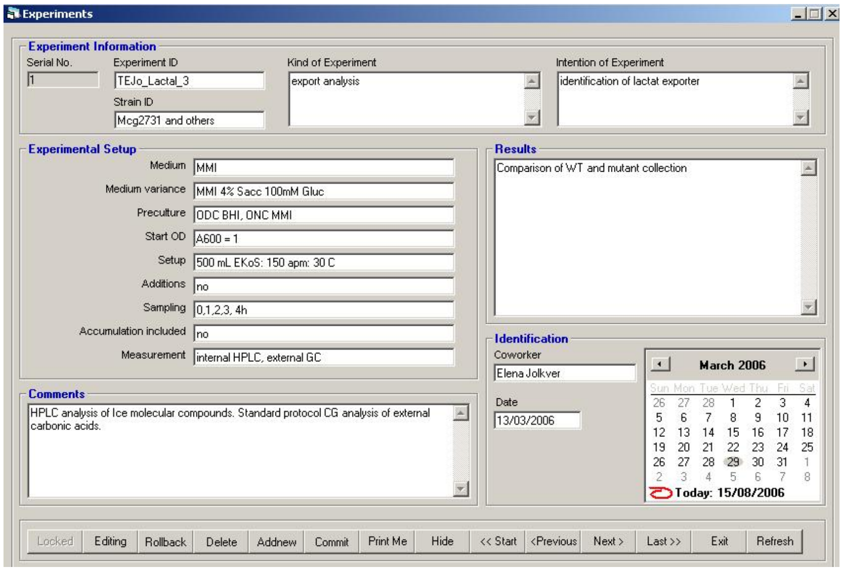
15 August 2006

	Spender		Empfanger		Vector			Nukleinsaure		GVO				Unterschrift	
Nr.	Name	RG	Name	RG	Bezeichnung	KV	EV	Bezeichnung	RP	Bezeichnung	RG	Erzeugt	Entsorgt	Coworker	Datum
2	C. glutamicum	1	DH5alpha mcr	1	pK18MobSacB	x	x	PCR_Cgl1030	x	Ec_deltaCgl1030	1	04/07/2006	24/03/2006	Trötschel	04/07/200
2 a	C. glutamicum	1	C. glutamicum	1	pK18MobSacB	x	x	PCR_Cgl1030	x	MdCg1030	1	08/03/2006	25/03/2006	Trötschel	08/03/200
3	C. glutamicum	1	DH5alpha mcr	1	pK18MobSacB	x	x	PCR_Cgl1030	x	Ec_deltaCgl1030	1	04/07/2006	24/03/2006	Trötschel	04/07/200

Figure 4.6[a]: Report for risk evaluation of genetic work

there is a provision of submitting user comments as well as viewing the comments regarding mutants given by other users of the application. All the user data entry validation control that prevent invalid data entries have been implemented, too.

4.7 Experiments



Experiment Information

Serial No.: 1

Experiment ID: TEJo_Lactal_3

Strain ID: Mcg2731 and others

Kind of Experiment: export analysis

Intention of Experiment: identification of lactat exporter

Experimental Setup

Medium: MMI

Medium variance: MMI 4% Sacc 100mM Gluc

Preculture: ODC BHI, ONC MMI

Start OD: A600 = 1

Setup: 500 mL EKoS: 150 apm: 30 C

Additions: no

Sampling: 0,1,2,3, 4h

Accumulation included: no

Measurement: internal HPLC, external GC

Results: Comparison of WT and mutant collection

Identification

Coworker: Elena Jolkver

Date: 13/03/2006

March 2006

Sun	Mon	Tue	Wed	Thu	Fri	Sat
26	27	28	1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	1
2	3	4	5	6	7	8

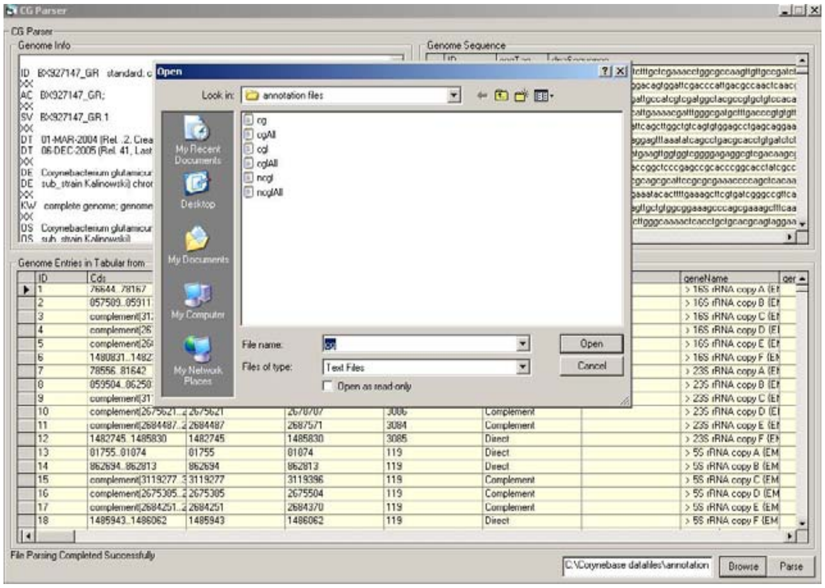
Today: 15/08/2006

Comments: HPLC analysis of Ice molecular compounds. Standard protocol CG analysis of external carbonic acids.

Figure 4.7: Experimental protocols management window

Figure 4.7 show a snap shot from the module for Experiments setups or protocols information management which provides data regarding experiment ids, type and purpose of the experiment, strain ids, complete information regarding experimental setups, along with results information in *C. glutamicum*. The function of adding new records, deleting/editing the existing information and printing are also implemented in this module. This module also contains the provision of submitting user comments as well as viewing the comments concerning a particular experiment given by other users of the application. The user data entry validation is performed to restrict the invalid data entries.

4.8 Parsing Facilities



```
ID   BX927147_GR   standard; c
XX
AC   BX927147_GR;
XX
SV   BX927147_GR.1
XX
DT   01-MAR-2004 (Rel. .2, Crea
DT   06-DEC-2005 (Rel. 41, Last
XX
DE   Corynebacterium glutamicur
DE   sub_strain Kalinowski) chror
XX
KW   complete genome; genome
XX
OS   Corynebacterium glutamicur
OS   sub_strain Kalinowski)
```

```
tctttgctcgaaacctggcgccaagttgttgccgatcl
ggacagtggattcgacccattgacgccaactcaac
gattgccatcgtcgatggctacgccgtgctgtccaca
cattgaaaacgatttgggcgatgctttgacccgtgtgtt
attcagcttggctgtcagtgtggagcctgagcaggaa
aggagtttaaatatcagcctgacgcacctgtgatctcl
atgaagttggtggtcggggagaggcgtcgacaagc
accggctcccgagccgcacccggcacctatcgcc
cgcagcgcattccgcgcgaaaccccagctcacaa
gaaatacacttttgaaagcttcgtgatcgggccgttca
agttgctgtggcggaaagcccagcgaaagctttcaa
cttgggcaaaactcacctgctgcacgcagtaggaa
```

Look in: annotation files

cg
cgAll
cgl
cglAll
ncgl
ncglAll

File name: cg

Files of type: Text Files

☐ Open as read only

ID	Cds						geneName	ger
1	76644..78167						> 16S rRNA copy A (EI	
2	057509..05911						> 16S rRNA copy B (EI	
3	complement(31:						> 16S rRNA copy C (EI	
4	complement(26						> 16S rRNA copy D (EI	
5	complement(26						> 16S rRNA copy E (EI	
6	1480831..1482						> 16S rRNA copy F (EI	
7	78556..81642						> 23S rRNA copy A (EI	
8	059504..06250						> 23S rRNA copy B (EI	
9	complement(31						> 23S rRNA copy C (EI	
10	complement(2675621..2	2675621	2678707	3086	Complement		> 23S rRNA copy D (EI	
11	complement(2684487..2	2684487	2687571	3084	Complement		> 23S rRNA copy E (EI	
12	1482745..1485830	1482745	1485830	3085	Direct		> 23S rRNA copy F (EI	
13	01755..01074	01755	01874	119	Direct		> 5S rRNA copy A (EM	
14	862694..862813	862694	862813	119	Direct		> 5S rRNA copy B (EM	
15	complement(3119277..3	3119277	3119396	119	Complement		> 5S rRNA copy C (EM	
16	complement(2675305..2	2675305	2675504	119	Complement		> 5S rRNA copy D (EM	
17	complement(2684251..2	2684251	2684370	119	Complement		> 5S rRNA copy E (EM	
18	1485943..1486062	1485943	1486062	119	Direct		> 5S rRNA copy F (EM	

File Parsing Completed Successfully

C:\Corynebase datafiles\annotation

Figure 4.8: Parsing function window

The last module illustrates the parsing facilities which can be used for updating the Corynebase data after the re-annotations. The parsing module works for all the three main publicly available annotations, namely Cg, Cgl and NCgl. The screen shot in figure 4.8 from the parser contains a 'browse' button which loads the input text file containing the complete genome as single line database entries. The button 'parser' validates the input text file, whether it is the right genome annotation, and then generates data in a format required for insertion into the relational database tables automatically. The parser extracts all the parameters along with the amino acid as well as the DNA sequence. This module was built in the beginning of the project when we were facing a lot of problems regarding data reliability. Anyhow, the parsing component of Corynebase solved this problem and provided a way to get data directly from the source of annotation.

5 Discussion

Corynebacterium glutamicum is one of the important microorganisms in biotechnology [1, 19]. Besides the production of glutamate [4, 5, 25] as flavour enhancer, other amino acids and nucleotides are central components of the increasing million dollar market of fermentative substrate production by heterotrophic bacteria. Because of the significance of this particular strain the genome sequence was determined in more than three different independent projects. However, in spite of, or because of, the industrial use of *C. glutamicum* the derived genome information is not assembled in a central public database.

In this project we tried to cover the limitations and problems previously faced by the scientists working on *C. glutamicum* while accessing a variety of information through various public databases. An important example of such information is the whole genome itself, which has been annotated in three different countries [41-42, 55]. All the other relevant information, such as protein functions, transporters, substrates, homologues, mutants and experiments was also redundant and stored in various data sources. With this work, it is now possible to view the complete genome information from three annotations. The project implementation covered a graphical viewing of the genome. It is also possible to search a variety of information such as gene id's, enzyme classification numbers, trans mebrane helices, biological process, cellular components, transporter substrates, mutants, etc. All this was not easy previously while scanning through different text files and websites. This project also implemented a parsing module to make the software updatable instantly if some new data becomes available in case of the re-annotation.

In order to shed light on metabolic or regulatory networks the integrated understanding of such information is essential for research at the systems biology level. Thus, within this project we summarized most of the relevant *in silico* information about three different genome data sets, derived protein data regarding structures and functions as well as the relation to other proteins of related organisms. In combination with an customized search function for different parameters and a graphical user interface, we provide a tool for daily laboratory purpose. The information is helpful for the comparison of all genome annotations regarding gene locations, neighboring genes, gene orientation or lengths important for construction of mutants and promotor analyses.

In collaboration with the Kraemer lab a special focus was directed to the identification and description of membrane proteins that could act as transporters. The comparison of three different trans-membrane domain predictions as well as the categorization of membrane proteins according to the prediction of transporter classes and substrates is included. Additionally, the developed software can be used for managing mutant stock information and collection of experimental setups. The standalone designed and developed relational database application, named Corynebase, ultimately represents a tool for planning experiments in molecular biology, biochemistry as well as physiology of the microorganism *C. glutamicum* comprising all given objectives.

6 Conclusions

In conclusion we will say that, according to our aims and objectives, we completed this project along with its setup accessible over the local area network. We conclude that with this work, now it is easy to search any gene or protein of *C. glutamicum* regardless of the fact which annotation source the gene belongs to. We learned from this project that there are variations in different gene and protein parameters among three annotations which can be identified using Corynebase for further scientific investigation. We also conclude that integrated information regarding genes, proteins, mutants, transporter classes and substrates, provided by the application may help research at a systems biology level. This kind of information is used for the construction of mutants and promoter analysis, so we can conclude that Corynebase could be a handy tool to use in the laboratory. We also learned form this project that it made the identification of membrane proteins which may act as transporters by providing the protein information from three different prediction servers. This also led to the categorization of membrane proteins according to the prediction of transporter classes and substrates.

We also learned that, with Corynebase, managing mutant stock information, collection of experimental protocols and parsing the annotation files became easier. This work also provided a combination of extracting available information from already available public data sources and from the data generated by the labs itself. This integration also included the automatic reporting service for 'Risikobewertung gentechnischer Arbeiten' (Risk evaluation of genetic work) which was done manually before. This reduced the time for the group leaders for this administrative task done in an automatic way.

Future Work

We handled three main public annotations in our project, so in the future further annotation can be incorporated to have a further comparison. We covered the trans-membrane prediction and protein motifs from different sources only for the cg annotation due to the time limitations. This project can further be expanded to cover the same for the Cgl and Ncgl annotations, too. Software project is never free of complications. System is under continuous testing by the group members. We are sure this testing will definitely come up with new ideas and innovations to be added. So there shall be chances of improving or extending the current work.

References

[1] Kinoshita, S. (1985) Glutamic acid bacteria, p. 115-146. *In* A. L. Demain and N. A. Solomon (ed.), Biology of industrial microorganisms. Cummings, London, United Kingdom

[2] Malumbres, M., L. M. Mateos, and J. F. Martin. (1995) Microorganisms for amino acid production: *Escherichia coli* and corynebacteria, p. 423-469. *In* Y. H. Hui and G. G. Kachatourians (ed.), Food biotechnology microorganisms, vol. 2. VCH Publishers, New York, N.Y.

[3] Nakayama K, Kitada S, Kinoshita S. (1961) Studies on lysine fermentation I. The control mechanism on lysine accumulation by homoserine and threonine. *J Gen Appl Microbiol*. 7:145–154.

[4] Udaka S. (1960) Screening method for microorganisms accumulating metabolites and its use in the isolation of *Micrococcus glutamicus. J Bacteriol*. 79:754–755.

[5] Kinoshita S, Udaka S, Shimono M. (1957) Studies on the amino acid fermentation. Part I. Production of L-glutamic acid by various microorganisms. *J Gen Appl Microbiol*. 3:193–205.

[6] Shiio I, Nakamori S. (1970) Microbial production of L-threonine. Part II. Production by α-amino-β-hydroxyvaleric acid resistant mutants of glutamate producing bacteria. *Agric Biol Chem*. 34:448–456.

[7] Udaka S, Kinoshita S. (1958) Studies on L-ornithine fermentation. I. The biosynthetic pathway of L-ornithine in *Micrococcus glutamicus. J Gen Appl Microbiol*. 4:272–282.

[8] Nunheimer TD, Birnbaum J, Ihnen ED, Demain AL. (1970) Product inhibition of the fermentative formation of glutamic acid. *Appl Microbiol*. 20:215–217.

[9] Shiio I, Otsuka S, Takahashi M. (1962) Effect of biotin on the bacterial formation of glutamic acid. I. Glutamate formation and cellular permeability of amino acids. *J Biochem*. 51:56–62.

[10] Takinami K, Yoshii H, Tsuri H, Okada H. (1965) Biochemical effects of fatty acid and its derivatives on L-glutamic acid fermentation. Part III. Biotin-Tween 60 relationship in the accumulation of L-glutamic acid and the growth of *Brevibacterium lactofermentum. Agric Biol Chem*. 29:351–359.

[11] Hoischen C, Krämer R. (1990) Membrane alteration is necessary but not sufficient for effective glutamate secretion in *Corynebacterium glutamicum. J Bacteriol*. 172:3409–3416.

[12] Gutmann M, Hoischen C, Krämer R. (1992) Carrier-mediated glutamate secretion by *Corynebacterium glutamicum* under biotin limitation. *Biochim Biophys Acta*. 1112:115–123.

[13] Kimura E, Yagoshi C, Kawahara Y, Ohsumi T, Nakamatsu T, Tokuda H. (1999) Glutamate overproduction in *Corynebacterium glutamicum* triggered by a decrease in the level of a complex comprising DtsR and a biotin-containing subunit. Biosci Biotechnol Biochem. 63:1274–1278.

[14] Wolfgang L, Klaus H and Karlheinz D. (2005) Biotechnological production of amino acids and derivatives: current status and prospects. *Applied Microbiology and Biotechnology*. 69: 1-8.

[15] Hermann, S, 2005, Institute of Biotechnology, Forschungszentrum Julich GmbH, Julich, Germany 'Foreword' in 'Handbook of *Corynebacterium glutamicum'* CRC press Taylor & Francis Group, Boca Raton, FL, USA.

[16] Lothar, K and Michael, B (ed.) 2005, 'Handbook of *Corynebacterium glutamicum*' CRC press Taylor & Francis Group, Boca Raton, FL, USA.

[17] Fudou R, Jojima Y, Seto A, Yamada K, Kimura E, Nakamatsu T, Hiraishi A, and Yamanaka S. (2002) *Corynebacterium efficiens* sp. Nov., a glutamic acid producing species from soil and vegetables. *Int. J. Syst. Evol. Micorbiol*. 52:1127-1131.

[18] Ikeda, K. (2002) New seasonings [translation]. *Chem. Senses* 27:847-849.

[19] Nobuaki S, Satoshi O, Hiroshi N, Yota T, Masayuki I, and Hideaki Y. (2005) Large-Scale Engineering of the *Corynebacterium glutamicum* Genome. *Applied and Environmental Microbiology*. 6: 3369-3372.
[20] Takashi H, Masaaki W, and Kazuo N. (2001) L-Glutamate production by lysozyme-sensitive *Corynebacterium glutamicum ltsA* mutant strains. *BMC Biotechnol*. 1: 9.
[21] Ikeda M and Nakagawa S. (2003) The *Corynebacterium glutamicum* genome: features and impacts on biotechnological processes. *Appl. Microbiol. Biotechnol*. 62: 99-109.
[22] Tauch A, Homann I, Mormann S, Ruberg S, Billault A, Bathe B, Brand S, Brockmann Gretza O, Ruckert C, Schischka N, Wrenger C, Hoheisel J, Mockel B, Huthmacher K, Pfefferle W, Puhler A, and Kalinowski J. (2002) Strategy to sequence the genome of *Corynebacterium glutamicum* ATCC 13022: use of a cosmid and a bacterial artificial chromosome library. *J. Biotechnol*. 95: 25-38.
[23] Nishio Y, Nakamura Y, Kawarabayasi Y, Usuda Y, Kimura E, Sugimoto S, Matsui K, Yamagishi A, Kikuchi H, Ikeo K, and Gojobori T. (2003) Comparative complete genome sequence analysis of the amino acid replacements responsible for the thermostability of *Corynebacterium efficiens*. *Genome Res*. 13:1572-1579.
[24] Kinoshita, S. Thom Award Address. (1987) Amino acid and nucleotide fermentations: From their genesis to the current state. *Developments in Industrial Microbiology* 28:1-12.
[25] Kinoshita S, Udaka S, and Shimono M. (1957) Studies on the amino acid fermentation. I. Production of L-glutamic acid by various microorganisms. *J. Gen. Appl. Microbiol*. 3: 193-205.
[26] Schrumpf B, Eggeling L, and Sahm H. (1992) Isolation and prominent characteristics of an l-lysine hyperproducing strain of *Corynebacterium glutamicum*. *Appl. Microbiol. Biotechnol*. 37:566-571.
[27] Lange C, Rittmann D, Wendisch V.F, Bott M, and Sahm H. (2003) Global Expression Profiling and Physiological Characterization of *Corynebacterium glutamicum* Grown in the Presence of L-Valine. *Appl Environ Microbiol*. 69(5): 2521–2532.
[28] Radmacher, E. Vaitsikova, A Burger, U. Krumbach, K. Sahm, H. and Eggeling, L. (2002) Linking central metabolism with increased pathway flux: l-valine accumulation by *Corynebacterium glutamicum Appl. Environ. Microbiol*. 68:2246-2250.
[29] Liebl W, Ehrmann M, Ludwig W, and Schleifer KH. (1991) Transfer of *Brevibacterium divaricatum* DSM 20297, *B. falvum* DSM 20411, *B. lactofermentum* DSM 20412 & 1412 and *C. lilium* DSM 20137 to *C. glutamicum* and their distinction by rRNA gene restriction patterns. *Int J. Syst. Bacteriol*. 41:255-260.
[30] Barksdale L. (1970) *Corynebacterium diphtheriae* and its relatives*. Bacteriol. Rev*. 34:378-422.
[31] Nikaido H. (1994) Prevention of drug access to bacterial targets: permeability barriers and active efflux. *Science* 264:382.
[32] Kartmann B, Stengler S, and Niederweis M. (1999) Porins in the cell wall of *Mycobacterium tuberculosis. J. Bacteriol*. 181:6543.
[33] Lichtinger T, Burkovski A, Niederweis M, Kramer R, and Benz R. (1998) Biochemical and biophysical characterization of the cell wall porins of *Corynebacterum glutamicum*: the channel is formed by a low molecular mass polypeptide. *Biochemistry* 37:15024.
[34] Ikeda M. (2003) Amino acid production processes. *Adv. Biochem. Eng*. 79:2-35.
[35] Mori M and Shiio I. (1983) Glutamate transport and production in *Brevibacterium flavum. Agric. Biol. Chem*. 47:983-990
[36] Clement Y, Escoffier B, Trombe MC, and Laneelle G. (1984) Is glutamate excreted by its uptake system? A working hypothesis. *J. Gen. Microbiol*. 130:2589-2594
[37] Luntz MG, Zhdanova NI, and Bourd GI. (1986) Transport and excretion of L-Lysine in *Corynebacterium glutamicum. J. Gen. Microbiol*. 132:2137-2146.
[38] Milner JL and Wood JM. (1987) Transmembrane amino acid fluxes in bacterial cells. CRC *Crit. Rev. Biotechnol*. 5:1-47.
[39] Kramer R. (1994) Secretion of amino acids by bacteria: physiology and mechanism *FEMS Microbiol. Rev*. 13:75-94.
[40] Hoischen C and Kramer R. (1989) Evidence for an efflux carrier system involved in the secretion of glutamate by *Corynebacterium glutamicum. Arch. Microbiol*. 172:3409-3416.
[41] Nakagawa S. (2000) Complete genomic sequence of *Corynebacterium glutamicum* ATCC 13032. submitted to EMBL, GenBank, DDBJ databases.
[42] Kalinowski J, Bathe B, Bartels D, Bischoff N, Bott M, Burkovski A, Dusch N, Eggeling L, Eikmanns BJ, Gaigalat L, Goesmann A, Hartmann M, Huthmacher K, Kraemer R, Linke B, McHardy AC, Meyer F, Moeckel B, Pfefferle W, Puehler A, Rey DA, Rueckert C, Rupp O, Sahm H, Wendisch VF, Wiegraebe I

and Tauch A. (2003) The complete *Corynebacterium glutamicum* ATCC 13032 genome sequence and its impact on the production of L-aspartate-derived amino acids and vitamins' *J. Biotechnol*. 104:5-25.

[43] Hector GM, Jeffrey DU and Jennifer DW. (2002) The Worlds of Database Systems in '*Database Systems: The Complete Book'*

[44] Codd EF. (1970) 'A relational model for large shared data banks,' *Comm. ACM,* 13(6): 377-387.

[45] Abiteboul S, Hull R, and Vianu V. (1995) 'Foundations of Databases', Addison Wesley, Reading, MA.

[46]Stonebraker M, and Hellerstein JM. (1998) 'Readings in Database Systems' (eds.), Morgan Kaufmann, San Francisco.

[47] Peter N, and Brian. R, (1969) 'Software Engineering' (eds.) Garmisch, Germany, 7-11 October 1968, Brussels, Scientific Affairs Division, NATO. 231pp.

[48] Roger S Pressman and Associates, 'Softwares and Software Engineering' in *Software Engineering: A Practitioner's Approach*, 6th Edition. McGraw-Hill Higher Education.

[49] Roger S Pressman and Associates, 'Process: A Generic View' in *Software Engineering: A Practitioner's Approach*, 6th Edition. McGraw-Hill Higher Education.

[50] Roger S Pressman and Associates, 'Prescriptive Process Models' in *Software Engineering: A Practitioner's Approach*, 6th Edition. McGraw-Hill Higher Education.

*[3]

[3] * Reference numbers [51] to [70] are in section 'Web Links' on Page 22.

Web Links

[51] Misra R, Richard H and Thomas G. (2005) EchoBase an integrated post genomic database for *E.Coli*. Version 1.3 Viewed on June 14, 2006, Available at: <http://www.biolws1.york.ac.uk/echobase/>

[52] CyanoBase, The genome database for *Cyanobacteria*. Viewed on June 14, 2006, Available at: <http://www.kazusa.or.jp/cyano/ >

[53] *Bacillus subtilis* Genome Database, GenomeNet, Bioinformatics Center, Institute of Chemical Research, Kyoto University, Japan. Viewed on June 14, 2006, Available at: <http://bacillus.genome.jp/ >

[54] ARAMEMNON, Plant membrane protein database for *Arabidopsis* version 3.2. Viewed on June 15, 2006, Available at: <http://aramemnon.botanik.uni-koeln.de/index.ep >

[55] National Center for Biotechnology Information NCBI. (2001) *Corynebacterium glutamicum* ATCC 13032, complete genome. Viewed on July 12, 2006, Available at:
<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=genome&cmd=search&term=NC_003450 >

[56] European Bioinformatics Institute, EBI, EMBL, Sequence Retrieval System (SRS Release 7.1.3.1. Viewed on July 12, 2006 Available at: <http://srs.ebi.ac.uk >

[57] The Universal Protein Resource, UniProt Knowledge Base (UniProtKB). Viewed on July 12, 2006, Available at: <http://www.pir.uniprot.org/database/knowledgebase.shtml >

[58] Expert Protein Analysis System (ExPASy). Viewed on July 12, 2006, Available at: <http://www.expasy.org/sprot >

[59] Prediction of trans-membrane helices in proteins, TmHMM Server v.2.0. Viewed on July 13, 2006, available at: <http://www.cbs.dtu.dk/services >

[60] A combined trans-membrane topology and signal peptide predictor, Phobius. Viewed on July 13, 2006, Available at: <http://phobius.cgb.ki.se/ >

[61] Hirokawa T. Boon-Chieng S. and Mitaku S. (1998) SOSUI Classification and secondary structure prediction for membrane proteins, *Bioinformatics* Vol.14 S.378-379. Tool available at:
<http://www.proteome.bio.tuat.ac.jp/sosui_submit.html >

[62] Protein families database of alignments and HMMs, Pfam Version 19.0. Viewed on July 14, 2006. Available at: <http://pfam.cgb.ki.se/ >

[63] European Bioinformatics Institute, EBI, EMBL, InterPro a database of protein families, domains and functional sites. Viewed on July 14, 2006 Available at: <http://www.ebi.ac.uk/interpro/ >

[64] Genomic Comparisons of Membrane Transport Systems, TransportDB. Viewed on July 14, 2006. Available at: <http://www.membranetransport.org/ >

[65] Basic Local Alignment Search Tool, BLAST version 2.2.13. Viewed on July 14, 2006. Available at: <http://www.membranetransport.org/ >

[66] Web tool for *Corynebacterium glutamicum* ATCC 13032. Viewed on October 26, 2006. Available at: <http://gib.genes.nig.ac.jp/single/keywrd/main.php?spid=Cglu_ATCC13032 >

[67] The Condensed Genome Display for the *Corynebacterium glutamicum* ATCC 13032 Bielefeld. Viewed on October 26, 2006. Available at:
http://rice.tigr.org/tigr-scripts/CMR2/PseudoGenomeDisplay2.spl?asmbl_id=861

[68] The MEROPS peptidase database, Version 7.50. Viewed on October 26, 2006. Available at:
<http://merops.sanger.ac.uk/cgi-bin/speccards?sp=sp000294&type=P>

[69] Proteins of *Corynebacterium glutamicum* database in Encyclopedia of Life Viewed on October 26, 2006. Available at: <http://pat.sdsc.edu/perl/browser.pl?tax=Corynebacterium%20glutamicum>

[70] Todd ML and Sean RE. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research,* 25(5): 955–964. Available at: <http://lowelab.ucsc.edu/GtRNAdb/ >