

©BRAND X, PHOTODISC

A Comprehensive Human Chromosome 21 Database

A User-Friendly Database Covering This Important Chromosome That Allows Efficient and Effective Use by Researchers from Different Areas of Expertise

BY CAO NGUYEN, SUPPHACHAI THAICHAROEN,
THOMAS LACROIX, KATHELEEN GARDINER,
AND KRZYSZTOF J. CIOS

Although there exist databases containing information on the genes of human chromosome 21, they are either limited in scope for chromosome 21-specific data or include this information buried within similar information for the entire human genome or multiple organism genomes. For example, the Human Chromosome 21 cSNP and MAP database contains single nucleotide polymorphisms within cDNA sequences and the map of the chromosome 21 but lacks protein function data. The Human Genome Organisation (HUGO) stores chromosome 21 data as a subset of whole human genome information. In addition, the user-interfaces of these databases were often designed in such a way that only experienced users are able to easily or fully utilize them. These non chromosome 21-dedicated or “expert-only” databases pose significant challenges for chromosome 21 research for both biologists and non-biologists alike. To ameliorate these problems we are creating a comprehensive chromosome 21 database with the goal of storing all chromosome 21 gene and protein related information. It was created using information from existing databases and from the literature. Most importantly, however, we designed a novel and easy-to-use user-interface, called GeneQuest, that enables even inexperienced users to fully utilize the database in an efficient and effective manner. The database embraces a wide range of information including chromosome 21 gene structures, protein post-translational modifications and interactions, plus chromosome 21 orthologs in model organisms plus their phenotypes of RNAi and their protein–protein interactions. In addition, we added a built-in protein–protein interaction function prediction based on Markov random field. The database can be accessed at <http://chr21db.cudenver.edu>.

Chromosome 21 is the smallest human chromosome, with the long arm (21q) consisting of only 34 Mb [1]. Recent review of genomic sequence annotation identified 384 genes and gene models [2], and current estimates have increased this to ~430 (Gardiner, unpublished, and International Human Chr21 Annotation Consortium, manuscript in preparation). Because human chromosome 21 is associated with significant genetic disorders such as Down syndrome, studying its genetic content can lead to a development of therapeutics for preventing and ameliorating these abnormalities. Several chromosome 21-related databases have been developed to aid chromosome 21-related

research. The Human Chromosome 21 cSNP and Map database stores the single nucleotide polymorphisms within cDNA sequences that map to chromosome 21. These data were generated by applying both bioinformatics and experimental approaches to the complete DNA sequence and Expressed Sequence Tags (ESTs) from public databases. The JST-ALIS database contains map and sequencing data for human chromosome 21. HUGO chromosome 21 includes links to other chromosome 21 resources. CroW21 is a user-friendly program for accessing existing specific chromosome 21 databases. However, none of these databases or programs is all inclusive. Users may have to search multiple databases to acquire all the information for their research and/or may be unaware that additional information exists. As a result, a comprehensive database is necessary.

In addition to comprehensive content, ease of use for a broad range of investigator expertise is an indispensable feature. Nowadays, researchers who lack a strong biological background as well as biologists who lack a strong statistical or evolutionary studies background increasingly become key users of the same biological databases. These users include computer scientists, statisticians, and bioinformaticians. Highly specialized user interfaces make it a challenge for many types of users to obtain all desirable information.

In this article, we present a new, comprehensive, user-friendly chromosome 21 database driven by a built-in protein interaction prediction tool based on Markov random field (MRF) and GeneQuest, a novel easy-to-use user interface. The database contains a wide range of data including both in-house generated data (i.e., chromosome 21 orthologs) and data collected from other existing chromosome 21 databases (i.e., protein interactions, pathways, etc.). This article will describe the underlying overall database design and data preparation, explain the underlying process of our built-in protein–protein interaction prediction tool based on MRF including experimental results on our data, present details of the GeneQuest facilities, and illustrate the results and typical uses of the database.

Data

Data Sources

The database contains chromosome 21-related data collected from a number of reliable sources: 1) chromosome 21 genes in

Researchers who lack a strong biological background as well as biologists who lack a strong statistical or evolutionary studies background increasingly become key users of the same biological databases.

ENTREZ, and the recent H-Inv database [3]; 2) sequence data from NCBI RefSeq, UniGene, dbEST, and SwissProt; 3) RNAi data from Flybase [4], Wormbase, and the yeast database; 4) proteins interaction data for *Drosophila*, *C. elegans* and *S. cerevisiae* from “The Grid”; 5) mammalian protein interaction data from the Database of Interacting Proteins (DIP); and, finally, 6) data on post-translational modifications, interactions, substrates, and targets from PubMed and with assistance of Dr. Akhilesh Pandey from Human Protein Reference Database (HPRD). Additionally, the database includes gene structure and alternative splicing information on genes from chromosome 21 and the orthologous regions on mouse chromosomes 16, 17, and 10, derived from direct annotation of genomic sequences. Mouse data are an important inclusion because this is the major model organism in the study of human diseases. The Web addresses of the above-mentioned databases are provided in a supplementary table which can be viewed at <http://chr21db.cudenver.edu/EMB>.

The Database

The database was designed using a relational model and implemented in an open-source MySQL database. A combination of HTML/Perl/Java languages was used to extract source information. The high-level model in the form of the entity-relationship diagram is shown in Figure 1. *Genes* contains information on human chromosome 21 genes and their orthologous genes in mouse such as: gene id, name, alias, description, strand, start and stop location, contig, etc. *Nas* includes nucleotide information for each gene; i.e., accession number, nucleotide length, coding start location, coding stop location, nucleic acid type, etc. *Proteins* contains protein features such as swiss protein id, GenBank protein id, accession number, protein sequence, etc. Ortholog information such as organism accession number, identity, similarity, etc., is included in *Orthologs*. *Domain* and *Interactions* contain information of protein domains and protein-protein interactions, respectively. More detailed description is provided in a supplementary table at <http://chr21db.cudenver.edu/EMB>. To keep the database current, we write java scripts to automatically extract data from external sources and feed data into the database monthly. In addition, to enhance the accuracy and completion of the information, the database will be manually curated by experts in Down syndrome and chromosome 21 gene functions. This is an important and distinguishing feature of this database.

Ortholog Data

Orthologs are genes in different species that are descended from a common ancestral gene and diverged in sequence during the speciation process. They are assumed to have retained

the functions of the ancestral gene. To obtain orthologs for human chromosome 21 genes in different organisms, a combination of BLAST (Basic Local Alignment Search Tool, a speed algorithm for comparing biological sequences and identifying similar sequence [5]) searches were conducted using human chromosome 21 protein sequences against the protein databases of *P. troglodytes*, *R. norvegicus*, *C. familiaris*, *G. gallus*, *D. rerio*, *T. rubripes*, *D. melanogaster*, *C. elegans*, *S. cerevisiae*, and *S. pombe*. A high-level diagram describing the ortholog-data acquisition process is shown in the Figure 2.

Methods

GeneQuest Method

We developed a new method, called GeneQuest, that enables users to exploit the database in an easy to use and efficient manner. The method was designed to hide the complexity of the database from the end-users. GeneQuest, a meta-data method, can work with any relational database, not just with the chromosome 21 database. In GeneQuest, the data are divided into groups of information from the chromosome 21 database. Each group has a relation with another group in the database. Using GeneQuest one can walk through an ontology tree and select items of interest for extraction. GeneQuest allows users to collect the information most relevant to their needs. The advantage of GeneQuest is that both novices and experts can use it to generate simple and complex queries (Figure 3). The following points describe some of the details of GeneQuest:

- **Groups.** Each group may contain one or many tables in the database. Currently, there are nine groups: Gene information, Ortholog gene, Nucleotide, Nucleotide features, Proteins, Ortholog proteins, Interaction proteins, Domains, and UniGene profiles.
- **Items.** Each item is an attribute of a table. An item has a name, caption, type, group which the item belongs to, table, etc.

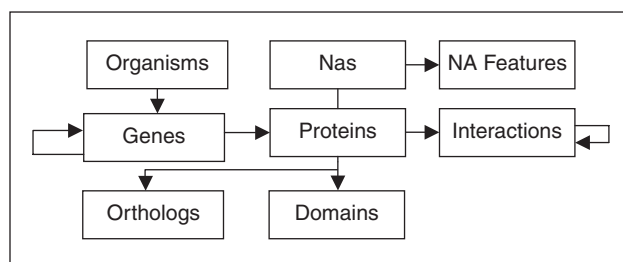


Fig. 1. Human Chromosome 21 database entity relationship diagram consists of eight tables. The Genes table has a relation with itself as each chromosome 21 gene may have ortholog genes in other organisms. The Interactions table also has a relation with itself.

Protein-protein interaction networks can be a powerful tool for generating novel functional information and generating new hypotheses regarding chromosome 21 proteins.

- **Relations.** Relations define relations between groups. For example, when users come to the Gene information group, they see only groups that have a relation with the Gene information group such as Ortholog gene, Nucleotide acid, Proteins, and so on.

After selecting items, the users can filter extracted data (Figure 3). With the aid of an AND/OR tree they can easily set up simple and complex conditions. In conditions, the users can set parameters to answer a general question. For example, to parameterize the condition involving organisms, the user can set a condition such as OrganismName=?[Organism Name], and whenever the question is executed, GeneQuest asks the user to enter the value for the organism. The method is original and supports a template concept; i.e., the user can save the question into a template for later use. The pseudo code for construction of the ontology tree is summarized as follows:

Algorithm 1: Construction of the Ontology Tree

Read data set from items, groups and relations.

For each *group-item* in data set

 Add *group-item* to the tree

 If the group changed

 set up relation with previous group

 End If

End For

Example 1. Given a specific gene, find the orthologous protein in a specific organism. The extracted information is: *gene id, protein accession number, ortholog gene id, ortholog protein accession number, identity %, similarity %, expect value and the blast alignment picture.*

The users should go to *Genes information* to select information of gene (*gene id*) then continue to *Proteins* to get protein

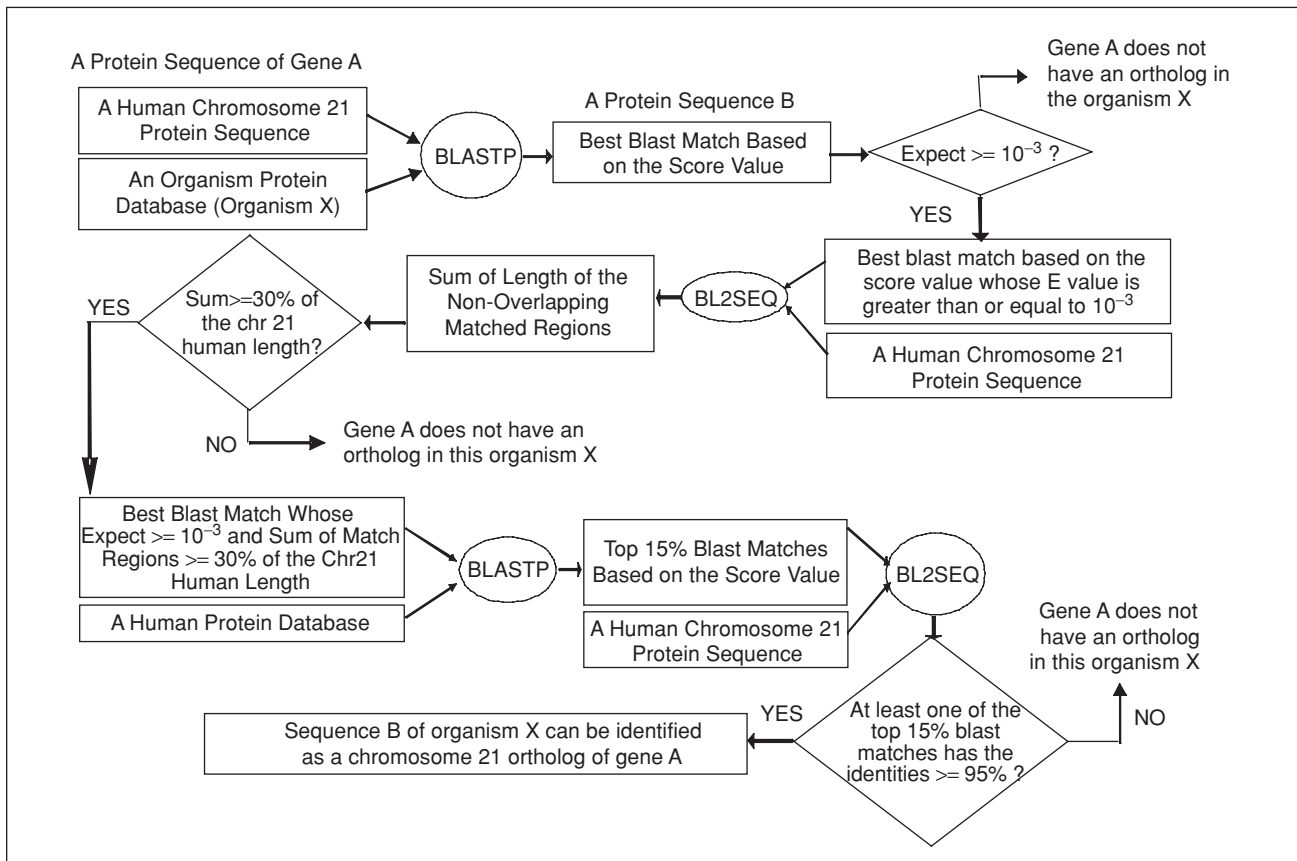


Fig. 2. A high-level diagram of ortholog data acquisition process: 1) use BLASTP for each chromosome 21 protein sequence against an organism protein database, 2) select the best blast match sequences using BLASTP (reciprocal blast) against human protein database, and 3) use BL2SEQ between the best BLAST match sequences in step 2 against the chromosome 21 sequence to perform the cutoff and ortholog identification.

accession number, and stop at *Ortholog proteins* to select the rest of extracted information.

In the *Filter* section, the user sets under the *AND* condition:

Gene name =?[Gene Name]
 chromosome=21,
 Ortholog description like %?
 [Organism]%.

The result for this query is answered by GeneQuest; for instance, if the Gene name is input as *ITSN1* and Organism is input as *drosophila*, the output is shown in Table 1.

Example 2. We illustrate how to extract a UniGene expression profile for a specified gene in a specific tissue. To create such a question, the user needs to go to *Genes information* to select *Gene id* and *Gene name*, then stop at UniGene profile to obtain *Sequences per million* and *Clones sequenced/ Total sequences*.

In the *Filter* section, the user sets under the *AND* condition:
 Gene name =?[Gene Name]
 Breakdown value like %?[Tissue]%.

Table 2 shows the answer generated by GeneQuest when Gene name is input as *SYNJ1* and Tissue is entered as brain.

Protein Function Prediction Tool

Our goal is not only to have the chromosome 21 database as comprehensive as possible but also user friendly. To provide comprehensive content we continue to add new functionalities to the database. One of these additions is a program for protein–protein interaction, based on the MRF method [6], which is described below.

Methods for Protein Function Prediction

The classical way of approaching the problem of the protein function prediction is to find similarities between the protein of interest and other proteins using programs such as BLAST, and then to assign predictions of function to the query based on known functions of the matches [5]. Another approach is

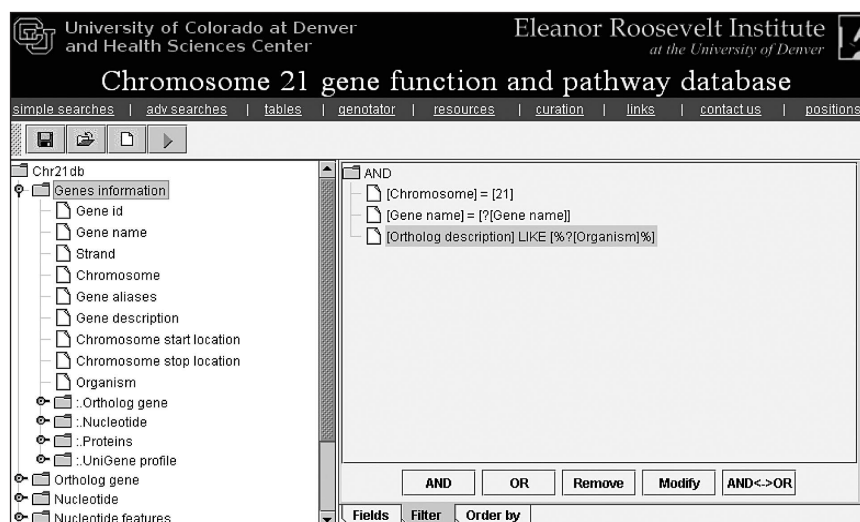


Fig. 3. GeneQuest interface: the left panel shows the ontology tree and the right panel shows the conditional (AND/OR) tree.

based on a heuristic combination of different types of data. Troyanskaya et al. [7] used Bayesian reasoning to integrate heterogeneous types of high-throughput biological data (e.g., large-scale two-hybrid screens and multiple microarray analyses) for gene function prediction.

The approach we are interested in, which we describe below, is based on protein–protein interaction networks. The protein–protein interaction network describes a neighborhood structure among the proteins. If two proteins interact, they are neighbors of each other. For a nonannotated protein, the functions of its neighbors inform about its function. For a given function, if most of the neighbors of a protein have that function, we assume that the protein may have the same function as well. Schwikowski et al. [8] proposed a method to infer the functions of a nonannotated protein based on the frequencies of its neighbors having certain functions. They assign k functions to the nonannotated protein with the k largest frequencies among its neighbors. The approach of Hishigaki et al. [9] inferred protein functions by using the chi-square χ^2 -statistics. For a protein P_i , let $n_i(j)$ be the number of proteins interacting with P_i and having function F_j . Let $e_i(j) = (\text{Number of neighbors of protein } P_i) \times (\pi_j)$, the expected number of proteins in neighborhood of P_i having function F_j . For a fixed k , they assigned a

Table 1. Orthologous protein of gene *ITSN1* in *drosophila*.

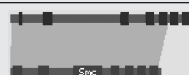
Gene ID	Protein Accession no	Ortholog		Identity	Similarity	Ortholog Picture	Reciprocal Blast E value
		Gene ID	Accession No.				
6453	NP_001001132	35378	AAC39139	41	58		7e-72

Table 2. UniGene expression profile for *SYNJ1* in brain.

Gene ID	Gene Name	UniGene ID	Sequences per Million	Clones Sequenced/Total Sequences
8867	SYNJ1	HS473632	70	33/469643

To aid in hypothesis generation, a mouse click “humanizes” the interaction network, displaying gene name and ID for the best human matches for each model organism protein in the network.

nonannotated protein with k functions having the top k value $(n_i(j) - e_i(j))^2 / e_i(j)$. However, it seems logical that proteins far away from P_i contribute less information than the close neighbors. Thus, we should assign less weight to proteins far away from protein P_i than to those who are close neighbors.

Markov Random Field Method

In our chromosome 21 database, predictions of protein functions based on the functions of their interacting proteins were derived using a combination of the MRF of Deng et al. [6] and Rosetta stone translation. Suppose we have N proteins and $X = (X_1, X_2, \dots, X_N)$ is the functional annotation of a function F_j . Let X_1, \dots, X_k be the set of nonannotated proteins and X_{k+1}, \dots, X_N be the set of annotated proteins (with function F_j). We know that protein $P_{i,i=1..N}$ interacts with $P_{j,j=1..N}$, and that P_j interacts with $P_{k,k=1..N}$, and so on. Such an arrangement can be depicted as a protein-protein interaction network. Once we have defined the network, we need to find a posteriori distribution of (X_1, \dots, X_k) . We do this using a Bayesian belief network with the following quasi-likelihood function:

$$P(X_1, \dots, X_k) = \prod_i P(x_i = 1 | x_{[-i]}, \theta) = \prod_i \frac{e^{\alpha + (\beta - 1)M_0^i + (\gamma - \beta)M_1^i}}{1 + e^{\alpha + (\beta - 1)M_0^i + (\gamma - \beta)M_1^i}} \quad (1)$$

where $X_{[-i]} = (X_1, X_2, \dots, X_{i-1}, X_{i+1}, \dots, X_N)$, $M_0^{(i)}$ = cardinality of $\{j \in \text{Neighbor set of } i, X_j = 0\}$, $M_1^{(i)}$ = cardinality of $\{j \in \text{Neighbor set of } i, X_j = 1\}$, and $\theta = (\alpha, \beta, \gamma)$ are

parameters in the equation.

To maximize this quasi-likelihood function we need to solve the nonlinear equation system:

$$\text{Ln}(P) = \sum_i \left\{ \alpha + (\beta - 1)M_0^i + (\gamma - \beta)M_1^i - \ln \left(1 + e^{\alpha + (\beta - 1)M_0^i + (\gamma - \beta)M_1^i} \right) \right\} \rightarrow \max \quad (2)$$

$$\text{when } \frac{\partial \text{Ln}(P)}{\partial \beta} = 0, \quad \text{and, } \frac{\partial \text{Ln}(P)}{\partial \gamma} = 0$$

Below we provide pseudo-codes of the two algorithms. The first algorithm, the MRF method, predicts the probability that proteins have a function of interest; the second algorithm, the leave-one-out cross-validation method, evaluates the accuracy of our approach.

Algorithm 2: MRF

Step1: For a function in a list of functions of interest, estimate the probability, π , that a protein has the function: $\pi = \text{\#number of proteins with the function} / \text{\#total number of protein}$ $s = (N - k) / N$, set $\alpha = \ln(\frac{\pi}{1-\pi})$

Step2: Solve the nonlinear equation system in (2) to estimate the parameters β and γ .

Step3:

- 1) Randomly set the value of nonannotated proteins $X_i = 1$ ($i = 1..k$) with probability π .

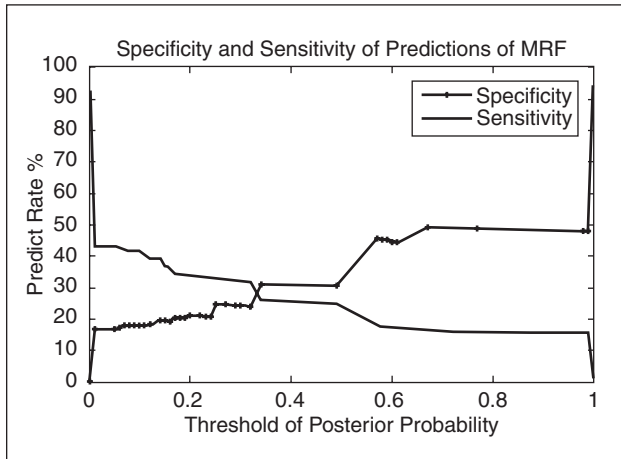


Fig. 4. Specificity and sensitivity of the MRF method for different thresholds on the raw data set.

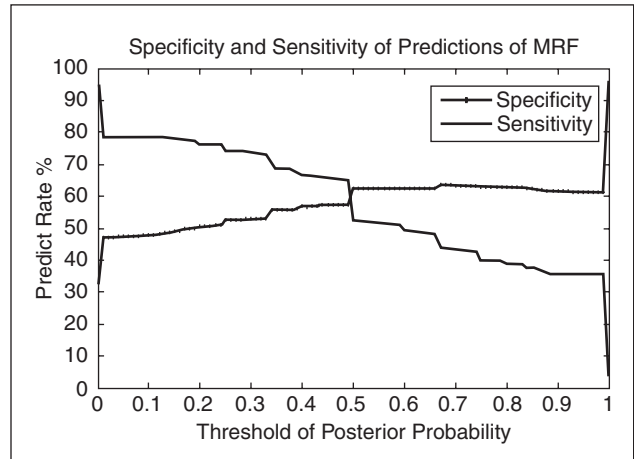


Fig. 5. Specificity and sensitivity of the MRF method for different thresholds on the refined data set.

Protein-protein interaction networks can be a powerful tool for generating novel functional information and generating new hypotheses regarding chromosome 21 proteins.

- 2) For each protein P_i , update the value X_i using the quasi-likelihood function.
- 3) Repeat step 2 n times until all the posterior probabilities $P(X_i|X_{[-i]})$ are stabilized.

For a protein that belongs to an interaction network, we directly apply the MRF method to predict the function of the protein. For a protein that does not belong to any interaction network, we first apply a Rosetta stone translation: we search for orthologous proteins in organisms such as *D. melanogaster*, *C. elegans*, and *S. cerevisiae*. Then we humanize the ortholog's protein-protein interaction network by using BLASTP to identify the best human matches for each model organism protein in the network, and, after choosing cutoff values (i.e., percent similarity and expectation value) to the humanized network, we apply the MRF method to predict functions for the protein.

Experimental Results

We applied the MRF method to predict functions of nonannotated proteins in the recently published human protein-protein interaction network [10]. Functions of those proteins are extracted from Gene Ontology (GO) at the NCBI database (see supplementary table at <http://chr21db.cudenver.edu/EMB>). We tested this method on two data sets. The first is the raw data that include 6,726 interaction pairs involving 3,134 proteins and 2,218 GO functions. The second is a refined data set in which we selected only the top five, highest frequency, functions of proteins. Those functions and their frequencies are, respectively, GO:0005634:711, GO:0005515:636, GO:0000166:365, GO:0046872:323, GO:0006355:302. The refined data include 2,615 interaction pairs involving 1,304 proteins and 5 GO functions. For both data sets we calculated the accuracy of the predictions by using the leave-one-out cross-validation method, which is summarized as follows:

Algorithm 3: Leave-One-Out Cross-Validation Method with a Threshold

Loop

Randomly select an annotated protein P_i and assume it is nonannotated. Let n_i be the number of functions of P_i
 For each function F_i in the function list :
 Apply Algorithm 1 to predict F_i for P_i : if the probability of the function F_i is above a *threshold* (range from 0..1) then assign the function F_i for P_i . For each *threshold* in range: let m_i be the predicted functions for protein P_i , and k_i be the overlap between the set m_i predicted functions and set n_i observed functions.

End For

End Loop (K times)

For each *threshold* we calculate the specificity (SP) and the sensitivity (SN) as follows:

$$\text{Specificity } SP = \frac{\sum_i^K k_i}{\sum_i^K m_i} \quad \text{Sensitivity } SP = \frac{\sum_i^K k_i}{\sum_i^K n_i}.$$

Figure 4 and Figure 5 show the relationship between specificity and sensitivity for the two data sets, respectively. The accuracy is slightly worse on the raw data set. With the threshold of 0.37 we obtain an accuracy of approximately 30%.

The accuracy is better on the refined data set. We achieved accuracy of 60.2% with the threshold of 0.49. The results can be considered as valid predictions, but they remain predictions only and require inspection for biological reasonableness before they are used. Deng et al. [6] applied their method to infer the functions of nonannotated proteins in yeast. They considered only 43 functional categories based on cellular role involving 1,877 proteins and 2,442 interaction pairs. With the threshold equal to 0.17, they obtained the accuracy of 47%, which is an improvement over both the Schwikowski et al. [8] and the Hishigaki et al. [9] methods (Figure 6).

Results

In this section we show the results of using the GeneQuest program for searching the chromosome 21 database, as well as results of the protein function prediction program.

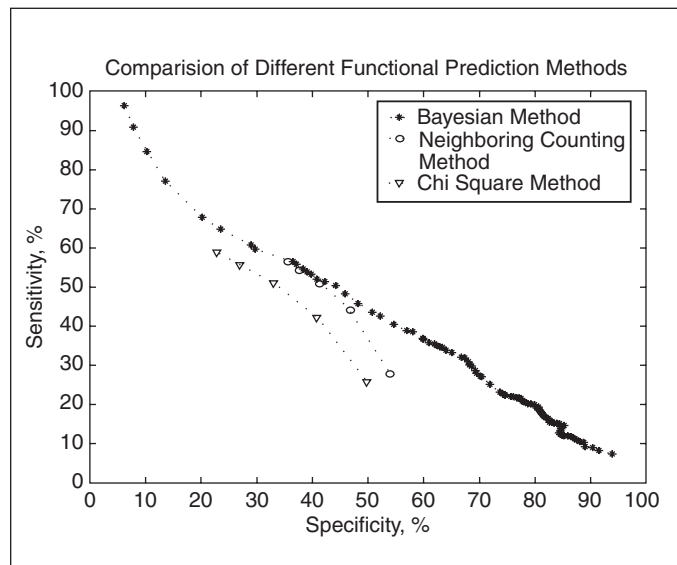


Fig. 6. Specificity and sensitivity of prediction for the three methods (taken from (6)).

GeneQuest enables users to utilize the database in a very efficient manner. Our comprehensive database contains data collected from several reliable sources and those generated in-house, including sequence data, RNAi data, post-translational modification data, and ortholog data.

Simple Database Search

A database search by gene name (or alternatively by alias, description, or accession number) immediately produces schematics and accession numbers of major mRNA splice variants, the domain compositions of encoded proteins, references to literature reports in PubMed for gene identification and functional analysis, plus a list of similar proteins in ten major model organisms, again with domain correlations (see supplementary figure at <http://chr21db.cudenver.edu/EMB>). From this gene information, a single click moves us to the adjacent gene on chromosome 21 or to the orthologous gene, with similar information, in the mouse genome. Users must be aware that knowledge limitations may exist; these can include incomplete gene structures, in particular at the 5' end and incomplete data on alternative splice variants. Biologists in particular must be aware that, in some genes, some exons overlap with repetitive sequences, and that there are a large number of gene models based on spliced ESTs that lack sequence conservation in the mouse genome and/or that lack obvious open reading frames. While computational methods may be valid, biological interpretations are not uniformly assured.

Advanced Search

Options are also provided for more complex queries. The user can obtain lists of genes with predefined characteristics; for example, all chromosome 21 proteins with similar proteins in any of the ten model organisms, with each entry described by accession numbers, percent and length of identity and domain composition. Alternatively, the user can use GeneQuest to formulate novel queries and to retrieve lists of genes with specified characteristics; for example, proteins with specific functional domains and a user defined set of associated data, proteins of chromosome 21 conserved in *Drosophila* and encoding SH3 domains with expected value $>10^{-5}$, all proteins interacting with a specified gene, etc.

Protein Interaction Data

Protein-protein interaction networks can be a powerful tool for generating novel functional information and generating new hypotheses regarding chromosome 21 proteins. Protein interaction data from mammals (human, rat, and mouse) have been consolidated from multiple databases and, from a query on gene name, are provided either as a list or in graphical form. Complete interactomes are being developed for *Drosophila*, *C. elegans*, and yeast [8]–[10]. For each organism, the user can access a list of all chromosome 21 orthologs for which interaction data are available. Graphical representations for primary and secondary interactions, of high and low confidence, are provided for each protein, with accession number, gene name, and gene ID (see supplementary figure at <http://chr21db.cudenver.edu/EMB>). To aid in hypothesis generation, a mouse

click “humanizes” the interaction network, displaying gene name and ID for the best human matches for each model organism protein in the network. The user can choose to display only those matches (default) with E values $< e^{-03}$ or less, or those with a user-defined more stringent cutoff. The challenge for automated data collection is to define biologically significant matches and interactions. While interactions cannot be assumed to be conserved among organisms, the information may provide new insights that can be tested via experiments with human genes. The user needs to use his or her own biological intuition; however, and is expected to perform additional BLAST analyses before designing further experiments. To predict functions of a novel protein, the user will select a specified function category and then perform predictions. The user can also set up a threshold for the posterior probability (see supplementary figure at <http://chr21db.cudenver.edu/EMB>). The results are purely predictive so their validity needs to be confirmed by additional experiments such as demonstrating protein function (e.g., kinases phosphatases), identifying cofactors, a role in signal transduction, or a phenotype of a mutation.

Conclusions and Future Work

The existing chromosome 21 databases are far from being comprehensive and are very often difficult to use for both biologists and nonbiologists. This article introduced a comprehensive, user-friendly chromosome 21 database driven by a built-in protein interaction prediction tool based on MRF and an easy-to-use user interface program called GeneQuest. The protein interaction prediction tool allows users to predict protein functions for nonannotated genes. GeneQuest enables users to utilize the database in a very efficient manner. Our comprehensive database contains data collected from several reliable sources and those generated in-house, including sequence data, RNAi data, post-translational modification data, and ortholog data. Additions that we are planning include protein functional consequences of human polymorphisms and of differences between humans and chimpanzee; schematics of pathways that directly, or indirectly, involve or are impacted by chromosome 21 genes; and tools for interpreting microarray data from a Down syndrome-relevant perspective. This latter will include analysis of existing microarray data to identify correlations of expression pattern changes among sets of chromosome 21 genes and among chromosome 21 genes and nonchromosome 21 genes. Additional new data will derive from novel proteomics approaches, e.g., phosphoproteomics, directed at comparing normal tissues with those deriving from trisomic mouse models of Down syndrome. Identifying perturbations in trisomy in systems at the RNA or the protein level will contribute to pathway information. The long term goal is prediction of novel pathway associations and gene networks relevant to chromosome 21 genes. The

database and GeneQuest are easily expandable for considering not only human and mouse data but also other organisms such as worm, yeast, *E. coli*, fly, rat, chimpanzee, etc.



Cao Nguyen is a Ph.D. student in Computer Science and Information Systems program, with the option in Computational Biology, at the University of Colorado at Denver and Health Sciences Center (UCDHSC). He conducts research with Dr. Cios in the area of mathematical modeling (hidden Markov models, clustering, and fuzzy cognitive maps) for prediction of protein-protein interfaces and protein functions. Nguyen holds an M.S. degree in Computer Science from the Vietnam National University. He has been awarded full scholarship for his study in the United States.



Supphachai Thaicharoen received the B.S. degree in electrical engineering from King Mongkut's Institute of Technology North Bangkok, Thailand. He holds two M.S. degrees, one in computer information systems and the other in computer science, both from Colorado State University. Currently he is a Ph.D. candidate in the College of Engineering and Applied Science at the UCDHSC. His research interests include machine learning, bioinformatics, and text mining.



Thomas Lacroix graduated with an M.S. degree in bioinformatics and genomics at the University of Paris VII and Evry, Paris, France. He currently works as bioinformatician and molecular biologist at the Eleanor Roosevelt Institute at the University of Denver under supervision of Dr. Gardiner. His main professional experiences include six months at the Applied Biosystems/Celera Genomics (San Francisco) in the department of protein informatics, six months at the Roslin Institute (Edinburgh) in the department of gene expression and development, three months as a laboratory technician for the J. Minjoz Hospital (Besançon), and three months at the University of Besançon in the Biochemistry-Molecular Biology Laboratory.

Katheleen J. Gardiner received the Ph.D. degree from the University of Colorado and is currently a Professor at the Eleanor Roosevelt Institute at the University of Denver and an Adjoint Associate Professor in the department of biochemistry and molecular genetics at the UCDHSC. Dr. Gardiner is an internationally recognized researcher in the field of Down syndrome. Her specific research interests are in the identification of genes encoded by human chromosome 21 that contribute to learning and memory deficits in Down syndrome and the use of data from mouse and other model organisms to predict associated pathway perturbations. Dr. Gardiner has authored over 100 journal articles, meeting reports, and book chapters. Her research is currently funded by the National Institutes of Health and the Fondation Jerome Lejeune.



Krzysztof J. Cios received the M.S. and Ph.D. degrees from the AGH University of Science and Technology, Krakow, the MBA degree from the University of Toledo, Ohio, and the D.Sc. degree from the Polish Academy of Sciences. He is currently a Professor at the University of Colorado at Denver and Health Sciences Center, and Associate Director of the University of Colorado Bioenergetics Institute. He directs Data Mining and Bioinformatics Laboratory. Dr. Cios is a well-known researcher in the areas of learning algorithms, biomedical informatics, and data mining. NASA, NSF, American Heart Association, Ohio Aerospace Institute, NATO, US Air Force, and NIH have funded his research. He has published three books, about 150 journal and conference articles, and 12 book chapters; serves on editorial boards of *Neurocomputing*, *IEEE Engineering in Medicine and Biology Magazine*, *International Journal of Computational Intelligence*, and *Biodata Mining*; and has edited five special issues of journals. Dr. Cios has been the recipient of the Norbert Wiener Outstanding Paper Award, the Neurocomputing Best Paper Award, the University of Toledo Outstanding Faculty Research Award, and the Fulbright Senior Scholar Award. Dr. Cios is a Foreign Member of the Polish Academy of Arts and Sciences.

Address for Correspondence: Krzysztof J. Cios, Department of Computer Science and Engineering, University of Colorado at Denver and Health Sciences Center, P.O. Box 173364, Campus Box 109, Denver, CO 80217-3364 USA. E-mail: krys.cios@cudenver.edu.

References

- [1] M. Hattori, A. Fujiyama, T.D. Taylor, H. Watanabe, et al, (Chromosome 21 Mapping and Sequencing Consortium), "The DNA sequence of human chromosome 21," *Nature*, vol. 405, pp. 311–319, 2000.
- [2] K. Gardiner, A. Fortna, L. Bechtel, and M.T. Davison, "Mouse models of Down syndrome: How useful can they be? Comparison of the gene content of human chromosome 21 with orthologous mouse genomic regions," *Gene*, vol. 318, pp. 137–147, 2003.
- [3] T. Imanishi, T. Itoh, Y. Suzuki, C. O'Donovan, et al., "Integrative annotation of 21,037 human genes validated by full-length cDNA clones," *PLoS Biol.*, vol. 2, p.856–875, 2004.
- [4] R.A. Drysdale, M.A. Crosby, and the FlyBase Consortium, "FlyBase: Genes and gene models," *Nucleic Acids Res.*, vol. 33, pp. 390–395, 2005.
- [5] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman, "Gapped BLAST and PSI-BLAST: A new generation of protein database search programs," *Nucleic Acids Res.*, vol. 25, pp. 3389–3402, 1997.
- [6] M. Deng, K. Zhang, S. Mehta, T. Chen, and F. Sun, "Prediction of protein function using protein-protein interaction data," *J. Comput. Biol.*, vol. 10, pp. 947–960, 2003.
- [7] O. Troyanskaya, K. Dolinski, A. Owen, R. Altman, and D. Botstein, "A Bayesian framework for combining heterogeneous data sources for gene function prediction (in *Saccharomyces cerevisiae*)," in *Proc. Nat. Acad. Sci. USA*, vol. 100, pp. 8348–8353, 2003.
- [8] B. Schwikowski, P. Uetz, and S. Fields, "A network of protein-protein interactions in yeast," *Nature Biotechnol.*, vol. 18, pp. 1257–1261, 2000.
- [9] H. Hishigaki, K. Nakai, T. Ono, A. Tanigami, and T. Takagi, "Assessment of prediction accuracy of protein function from protein-protein interaction data," *Yeast*, vol. 18, pp. 523–531, 2001.
- [10] J.F. Rual, K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G.F. Berriz, F.D. Gibbons, M. Dreze, N. Ayivi-Guedehoussou, N. Klitgord, C. Simon, M. Boxem, S. Milstein, J. Rosenberg, D.S. Goldberg, L.V. Zhang, S.L. Wong, G. Franklin, S. Li, J.S. Albala, J. Lim, C. Fraughton, E. Llamas, S. Cevik, C. Bex, P. Lamesch, R.S. Sikorski, J. Vandenhaute, H.Y. Zoghbi, A. Smolyar, S. Bosak, R. Sequerra, L. Doucette-Stamm, M.E. Cusick, D.E. Hill, F.P. Roth, and M. Vidal, "Towards a proteome-scale map of the human protein-protein interaction network," *Nature*, vol. 437, pp. 1173–1178, 2005.