Using Database Search Algorithm Design Database for Interleukins

N.Deepak Kumar, Dr.A.Ramamohan Reddy, Dr.AnjanBabu

Abstract— Advancing our understanding of mechanisms of immune regulation in allergy,asthma,autoimmune diseases,tumor development, organ transplantation, and chronic infection are the some of the diseases where Interleukins play an important role. The immune and inflammatory cells interact through interleukins and reciprocal regulation with counter balance among TH and regulatory T cells,as well as the subsets of B cells will offer opportunities for immune interventions.
A database manages information and allows organizing data, ensuring completeness and integrity, and transforming the data from one form to another. It make search through the data efficiently to find the desired information. In the present work a database for Interleukins (proteins) have been created as the data related to Interleukins is increasing day by day, it has become difficult for researchers to manage their structure, classification and functions. There are 37 Interleukins have been discovered by the researchers and many more may be added in the future. Interleukins study is mostly useful for diagnosing and treating all the diseases of a human body.

Index Terms— Interleukin,Protein,Immune cells,Bioinformatics,ER diagram,DBMS.

1 INTRODUCTION

Interleukins are biologically active glycoproteins derived primarily from activated lymphocytes and macrophages. Tremendous insight into the biochemical and biological properties of interleukins has been gained with advances in recombinant DNA technology, protein purification, and cell-culture techniques. The biological properties of interleukins include induction of T-lymphocyte activation and proliferation, augmentation of neutrophil, macrophage, and T-lymphocyte cytooxicity, and promotion of B lymphocyte and multilineage bone marrow stem-cell precursor growth and differentiation. Interleukins may play a role in the pathogenesis of several important diseases. Interleukin therapy is likely to play an important role in the treatment of cancer, infectious diseases, and immunodeficiency syndromes. [12,14].
Specified design processes are standard in the software development industry, and there are many design processes described in the software engineering literature. The details of the design process are less crucial than the use of a process. However, there are some crucial steps, such as gathering requirements. Requirements document what the database is trying to accomplish. Most databases have to make data model compromises. Databases have been used to manage and integrate large volumes of complex data in other disciplines for decades [1].
Development of a data model is another crucial step because this helps to identify potential problems in the design early on in the project, when they are still easy to correct. The most common tool used for this purpose in relational database design is the entity relationship diagram. This type of diagram represents the real-world entities about which the database will store information, and the relationships between those entities.
Use the database to enforce data integrity. A database should protect the integrity or consistency of the data that it stores. The strong theoretical basis of relational DBMS provides rules of normalization, which, if followed, will ensure basic data integrity. These rules ensure that all information is stored in the smallest meaningful pieces and is stored in only one place, preventing data duplication and the concomitant potential for internal inconsistencies. A database that obeys these rules is said to be normalized. Normalization splits related data across multiple tables, requiring queries to perform operations called joins to reassemble the data.
The recent bioinformatics literature includes numerous papers about databases, but these primarily focus on the need for integration across existing databases [2,3], report the design and use of specific databases [4–9], or argue for better large scientific databases and the systematic changes necessary to accomplish this goal [10,11]. All this information is valuable, but does not provide much help to novice database designers.

2 METHODOLOGY

The database design should be able to accommodate realistic test data early in the design process. It is important to include 'pathological examples' (i.e. example data that represent the most complex relationships in the future dataset). For instance, consider the design of a set of tables to store the relationship between genes and proteins. Without any knowledge of biology, one might erroneously assume this is a one-to-one relationship (each gene relates to one and only one protein, and vice versa). Realistic test data quickly reveal the flaw in this design: because of splicing variations, one gene can produce multiple proteins, resulting in a one-to-many

relationship. Further consideration reveals the pathological case: owing to genetic redundancy, one protein can also be produced by multiple genes. The gene–protein relationship is therefore most correctly modeled as a many-to-many relationship.

Some complicated relationships might be unimportant for the goal of a database and can be simplified. For in stance, a database concerned primarily with protein function might need to identify only one gene for a protein. However, any simplifying assumptions made in the data model should always be documented. Also, it must be certain that the more complicated relationship is truly outside of the scope of the database, and not merely absent from the requirements of the current application accessing the database (see 'Keep the database manageable' above). The magnitude of data is also an issue in biological databases. A good design for a database with one hundred rows can be a disaster for a database with one million rows. The design process must take the volume of data into account, particularly during the physical design of the database.

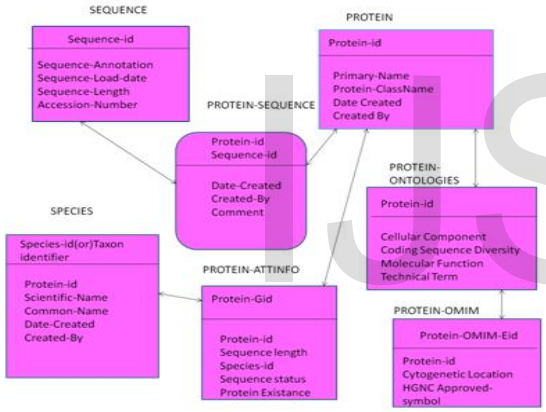

Figure I. ER Diagram for storing information about Proteins

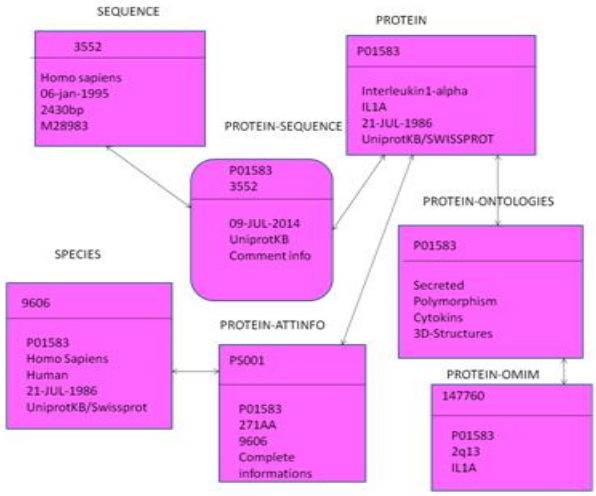

Figure II. Example ER Diagram for storing information about Interleukins. An hypothetical data was shown for Interleukin database.

2.1 . Organization of Database

The organization of membrane protein function database is illustrated in Fig. 1. Each entry in the database contains the following information:
(i) SEQUEENCE, (ii)PROTEIN (iii) PROTEIN SEQUENCE(iv)SPECIES (v) PROTEIN ONTOLOGIES (vi) PROTEIN ATTRIBUTE INFORMATIONS ,(vii)PROTEIN MEDICAL INFORMATIONS. We have provided the sequence and structure information in t he form of Uniprot and NCBI,OMIM,HGNC.

Depending on the type of stored information, the most importan databases in biotechnology are in the following:

(i)SEQUENCE:
a)Sequence-id
b)Protein-id
c)Sequence-Annotation
d)Sequence load date
e)Sequence-length
f)Accession Number

(ii)PROTEIN:
a)Protein-id
b)Primary-Name
c)Protein-Class Name
d)Date Created
e)Created By

(iii)PROTEIN SEQUENCE:
a)Protein-id
b)Sequence-id
c)Date Created
d)Created By
e)Comment

(iv)SPECIES:
a)Protein-id
b)Species-id (or)Taxon identifier
c)Scientific Name
d)Common Name
e)Date created
f)Created By

(v)PROTEIN ONTOLOGIES:
a)Protein-id
b)Cellular Component
c)Coding sequence Diversity
d)Molecular Function
e)Technical term

(vi)PROTEIN ATTRIBUTE INFORMATIONS:
a)Protein id
b)Species id
c)Sequence length
d)Sequence status
e)Protein Existence

(vii)PROTEIN MEDICAL INFORMATIONS:
a)Protein-id
b)Cytogenetic Location

c)HGNC Approved Symbol

2.2.Database searching Algorithm

How to search the Interleukin protein informations in the large databases like NCBI,UNIPROT,HGNC,OMIM.That point of view using below search algorithms.These algorithms are very very useful to collect all protein informations in the large databases.Because lakh of protein informations are available in the different databases,that regards using the below searching algorithms.i.e namely

HUNO Algorithm:

Huno Algorithm is the one of the easy way to access the protein information from HGNC,Uniprot,NCBI,OMIM.The Huno algorithm follow the below steps:

Stept1: Collect all Sequence,Species Entity informations from NCBI through Descriptive model.Descriptive algorithms are based on a mechanistic prediction of how peptides fragment in a tandem mass spectrometer, which is then quantified to determine the quality of the match between the prediction and the experimental spectrum. Mathematical methods such as correlation analysis have been used to assess match quality.SEQUEST is an example of a program that uses a descriptive model for peptide fragmentation and correlative matching to a tandem mass spectrum. It uses a two-tiered scoring scheme to assess the quality of the match between the spectrum and amino acid sequence from a database. The first score calculated, the preliminary score (Sp), The original Sp score is:

$$Sp=(\Sigma_k I_k)m(1+\beta)(1+p)/L$$

where the first term in the product is the sum of ion abundances of all matched peaks, m is the number of matches, β is a 'reward' for each consecutive match of an ion series is a 'reward' for the presence of an immonium ion and L is the number of all theoretical ions of an amino acid sequence.based on this algorithm we will get The sequence Entity attributes.i.e a)Sequence-id ,b)Protein-id,c)Sequence-Annotation,d)Sequence load date,e)Sequence-length,f)Accession Number and Species Entity attributes i.e. a)Protein-id,b)Species-id (or)Taxon identifier,c)Scientific Name,d)Common Name,e)Date created,f)Created By through Sequence ID.

Step2: Collect all Protein Entity informations from HGNC through Statistical and Probability model.Statistical and probability models determine the relationship between the tandem mass spectrum and sequences. The probability of peptide identification and its significance are then derived from the model.Recently, a group of database search algorithms have been implemented that use collective properties of database sequences to calculate the probability that a sequence match is a random event. Thus, we have proposed to divide all database fragment ions into two groups: matches and misses. Then, we assume that a hypergeometric probability models the frequencies of database peptides based on the number of matches. According to this model a probability that a peptide match is a random event is predicted from the hypergeometric probability of choosing $K1$ matches (number of matches of a peptide) in $N1$ trials (the number of fragment ions of the peptide) from a pool of fragments consisting of N fragments (number of all database fragments) K of which are matches shown in given graph clearly (number of matches of all fragment ions to a spectrum). The hypergeometric probability of this event is:

$$P_{K,N}(K1,N1)= C_k^{K1} * C_{N-K}^{N1-K1} / C_N^{N1}$$

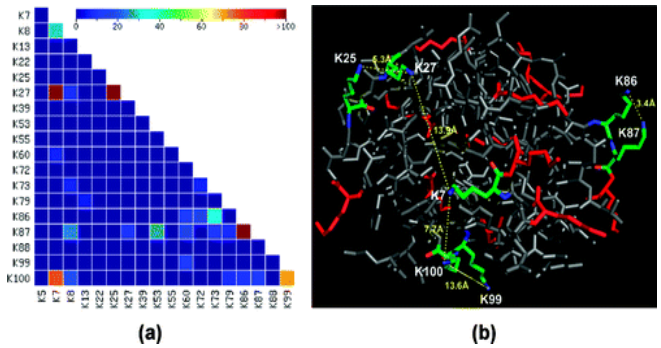

Fig II.Graph for Database search

Based on this algorithm we will get the Protein Entity attributes i.e. a)Protein-id,b)Primary-Name,c)Protein-Class Name,d)Date Created,e)Created By through Protein Class Name.

Step3: Collect all Protein Sequence,Protein Ontologies,Protein Attribute Informations Entity informations from Uniprot through Interpretative model.Interpretative approaches are based on manual or automated interpretation of a partial sequence from a tandem mass spectrum and incorporation of that sequence into a database search. Matches between the sequence and the spectrum have been scored using probabilities or correlation methods.A search engine has been fashioned using the partial sequence by dividing every candidate sequence into three parts:region 1 of unknown mass, region 2, containing the sequence tag and another region 3.Also, a probability is assigned to the amino acids at the cleavage sites. Complete tryptic cleavage of a protein results in peptides terminating in one of two amino acids—lysine or arginine.Therefore, if a sequence is tryptic, the random match probabilityis multiplied by 1/100, and if it is half tryptic by 1/10. Combiningthe probabilities of all regions and cleavage probabilities, a probability that a sequence match is random is set:

$$P_{random}=p_{NtermCleavage} * p_{m1} * p_{1sttagposition} * \ldots p_{lasttagposition} * p_{m3} * p_{Ctermcleavage}$$

The probability of a nonrandom match in a database with N amino acids would then be

$$p_{nonrandom} = (1-2 * p_{random})^N$$

In the above formula the random match probability is multiplied by 2 to account for the fact that the direction of the partial sequence is not known. As it is seen from the formula for a nonrandom match, the identification is dependent on the size of the database. In general, the larger the database, the longer the sequence tag should be for higher confidence matches.Based on this algorithm we will get the Protein Sequence,Protein Ontologies Entity attributes i.e a)Protein-id,b)Sequence-id,c)DateCreated,d)Created y,e)Comment Protein Ontologies Entity attributes i.e.. a)Protein-id,b)Cellular Component,c)Coding sequence Diversity,d)Molecular Function,e)Technical term and Protein Attribute Informations Entity attributes I,e. a)Protein id,b)Species id,c)Sequence length,d)Sequence status,e)Protein Existence through **ProteinId.**

Step4: Collect all Protein Medical Informations Entity informations from OMIM through Stochastic models.Stochastic models are based on probability models for the generation of tandem mass spectra and the fragmentation of peptides. Basic probabilities of fragment ion matches are obtained from training sets of spectra of known sequence identity. Stochastic models use statistical limits on the measurement and fragmentation process to create a likelihood that the match is correct.Stochastic methods are based on probability estimates for peptide fragmentation and the subsequent generation of tandem mass spectra., one of the early algorithms in this category, the MS/MS spectrum generation is modeled by a two-step stochastic process: fragmentation and measurement. The first step, fragmentation, enumerates all the

possible fragmentation patterns of a peptide, and it determines the empirical probabilities associated with the pattern. The second step, measurement, generates tandem mass spectra from the fragments obtained in the first step, according to the distribution of the instrument measurement error. Formally, the two-step process used by SCOPE can be described as follows: for a peptide p, a fragmentation pattern F and a tandem mass spectrum S, the first step of the algorithm estimates $Pr(F|p)$,the probability of obtaining the fragmentation pattern F from the collision-induced dissociation of peptide p. The second step of the algorithm determines $\psi(S \square F, p)$, the probability of fragmentation pattern F to generate spectrum S. Finally, the probability of obtaining the spectrum S from peptide p is computed combining the two steps:

$$\psi(S,p) = \Sigma\psi(S|F,p)Pr(F|p)$$

where $\psi(p)$ is the fragment space, which contains all of the fragmentation patterns theoretically generated from peptide p. SCOPE implements a dynamic programming algorithm to efficiently compute the above formula and reports the peptide p that maximizes the score, along with its corresponding P-value.

Stochastic models to determine the best fit between a tandem mass spectrum and a sequence have been used in other programs.

Based on this algorithm we will get the Protein Medical Informations Entity attributes i.e. a)Protein-id,b)Cytogenetic Location,c)HGNC Approved Symbol through Protein OMIM Eid.

Step5: Collect all protein information from various databases through above steps and design protein databases.

Step6: In this protein databases display all entity attribute information through Protein Id using join query.

3. RESULTS

The Interleukin database model consists of Protein, Protein Information, Protein OMIM information, Protein Sequence, Protein attribute informations , Sequence and Species. The entity relationship diagram is shown here for Interleukin database. In Protein, the information related to Interleukin ID, Protein primary name, Class name, date created and created by will be given. Sequence entity contain sequence id, sequence annotation, sequence load date, sequence length, accession number. Protein Sequence entity contain attributes are sequence id, protein id, date created, created By, Comment. Similarly Species entity contain the species id, protein id, scientific name, common name, Date created, created By attributes. Similarly Protein ontologies entity contain the protein id, cellular component, coding sequence diversity, Molecular function, technical term attributes. Similarly Protein Attribute entity contain Protein Gid, protein id, sequence length, species id, sequence status, protein existence attributes. Similarly Protein OMIM entity contain Protein Eid,protein id, cytogenetic location, HGNC approved symbol attributes. In each and every entity contain protein id attribute, through protein id attribute we will display the all related informations in the Database.

4. CONCLUSION

In this work an attempt performed for creation of Interleukin database for the first time by considering hypothetical information on the details of the Interleukins. As the number of discoveries on the Interleukins is increasing day by day the creation of a separate protein biological database for Interleukins along with effective robustic DBMS method is used. There are many applications of Interleukins in Biology which make them unique to understand and the information exclusively on them is very much useful to all the researchers.

5.REFERENCES

[1]. Achuthsankar S Nair Computational Biology & Bioinformatics – A gentle Overview, Communications of Computer Society of India, January 2007

[2]. Aluru, Srinivas, ed. Handbook of Computational Molecular Biology. Chapman & Hall/Crc, 2006.ISBN 1-58488-406-1 (Chapman & Hall/Crc Computer and Information Science Series)

[3]. Baldi, P and Brunak, S, Bioinformatics: The Machine Learning Approach, 2nd edition. MIT Press, 2001. ISBN 0-262-02506-X

[4]. Barnes, M.R. and Gray, I.C., eds., Bioinformatics for Geneticists, first edition. Wiley, 2003. ISBN 0-470-84394-2

[5]. Peri S, et al. (2003). "Development of human protein reference database as an initial platform for approaching systems biology in humans". Genome Research 13: 2363–71.doi:10.1101/gr.1680803.

[6]. Gandhi, T.K.B. et al. Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. Nature Genetics. 2006. 3, 285–293

[7]. Mathivanan, S. et al. An evaluation of human protein–protein interaction data in the public domain. BMC Bioinformatics. 2006. 7, S19

[8]. Mishra, G. et al. Human protein reference database—2006 update. Nucleic Acids Research. 2006. 34, 411–414

[9]. Mathivanan, S. et al. Human Proteinpedia enables sharing of human protein data. Nature Biotechnology. 2008. 26, 164–167

[10]. Amanchy, R. et al. A compendium of curated phosphorylation-based substrate and binding motifs. Nature Biotechnology. 2007. 25, 285–286

[11]. Mathivanan S, Periaswamy B, Gandhi TK et al. (2006). "An evaluation of human protein-protein interaction data in the public domain". BMC Bioinformatics. 7 Suppl 5: S19.doi:10.1186/1471-2105-7-S5-S19. PMC 1764475. PMID 17254303.

[12]. Brocker, C; Thompson, D; Matsumoto, A; Nebert, DW; Vasiliou, V (Oct 2010). "Evolutionary

divergence and functions of the human interleukin (IL) gene family.". Human Genomics 5 (1): 30–55. doi:10.1186/1479-7364-5-1-30. PMC 3390169. PMID 21106488.

[13]. Khadka, A (2014). "Interleukins in Therapeutics". PharmaTutor 2 (4): 67–72.

[14]. Priestle JP, Schär HP, Grütter MG (December 1989). "Crystallographic refinement of interleukin 1 beta at 2.0 A resolution". Proc. Natl. Acad. Sci. U.S.A. 86 (24): 9667–71. doi:10.1073/pnas.86.24.9667. PMC 298562. PMID 2602367.

[15]. Arai K, Yokota T, Arai N, Lee F, Rennick D, Mosmann T (1985). "Use of a cDNA expression vector for isolation of mouse interleukin 2 cDNA clones: expression of T-cell growth-factor activity after transfection of monkey cells". Proc. Natl. Acad. Sci. U.S.A. 82 (1): 68–72. doi:10.1073/pnas.82.1.68. PMC 396972. PMID 3918306

[16]. ALSTON, y., COOMBS, j. (1992), Biosciences, Information Sources and Services, New York: Stockton Press.

[17]. CRAFTS-LIGHTLY, A. (1986), Information Sources in Biotechnology, Weinheim: VCH Verlagsgesell-schaft.

[18]. GRINDLEY, J.N., BENNETT, D.J. (1993), Public perception and the socio-economic integration of biotechnology, in: Biotechnologia 20, 89-102.

[19]. LŰCKE, E.-M., POETZSCH, E. (1993), Biotechnology Directory Eastern Europe, Berlin-New York: de Gruyter.

[20]. MARCACCIO, K. Y. (1993), Gale Directory of Databases, Vol. 1: Online Database, Detroit: Gale Research Inc.

[21]. MEWES, H.-W. (1990), Workshop Computer Applications in Biosciences, Book of Abstracts, p. 11, Martinsried.

[22]. POETZSCH, E. (1986), Faktographische Informationsfonds zur Biotechnologie, Berlin: WIZ

[23]. Yates, J.R., Speicher, S., Griffin, P.R. & Hunkapiller, T. Peptide mass maps: a highly informative approach to protein identification. *Anal. Biochem.* **214**,397–408 (1993).

[24]. James, P., Quadroni, M., Carafoli, E. & Gonnet, G. Protein identification by mass profile fingerprinting. *Biochem. Biophys. Res. Commun.* **195**, 58–64 (1993).

[25]. Mann, M., Hojrup, P. & Roepstorff, P. Use of mass spectrometric molecular weight information to identify proteins in sequence databases. *Biol. Mass Spectrom.* **22**, 338–345 (1993).