Journal of
Proteomics & Bioinformatics

Review Article Open Access

Statistics and Bioinformatics Ethnicity

Siva Kishore Nandikolla[1]* and Mahaboobbi Shaik[2]

[1]Department of Biochemistry and Bioinformatics, GITAM University, Visakhapatnam, India
[2]Department of Biotechnology, Andhra University, Visakhapatnam, India

Abstract

The outcome of the information by bioinformatics is enormous nowadays. Due to this rapid increment in information, collecting and analysing became a quite hard job. To reduce this encumbrance, various statistical methods are developed. This review paper contains how statistics helped in tool designing for finding genes in genomic DNA where biologists needed in plenty, in Phylogenetic trees, in microarrays, in alignment, in sequence analysis, in BLAST and various other implementations.

Keywords: Statistics; Bioinformatics; Phylogenetic Trees; Microarrays; BLAST

Introduction

Statistics means collection, organization, analysis, and interpretation of given data. It includes planning of data collection in terms of the design of surveys and experiments."Manuscript on Deciphering Cryptographic Messages", is the first book written on statistics in 9th century written by Al-Kindi (801–873 AC). Al-Kindi described in detailed the use of **statistics** and frequency analysis; this was the birth of both cryptanalysis and statistics. Its mathematical foundations in this field were included in 17th century with the joint development by Blaise Pascal and Pierre de Fermat in probability theory. The scope of statistics increased in the early 19th century which included the collection and analysis of data in general. This probability theory aroused from the studies made in games of chance. Large-scale statistical computation is required in computation nowadays, and new methods that are tough to perform manually are done with the help of statistics.

Biostatistics plays a critical importance in the foundation of modern biology theories. The rediscovery of Mendel's work created gaps in understanding between genetics and evolutionary Darwinism and led to debate between biometricians like, Walter Weldon, Karl Pearson, Charles Davenport, William Bateson and Wilhelm Johannsen. Models built on statistical reasoning had helped to solve these variances and to generate the neo-Darwinian modern evolutionary synthesis.

Bioinformatics is a novel branch of science stands in-between biology and informatics, which is itself a new area of research. Therefore, bioinformatics is concerned with creation and application of information-based methodologies to analyze biological data sets and the contained information. The wide adoption of technologies like microarrays, genome sequencing projects has resulted in accumulation of large amount of data daily. Hence, to extract automatically extraction and analysis of these data sets is required. To fill this gap new tools are designed with the help of bioinformatics. Mathematical techniques and statistical methods are the natural solution to this problem.

Statistics is helpful in predicting unknown system with the application of mathematical model to the observations obtained from the unknown system. Various fields and recent applications which received the boon of statistics in bioinformatics are depicted in this review article. Some examples like

- In-gel digestion and mass spectrometry analysis [1].
- The prediction of functional single nucleotide polymorphism (SNP) is promising in modern genetics analysis. Computational biology technology has facilitated an increase in the successful rate of genetic association study and reduced the cost of genotyping [2].
- Early drug discovery genomics and proteomics driven [3]. Chromatographic peak areas were integrated and normalized to the internal standard. Log values of these 24 metabolites were averaged and compared between time points. Triplicate injections of each sample were analyzed and the results were averaged. Metabolite profiles were analyzed using JMP 5.1 Statistical Software (SAS Institute, Cary, USA). ANOVA was used to examine the statistical difference in the quantities of the representative metabolites across the three time points. Using the hierarchical cluster analysis (HCA) platform, metabolite data were clustered into similar groups where values are statistically close together relative to other clusters. For principal component analysis (PCA), the principal components were derived from an Eigen value decomposition of the correlation matrix. Optimization can be done via Response Surface Method (RSM) based on 5-level, 4-variable of Central Composite RotaTable Design (CCRD) [4].
- By Molecular Evolutionary Genetic Analysis (MEGA) software (version 4.0.02) [5], which uses UPGMA method, Phylogenetic analysis of the sequences can be done. Each node can be tested using the bootstrap approach to ascertain the reliability of nodes. The number is indicated in percentages against each node [6].

Application of Bayesian Method

Implementations

- Construction of PreLocABC by integrating five sub-modules

***Corresponding author:** Siva Kishore Nandikolla, Department of Biochemistry and Bioinformatics, GITAM University, India, E-mail: sivakishoren@gmail.com

Received November 07, 2011; **Accepted** December 06, 2011; **Published** December 20, 2011

Citation: Siva Kishore N, Mahaboobbi S (2011) Statistics and Bioinformatics Ethnicity. J Proteomics Bioinform 4: 294-301. doi:10.4172/jpb.1000204

based on the bayesian model. In this step, PreLocABC integrated five sub-modules mentioned above to assign the likelihood to the nine compartments for each protein. First, the training dataset was used to train and calculate weight coefficients of five sub-modules and threshold values of each sub-cellular localization score. The sub-cellular compartments are denoted as follows:

$$\boldsymbol{C} = \{\boldsymbol{C}_i\}£–\boldsymbol{i} = 1, 2, 3, 4, 5, 6, 7, 8, 9 \qquad (1)$$

Where *i* refers to nine compartments: cytoplasm, ER, extracellular space, Golgi apparatus, lysosome, mitochondrion, nucleus, peroxisome and plasma membrane.

The five integrated sub-modules are denoted as follows:

$$P =\{ P_j \}, =1, 2, 3 ,4, 5 \qquad (2)$$

Where *j* refers to the serial number of five sub-modules.

The scores of the Bayesian model are defined as follows:

$$Score[C_i] = \sum L_{P_j}[C_i] * W_{P_j}[C_i] \qquad (3)$$

Where L_{P_j} $[C_i]$ refers to the likelihood of a protein localized in $\boldsymbol{C}_\mathrm{i}$ predicted by $\boldsymbol{P}_\mathrm{j}$ and W_{p_j} $[C_i]$ refers to the weight coefficients of likelihood in $\boldsymbol{C}_\mathrm{i}$ predicted by $\boldsymbol{P}_\mathrm{j}$. There are 45 weights for 9 sub-cellular components of 5 sub-modules. Each weight received an initial value of 0.5. In each cycle, one weight was changed ranging from 0~5 with step length 0.01 and the other 44 weights had no change. We chose the weight value when the prediction accuracy was the highest, and replaced the initial value of the weight with this training result in the next cycle [7].

- In simple approach to model the spread of a disease in a field relying on survival analysis methods dealing with approximation of times to infection. This approach allows for further developments in the model and a Bayesian development could be a possible direction as practitioners might have prior information on the propagation mechanism [8].

Quantile regression

Quantile regression is a type of regression analysis. Quantile regression models are used in various applications [9]. It has flexibility for modeling data with heterogeneous conditional distributions. Whereas the least squares method results in approximation of the conditional *mean* of the response variable given values of the predictor variables, quantile regression results in approximating either the median or other quantiles of the response variable [10].

Quantiles [11]

Let *Y* be a real valued random variable with distribution function $F_Y(y)=P(Y\leq y)$. The τ^{th} quantile of Y is given by

$$Q_Y(\tau) = F_Y^{-1}(\tau) = \inf\{y : F_Y(y) \geq \tau\}$$

Where $\tau \in [0,1]$

Define the loss function as $\rho_\tau(y) = y(\tau - I(y < 0))$. A specific quantile can be found by minimizing the expected loss of $Y - u$ with respect to *u*:

$$\min_u E(\rho_\tau(Y-u)) = \min_u(\tau-1)\int_{-\infty}^{u}(y-u)dF_Y(y) +$$
$$\tau\int_u^{\infty}(y-u)dF_Y(y)$$

This can be shown by setting the derivative of the expected loss function to 0 and letting q_τ be the solution of

$$0 = (1-\tau)\int_{-\infty}^{q_\tau} dF_Y(y) - \tau\int_{q_\tau}^{\infty} dF_Y(y)$$

This equation reduces to

$$0 = F_Y(q_\tau) - \tau,$$

and then to

$$F_Y(q_\tau) = \tau.$$

Hence q_τ is τth quantile of the random variable Y.

Conditional quantile and quantile regression

Suppose the τ^{th} conditional quantile function is $Q_{Y|X}(\tau) = X\beta_\tau$. Given the distribution function of *Y*, β_τ can be obtained by solving

$$\beta_\tau - \underset{\beta\in R^k}{\arg\min}\, E(\rho_\tau(Y - X\beta))$$

Solving the sample analog gives the estimator of β.

$$\hat{\beta}_\tau - \underset{\beta\in R^k}{\arg\min} \sum_{i=1}^{n}(\rho_\tau(Y_i - X\beta))$$

Some statistics packages, such as R, Eviews (ver. 6), Stata (via qreg), gretl, SAS through proc quantreg (ver. 9.2), and RATS include implementations of quantile regression. R implements it through Roger Koenker's quantreg package.

Bayesian analysis helps to estimate the segregation ratio in nuclear systems when there is an ascertainment bias. Consider the situation in which the proband probabilities differ with the number of affected siblings, showed the effect of familial correlation among siblings within the same family [7].

In Microarray

Microarray experiments generate the sort of data where the number of measurements of each sample is much greater than the number of samples. These massively multivariate datasets are unable to be analyzed by traditional statistical methods. CSIRO developed GeneRave, a new statistical technique specifically for microarray data. GeneRave is able to cope with the multivariate nature of microarray data and extract meaningful information from it. GeneRave is also able to handle other types of multivariate data, including protein expression data and single nucleotide polymorphism (SNP) data. GeneRave can be used for developing:

- simpler clinical diagnostics
- efficient screening methods for potential drug candidates - 'toxicogenomics'
- Genetic tests for drug efficacy - 'pharmacogenomics'.

Implementations

By using Affymetrix Genechips, *CGR1, GOS1, ICS2, PCL5 and PLB1,* the high probabilities of being differentially expressed (up or down) were found to be in excellent agreement with the expression status determined by the independent, high precision confirmatory experiments are obtained by the probabilistic framework in *Saccharomyces cerevisiae* [12].

In nutrition new advanced methodologies for the analysis of DNA, RNA, protein, low-molecular-weight metabolites, microarray gene expression, real-time polymerase chain reaction, proteomics, and other bioinformatics technologies as well as access to bioinformatics databases are developed. Statistics, by making scientific inferences from data that contain variability, has played an important role in advancing nutritional sciences [13].

The capacity to differentiate from human mesenchymal stem cells (MSC) into osteoblasts, have good scope for the development of new treatment strategies, like improved healing of large bone defects. With 30,000 elements human oligonucleotide microarrays are developed with 30,000 elements and then large-scale expression profiling of long-term expanded MSC and MSC during differentiation into osteoblasts is performed. Microarray analysis of MSC during osteogenic differentiation identified three candidate genes for further examination and functional analysis: *ID4, CRYAB,* and *SORT1.* The expression levels of bone related genes like *RUNX2, SPP1, COL1A1, COL3A1, BGLAP, ALPL, and FOSL1)* and mesenchymal stem cells marker (*CD105*) were analyzed, using real time Reverse Transcription-Polymerase Chain Reaction [14].

Supported by new high-throughput methods (454 pyrosequencing, PhyloChip microarrays) and strategies (barcoding); the surveys of 16S gene in the human microbiota attempt to provide a comprehensive picture of the community differences between healthy and diseased states [15].

In Phylogenetic trees

The construction, or more accurately the estimation, of phylogenetic trees is of interest in its own right in evolutionary studies. It is also useful in many other ways, for example in the prediction of gene function. The evolutionary relationships between a set of species is represented by a binary tree, and in this book we consider only binary trees. "Species" may refer either to organisms or to sequences such as protein or DNA. It is required that the set of "species" have a common ancestor, so that construct a tree relating the various hemoglobins. The method of construction of phylogenetic tree based on molecular data is widely used to determine evolutionary relationships [16]. Several algorithmic procedures used in tree reconstruction are based on the concept of a *distance* between species. Let S be a set of points. The standard requirements for a distance measure

on S are that for all x and y in S, (i) $d(x, y) \geq 0$, (ii) $d(x, y) = 0$ if and only if $x = y$, (iii) $d(x, y) = d(y, x)$, and (iv) for all x, y, and z in S $d(x, y) \leq d(x, z) + d(z, y)$.

- Tree Reconstruction: The Ultrametric Case
- Tree Reconstruction: the Neighbor-Joining Approach
- Inferred Distances
- Tree Reconstruction: Parsimony
- Tree Estimation: Maximum Likelihood

Implementations

The TnrA and GlnA sequences of *B. clausii* and *B. halodurans* were compared with its ortholog sequences in the NCBI database using BLAST and aligned using the molecular evolutionary genetics analysis (MEGA) software, version 4.0. Phylogenetic trees were subsequently constructed by the neighbor-joining (NJ) method [17].

To know the evolutionary relationship of BZIB (BZIP are a class of dimeric sequence specific DNA-binding proteins) existing among the plants Phylogenetic tree is helpful [18].

In BLAST

The Basic Local Alignment Search Tool (BLAST), the program which compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches, was used to infer evolutionary relationships between RNA and peptide sequences (http://blast.ncbi.nlm.nih.gov/) [19]. In BLAST we compare a "query" sequence with a large number of database sequences, leading to a large number of match scores. In order to understand the theory for testing hypotheses and estimating parameters in bioinformatics applications, we must therefore consider the theory of many random variables. Geometric-like random variables are used in BLAST theory. The use of the maximum of several random variables as a test statistic arises in several areas in bioinformatics, and in particular in BLAST.

Implementations

The nucleotide and protein sequence of Indian isolate was analyzed using Blast program available at NCBI with test virus coat protein gene sequences from all around the world, available in the database of NCBI [20].

Amino acid sequences of NA and HA of swine influenza virus sub-type H1N1 strains were used for screening of 98-99% similar sequences available in non-redundant (nr) database situated at NCBI using BLASTP [21].

Sequence and structural similarity

Sequence similarity search is the principle technique in sequence analysis, adapted for understanding the biological significance of a sequence [22]. Homology-based tools such as BLAST [23], PSI-BLAST [24] and HMMER [25] are used to detect sequence homology between pairs of proteins or against protein family databases and infers functional similarity from homology. Template sequence were selected by search submits after sequencing with BLASTP [26]. Some studies [27,28] suggest that homology based tools are sufficient to determine the most probable EC number for the query sequence, but less coverage is achieved with these methods. However, simple pair-wise comparisons may be misleading due to the availability of redundant protein sequences in public databases [29]. Therefore, in addition to the sequence similarity, Tian et al [30] have used the functional similarity from homology to predict enzyme classes; yet, this method only works well when two sequences are very similar. Similarly, by combining sequence similarity with other functional features such as interacting partners, Espadaler et al [31] have shown that the protein sequences with sequence similarity are more likely to exhibit the same enzymatic activity if they share the same interacting partners. Otto et al [32] and Galperin et al [33] have developed methods for identification of analogous enzymes using sequence similarity by grouping proteins that share the same enzymatic activity (EC classes) [34]. For homology search of structurally similar sequences the BlastP was performed with the sequences obtained from protein data bank. For fold assignment GenThreader server was used. The alignment was done for target protein sequences with protein databank (PDB: 1A92) template using CLUSTALX [35].

In sequencing to reveal the similarity at protein level with other existing hsr *203J* like proteins of *Brassica juncea* cv Varuna BLASTx is used [36].

Glycine betaine (N, N, N-trimethylglycine) is a effective compatible solute, which maintains fluidity of membranes and protects the biological structure of the organisms under stress [37]. Similar sequences with respect to betaine aldehyde dehydrogenase (*GbsA*) and betaine alcohol dehydrogenase (*GbsB*) genes are obtained by BLAST.

In Alignments

As a segment of genetic material is passed on through the generations in some line of descent in a population, the sequence constituting this material will change through the process of mutation. The simplest mutations are of the form of a switch from one nucleotide to another, or in the form of an insertion or a deletion. Mutations can spread to an entire species, or nearly so, through the process of natural selection or random drift. When a switch in nucleotides spreads throughout most of a species we call it a *substitution*. (When in a population at a given site there does not exist a single nucleotide type, we say that a *polymorphism* exists at that site.) As substitutions, insertions, and deletions get passed along through two independent lines of descent, the two sequences will slowly diverge from each other. For example, the original sequence may have been

cggtatgcca,

where as the two descendants might be

cgggtatccaa

and

ccctaggtccca.

Many problems in bioinformatics relate to the comparison of two (or more) DNA or protein sequences. In order to compare sequences of nucleotides or amino acids, we use alignments.

Types of alignments:

- Global alignment
- Local alignment
- Gapped alignment
- Ungapped alignment
- Pairwise alignment
- Multiple alignment

Simple tests for significant similarity in an alignment

Exactly-Matching Sub-sequences

Well-Matching Sub-sequences

Approximations

Are there approximations for the mean and variance of $Y_{\max}$ more accurate than those given in Table 3.3? Waterman (1995, page 277) claims that for large *n*, good approximations for the mean and variance are

$$\mu_{\max} = \frac{\log n + \gamma + k\log\left(\frac{\log n}{\lambda}\right) + k\log\left(\frac{1-p}{p}\right) - \log(k!)}{\lambda} - \frac{1}{2} + r_1$$

$$\sigma^2_{\max} = \frac{\pi^2}{6\lambda^2} + \frac{1}{12} + r_2$$

where $\lambda = (-\log p)$, $n = N(1-p)$, and, for $p = 14$, $|r1| \leq 3.45 \times 10{-}4$ and $|r2| \leq 2.64 \times 10_{-2.}$

These approximations were first calculated assuming independence of sub-sequence lengths, and later shown not to change significantly in the dependent case.

Sequence analysis

Multiple sequence alignment of a representative set of 25 Pmk1 genes was done by Clustal-W method. Serine/Threonine protein kinases active-site signature & protein kinase domain enclosed in black colored rectangle which are also present in TiPmk1. Phylogenetic tree of Pmk1 was constructed by UPGMA method of MEGA version 4.0.02. Phylogenetic tree of Pmk1 showed four major clusters A, B, C and D as shown in *T indica* & *U maydis* are present in same cluster D [38]. To carry out homology modeling, Automated, Alignment, and Project modes of Swiss model server can be used [39].

Custom mutation in Protein and Nucleotide sequences can enable scientists to learn about the proteins and their expression. To facilitate this, an innovative computer tool is designed using various statistical methods, that is MUTATER. Even RAW input format can also be accessed by this tool [40].

Modified classification based on probabilities predicted by the RBF network [41]

RBF network can predict the probabilities of identifying mtSNP features in a person. By studying the relations between individual mtSNPs and the persons with high predicted probabilities different classes were identified. Based on these predictions made by RBF network, modern classification is made [42].

1) Selection of target class to be analyzed.

2) Rank individuals according to their predicted probabilities belonging to the target class.

3) Either select individuals whose probabilities are greater than a certain value or select the desired number of individuals and set them as a modified cluster.

Statistical validation of motifs [43]

By definition, a phosphorylation motif is associated with phosphorylation sites in the proteome. Accordingly, candidate motifs were evaluated by measuring this association, and comparing it to a null distribution obtained by permutation. Our measure of association is relative risk for the motif, given that a site is phosphorylated and is calculated as:

RRmotif = (NPM x NS) / (NM x NPS)

Where:

RRmotif = Relative risk for phosphorylation given motif

NPM = Number of times the given motif is associated with a phosphorylated site.

NS = Total number of non-phosphorylated sites

NM = Number of times the given motif is associated with a non phosphorylated site.

NPS = Total number of phosphorylated sites.

Relative risk >1 indicate good association between the motif and phosphorylation events. Significance is measured by comparing the relative risk for the candidate motif to a motif-specific null distribution of relative risk obtained by substituting random combinations of amino acids for those that define the motif.

The p-value calculated as the probability that randomly selected mock motifs has relative risk (RRmock) at least as great as the value observed for the predicted motif. Mock motifs are obtained from a predicted motif by replacing each amino acid with randomly selected ones. Because motifs are typically small, we have exhaustively checked all possible mock motifs rather than generating them randomly.

Thus the empirical p-value is calculated as p = N/D;

Where:

N = Number of mock motifs with RRmock ≥ RRmotif

D = Total number of possible mock motifs.

Logistic regression model [44]

Logistic regression is a well established statistical model suitable for probabilistic binary classification. In this study, we used the logistic regression model to differentiate whether a residue in antigens belongs to discontinuous epitope regions or not. Three logistic regression predictors were constructed using B-factor, RASA and the combination of these two features. Each predictor was input a structural window composed of a target residue and its ***N*** spatially nearest neighbors obtained by calculating the distances between the α-carbons of residues. The optimal value of ***N*** was determined by using different widow sizes as input for logistic regression model. Thus, each residue was represented by ***N***+1 input vectors if a single feature was used and by 2×(***N***+1) input vectors if the combined features was used. Assuming ***yr***∈{0, 1} and ***xr*** = { ***xr1*** , ***xr2*** ,…, ***xrj*** } are the class label and input vectors for a target residue ***r***, the logistic regression predictors assigned a probability ***θr*** = ***P***{ ***yr*** = 1|xr} to the target residue using the log it function:

$$\log\left(\frac{\theta_r}{1-\theta_r}\right) = \alpha + \beta_1 x_{r1} + \beta_2 x_{r2} + \cdots + \beta_j x_{rj}$$

Where ***α***, ***β***1, ***β***2 ,…, ***β***j are the model parameters. The logistic regression predictors were implemented with the LR-TIRLS package (http://komarix.org/ac/lr/#lr-tirls). Generally, the prediction threshold of standard logistic regression model was set to 0.5. However, in our study, the optimal threshold was determined when the predictor achieved the best Matthew's correlation coefficient (MCC) value of cross-validation [44].

Multivariate statistical analysis [45]

For the multivariate statistical analysis, the quantification parameters, generated by the commercial software, were organized in the form of a matrix, X (descriptor matrix), where the rows represent gel samples and each column (variable) represents one of the sub quadrants in which each gel image was partitioned and the integral of the intensities of the protein spots in a given sub quadrant the X-value.

Bi-linear regression models allow for more precise and accurate estimates of abundance, in comparison to methods that treat each spectrum independently, by taking into account the abundance of the molecule throughout the entire elution profile, with precision increased by one-to-two orders of magnitude [46].

Change in statistical method can affect the power profiles of Genome Wide Association (GWA) predictions. Older simulation studies of a single synthetic phenotype marker determined that the gene model or mode of inheritance (MOI) was a major influence on power [47].

Label-free shotgun proteomics is a semi-quantitative protein profiling method which can compare large number of samples in a single experiment [48].

Streptococcus mutans is a major microorganism for dental caries worldwide and is considered as the most cariogenic of all of the oral streptococci. Using HMM and BLAST, novel protein domains are identified and its function is predicted [49] which causes the dental caries.

Next generation sequencing is a new revolution in biological research. In this, Poisson modeling adopted solving a convex optimization problem in isoform expression using high throughput RNA sequencing (RNA-Seq) data [50]. ChIP-seq offers genom-wide coverage, the counts in ChIP-seq data in the two states were modeled by a generalized Poisson and a zero-inflated Poisson, hierarchical hidden Markov model to combine individual hidden Markov models are used. Statistics is also used in Protein sequence analysis [51].

Simplified models of the nervous system with neurons as simple processing units linked with weighted connections called synaptic efficacies are called Artificial Neural Networks (ANNs). Weights are gradually adjusted according to a *learning* algorithm To predict the α-helix, β-sheet and coil regions of this protein family Bayesian Regularization Feed-forward Back propagation Neural Network Technique was used. PSI-BLAST is used to study multiple-sequence alignment [52]. SCOP and PDB database has been used for searching the primary and secondary structure of proteins and for training the data set. Mathematic models for genetic mutation are used successfully in HIV [53]. The amino acid sequence alignment between the template and the final model of AHA1 was generated using CLUSTALW program [54].

Other Implementations of statistical methods

Huge databases with various molecular information of complex biological systems are opened up to researchers because of the

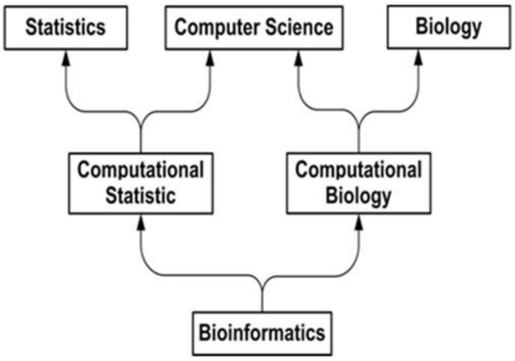

Figure 1: The interrelationship.

contribution made completion of complete genome sequencing of human and other organisms. A discrete event based stochastic modeling approach for studying the molecular dynamics of cells. Simulation methodology and present the mathematical formalism underlying the in silico system was developed called *iSimBioSys*, which interactively simulates the dynamics of a biological process [55].

Correlation dendrogram of m/z values built for the whole dataset. Average linkage clustering was used, distance defined as 1– |*r*|, where *r* is Pearson correlation coefficient [56].

Meta-analysis is an important method for integration of information from multiple studies. Various meta-analysis methods have been proposed for synthesizing information from multiple candidate gene studies and QTL mapping experiments, but there are several questions and challenges associated with these methods. Dirichlet Process Prior (DPP), which relaxes the normality assumptions about study specific outcomes. With a DPP model, the posterior distribution of outcomes is discrete, reflecting a clustering property that may have biological implications [57].

The difficulties found in the resolution of atomic level structures for interacting pairs, make the predictive power of molecular computational biology methods essential for the advancement of the field. Indeed, bridging the gap formed due to the lack of structural details can therefore transform systems biology into models that more accurately reflect biological reality [58].

Log-logistic and Weibull distributions have both accelerated survival time property [59]. The log-logistic distribution has also proportional odds property. Log-logistic distribution has unimodal hazard curve which changes direction. Link [60,61] presented a confidence interval estimate of survival function using Cox's proportional hazard model with covariates. Her idea more recently extended by to the exponential distribution and to exponential proportional hazard model, respectively [62]. The same idea has been extended to the Weibull proportional hazard regression model. In this study, it is formed on confidence interval for log-logistic distribution survival function for any values of the time provided that the survival times have a log-logistic distributed random variable. It is also extended the same results to the proportional odds regression. A Real time data and a simulation data example are also considered in the study for illustration the discussed confidence interval.

Kaplan-Meier survival function is the most commonly used statistical technique of survival analysis and has some drawbacks. To overcome it exact waiting time survival function is developed [63]**.** The proposed procedures are applied to a lung cancer data set.

Statistical solutions for the search of hotspots based on the "Peaksheight distribution", which account within the null hypothesis for the possible non-random behaviour of the integrations in finding Common Integration Sites (CIS) or hotspots in Gene therapy [64].

Using only the transcription network structure information, a probabilistic model was developed that computes the probabilities with which a pair of genes responds simultaneously (*SR*) or differentially (*DR*) to a random network perturbation is helpful in studying Yeast Gene Regulatory Network [65].

Statistical tools

Tools which are designed for statistical calculations in bioinformatics are

- SAS® 9.1.9
- MATLAB® 7.5.0 [66]
- Biogeme
- Dataplot™
- The BUGS Project
- ROSETTA
- JMP® Genomics
- SPSS®

Conclusion

The recent advancements in the field of Bioinformatics and the usage of various statistical methods like Bayesian method, Quantile regression, Logistic regression model in different fields like algorithms in BLAST, Phylogenetic, Sequence analysis, Microarrays are briefly depicted in the review. Bi-linear regression models, statistical validations of motifs, multivariate statistical analysis implementations are also depicted. Statistical techniques also have applications in: drug discovery, personalized medicine and clinical diagnostics.

References

1. Chen A, Guo Z, Zhou L, Yang H (2010) Hepatic Endosome Protein Profiling in Apolipoprotein E Defi cient Mice Expressing Apolipoprotein B48 but not B100. J Bioanal Biomed 2: 100-106.
2. Nuchnoi P, Nantakomol D, Chumchua V, Plabplueng C, Isarankura- Na-Ayudhya C, et al. (2011) The Identification of Functional Non-Synonymous SNP in Human ATP-Binding Cassette (ABC), Subfamily Member 7 Gene: Application of Bioinformatics Tools in Biomedicine. J Bioanal Biomed 3: 026-031.
3. Arun B (2009) Challenges in Drug discovery: Can We Improve Drug Development. J Bioanal Biomed 1: 050-053.
4. Syamsul Kamar MW, Salina MR, Siti Salhah O, Hanina MN, Mohd Basyaruddin AR (2011) Optimization of Lipase Catalyzed Synthesis of Nonyl Caprylate using Response Surface Methodology (RSM). J Biotechnol Biomaterial 1: 106.
5. Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) Software Version 4.0. Mol Biol Evol 24: 1596-1599.
6. Pandey S, Negi YK, Marla SS, Arora S (2011) Comparative Insilico Analysis of Ascorbate Peroxidase Protein Sequences from Different Plant Species. J Bioengineer & Biomedical Sci 1: 103.
7. Nandram B, Xu H (2011) Bayesian Corrections of a Selection Bias in Genetics. J Biomet Biostat 2: 112.
8. Gouno E (2011) Modelling Spread of Diseases Using a Survival Analysis Technique. J Biomet Biostat 2: 113.
9. Yu K, Lu Z, Stander J (2003) Quantile regression: applications and current research area. The Statistician 52: 331-350.
10. Alhamzawi R, Yu K, Pan J (2011) Prior Elicitation in Bayesian Quantile Regression for Longitudinal Data. J Biomet Biostat 2: 115.
11. Koenker R (2005) Quantile Regression. New York: Cambridge University Press.
12. Gelmi CA, Prakash P, Edwards JS, Ogunnaike BA (2011) Experimental Validation of a Probabilistic Framework for Microarray Data Analysis. J Biomet Biostat 2: 114.

13. Fu WJ, Stromberg AJ, Viele K, Carroll RJ, Wu G (2010) Statistics and bioinformatics in nutritional sciences: analysis of complex data in the era of systems biology. J Nutr Biochem 21: 561–572.

14. Sollazzo V, Palmieri A, Girardi A, Farinella F, Carinci F (2011) Trabecular Titanium Induces Osteoblastic Bone Marrow Stem Cells Differentiation. J Biotechnol Biomaterial 1: 102.

15. Fabrice A, Didier R (2009) Exploring Microbial Diversity Using 16S rRNA High-Throughput Methods. J Comput Sci Syst Biol 2: 074-092.

16. Dangre DM, Rathod DP, Gade AK, Rai MK (2009) An in Silico Molecular Evolutionary Analysis of Selected Species of Phoma: A Comparative Approach. J Proteomics Bioinform 2: 295-309.

17. Farazmand A, Yakhchali B, Shariati P, Ofoghi H (2011) Bacillus clausii and Bacillus halodurans lack GlnR but Possess Two Paralogs of glnA. J Proteomics Bioinform 4: 179-183.

18. Selvaraj D, Loganathan A, Sathishkumar R (2010) Molecular Characterization and Phylogenetic Analysis of BZIP Protein in Plants. J Proteomics Bioinform 3: 230-233.

19. Nahalka J (2011) Quantification of Peptide Bond Types in Human Proteome Indicates How DNA Codons were Assembled at Prebiotic Conditions. J Proteomics Bioinform 4: 153-159.

20. Neha S, Vrat BS, Kumud J, Thakur PD, Rajinder K, et al. (2011) Comparative In silico Analysis of Partial Coat Protein Gene Sequence of Zucchini Yellow Mosaic Virus Infecting Summer Squash (Cucurbita pepo L.) Isolated From India. J Proteomics Bioinform 4: 068-073.

21. Sharma DK, Rawat AK, Srivastava S, Srivastava R, Kumar A (2010) Comparative Sequence Analysis on Different Strains of Swine Influenza Virus Sub-type H1N1 for Neuraminidase and Hemagglutinin. J Proteomics Bioinform 3: 055-060.

22. Saravanan V (2010) Mass Blaster V1.0 – A Perl Gui Tool for Mass Sequence Blast and Gene Prediction. J Proteomics Bioinform 3: 302-304.

23. Altschul SF, Gish W (1996) Local alignment statistics. Methods Enzymol 266: 460-480.

24. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389-3402.

25. Durbin R, Eddy S, Krogh A, Mitchison G (1998) Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge University Press.

26. Vaseeharan B, Valli SJ (2011) In silico Homology Modeling of Prophenoloxidase activating factor Serine Proteinase Gene from the Haemocytes of Fenneropenaeus indicus. J Proteomics Bioinform 4: 053-057.

27. Shah I, Hunter L (1997) Predicting enzyme function from sequence: a systematic appraisal. Proc Int Conf Intell Syst Mol Biol 5: 276-283.

28. Audit B, Levy ED, Gilks WR, Goldovsky L, Ouzounis CA (2007) CORRIE: enzyme sequence annotation with confidence estimates. BMC Bioinformatics 8: S3.

29. Rost B (2002) Enzyme function less conserved than anticipated. J Mol Biol 318: 595-608.

30. Tian W, Skolnick J (2003) How well is enzyme function conserved as a function of pairwise sequence identity? J Mol Biol 333: 863-882.

31. Espadaler J, Eswar N, Querol E, Aviles FX, Sali A, et al. (2008) Prediction of enzyme function by combining sequence similarity and protein interactions. BMC Bioinformatics 9: 249.

32. Otto TD, Guimaraes AC, Degrave WM, de Miranda AB (2008) AnEnPi: identification and annotation of analogous enzymes. BMC Bioinformatics 9: 544.

33. Galperin MY, Walker DR, Koonin EV (1998) Analogous enzymes: independent inventions in enzyme evolution. Genome Res 8: 779-790.

34. Mohammed A, Guda C (2011) Computational Approaches for Automated Classification of Enzyme Sequences. J Proteomics Bioinform 4: 147-152.

35. Singh S, Gupta SK, Nischal A, Khattri S, Nath R, et al. (2010) Comparative Modeling Study of the 3-D Structure of Small Delta Antigen Protein of Hepatitis Delta Virus. J Comput Sci Syst Biol 3: 001-004.

36. Mishra A, Pandey D, Goel A, Kumar A (2010) Molecular Cloning and In silico Analysis of Functional Homologues of Hypersensitive Response Gene(s) Induced During Pathogenesis of Alternaria Blight in Two Genotypes of Brassica. J Proteomics Bioinform 3: 244-248.

37. Rajan LA, Vinodhini K, Rajalakshmi Y, Umashankar V (2011) Molecular Cloning and In Silico Sequence Analysis of Glycine Betaine Biosynthesis Genes in Bacillus subtilis. J Biotechnol Biomaterial 1: 103.

38. Gupta AK, Goel A, Seneviratne JM, Joshi GK, Kumar A (2011) Molecular Cloning of MAP Kinase Genes and In silico Identification of their Downstream Transcription Factors Involved in Pathogenesis of Karnal bunt (Tilletia indica) of Wheat. J Proteomics Bioinform 4: 160-169.

39. Barh D, Misra AN, Kumar A (2010) In Silico Identification of Dual Ability of N. gonorrhoeae ddl for Developing Drug and Vaccine Against Pathogenic Neisseria and Other Human Pathogens. J Proteomics Bioinform 3: 082-090.

40. Butt AM, Ahmed A (2009) MUTATER: Tool for the Introduction of Custom Position Based Mutations in Protein and Nucleotide Sequences. J Proteomics Bioinform 2: 344-348.

41. Takasaki S (2011) Mitochondrial Haplogroups Associated with Japanese Centenarians, Alzheimer's Patients, Parkinson's Patients, Type 2 Diabetes Patients, Healthy Non-Obese Young Males, and Obese Young Males. J Proteomics Bioinform 4: 106-112.

42. Takasaki S (2009) Mitochondrial haplogroups associated with Japanese centenarians, Alzheimer's patients, Parkinson's patients, type 2 diabetic patients and healthy non-obese young males. J Genetics and Genomics 36: 425-434.

43. Amanchy R, Kandasamy K, Mathivanan S, Periaswamy B, Reddy R, et al. (2011) Identification of Novel Phosphorylation Motifs Through an Integrative Computational and Experimental Analysis of the Human Phosphoproteome. J Proteomics Bioinform 4: 022-035.

44. Liu R, Hu J (2011) Prediction of Discontinuous B-Cell Epitopes Using Logistic Regression and Structural Information. J Proteomics Bioinform 4: 010-015.

45. Mazzara S, Cerutti S, Iannaccone S, Conti A, Olivieri S, et al. (2011) Application of Multivariate Data Analysis for the Classification of Two Dimensional Gel Images in Neuroproteomics. J Proteomics Bioinform 4: 016-021.

46. Eckel-Passow JE, Mahoney DW, Oberg AL, Zenka RM, Johnson KL, et al. (2010) Bi-Linear Regression for 18O Quantification: Modeling across the Elution Profile. J Proteomics Bioinform 3: 314.

47. Cooley P, Clark R, Folsom R, Page G (2010) Genetic Inheritance and Genome Wide Association Statistical Test Performance. J Proteomics Bioinform 3: 321-325.

48. Zhang R, Barton A, Brittenden J, Huang JT, Crowther D (2010) Evaluation for Computational Platforms of LC-MS Based Label-Free Quantitative Proteomics: A Global View. J Proteomics Bioinform 3: 260-265.

49. Varsale AR, Wadnerkar AS, Mandage RH, Jadhavrao PK (2010) Cheminformatics. J Proteomics Bioinform 3: 253-259.

50. Datta S, Datta S, Kim S, Chakraborty S, Ryan SJ (2010) Statistical Analyses of Next Generation Sequence Data: A Partial Overview. J Proteomics Bioinform 3: 183-190.

51. Dangre DM, Deshmukh SR, Rathod DP, Umare VD, Ullah I (2010) Prediction and Comparative Analysis of MHC Binding Peptides and Epitopes in Nanoviridae Nano-organisms. J Proteomics Bioinform 3: 155-172.

52. Yadav BS, Pokhariyal M, Ratta B, Rai G, Saxena M, et al. (2010) Predicting Secondary Structure of Oxidoreductase Protein Family Using Bayesian Regularization Feed-forward Backpropagation ANN Technique. J Proteomics Bioinform 3: 179-182.

53. Towfic G, Munshower J, Kettoola S, Towfic F, Graziano F, et al. (2009) Genetic Mutations Affecting the Success and Failure of HIV Regimens. J Proteomics Bioinform 2: 372-379.

54. Kumar S, Sahu BB, Tripathy NK, Shaw BP (2009) In Silico Identification of Putative Proton Binding Sites of a Plasma Membrane H+-ATPase Isoform of Arabidopsis Thaliana, AHA1. J Proteomics Bioinform 2: 349-359.

55. Ghosh S, Ghosh P, Basu K, Das SK, Daefler S (2011) A Discrete Event Based Stochastic Simulation Platform for ‘In silico’ Study of Molecular-level Cellular Dynamics. J Biotechnol Biomaterial S6: 001.

56. Pyatnitskiy M, Karpova M, Moshkovskii S, Lisitsa A, Archakov A (2010) Clustering Mass Spectral Peaks Increases Recognition Accuracy and Stability of SVM-based Feature Selection. J Proteomics Bioinform 3: 048-054.

57. Wu XL, Gianola D, Hu ZL, Reecy JM (2011) Meta-Analysis of Quantitative Trait Association and Mapping Studies using Parametric and Non-Parametric Models. J Biomet Biostat S1: 001.

58. Piccoli S, Giorgetti A (2011) Perspectives on Computational Structural Bio-Systems. J Bioprocess Biotechniq 1: 104e

59. Alukas K, Erilli NA (2011) Confidence Intervals Estimation for Survival Function in Log-Logistic Distribution and Proportional Odds Regression Based on Censored Survival Time Data. J Biomet Biostat 2: 116.

60. Link CL (1984) Confidence Intervals for The Survival Function Using Cox’s Proportional Hazard Model with Covariates Biometrics 40: 601-609.

61. Link CL (1986) Confidence Intervals for the Survival Function in the Presence of Covariates. Biometrics 42: 219-220.

62. Alakus K (2010) Confidence Intervals Estimation for Survival Function in Weibull Proportional Hazards Regression Based on Censored Survival Time Data. Science Research and Essays 5: 1589-1594.

63. Zaman Q, Strasak AM, Pfeiffer KP (2011) Exact Waiting Time Survival Function. J Biomet Biostat 2: 117.

64. Alessandro A, Di Serio C (2009) Vectors and Integration in Gene Therapy: Statistical Considerations. J Comput Sci Syst Biol 2: 117-123.

65. Pinto FR (2009) A Probabilistic Approach to Study Yeast’s Gene Regulatory Network. J Comput Sci Syst Biol 2: 044-050.

66. Apraiz I, Leoni G, David L, Persson JO, Cristobal S (2009) Proteomic Analysis of Mussels Exposed to Fresh and Weathered Prestige’s Oil. J Proteomics Bioinform 2: 255-261.