International Journal of Applied Biotechnology and Biochemistry.
ISSN 2248-9886 Volume 2, Number 4 (2012) pp. 375-379

http://www.ripublication.com/ijabb.htm

Deduce the functional profile from sequence: Role of integrative platform for sequence analysis

L Tadiparthi[1*], Venkata S.P. Dendukuri[1], M.Maheshwara reddy[1]

K L University, Department of Biotechnology, Green fields, Vaddeswaram, Pincode: 522502, Fax: 08645-247249
E-mail: leela_bt@kluniversity.in, pandit@kluniversity.in
Mahesh_bt@kluniversity.in

ABSRTACT:

Functional profiling of sequences is very important to characterise large datasets generated from different sequencing methods. While alignments are accurate for sequences of high similarity, they become unreliable as sequences fall into 'twilight zone'. In such cases alignment based on structure is a better option to explain homology of 'twilight zone' sequences. However presently, majority sequence alignment is done only based on consensus without considering structural information. These circumstances demand development of computational algorithms that can produce alignment considering both consensus and structural information there by explaining functional profile of distinctly related homologs in a better way.

KEY WORDS: Twilight zone, Homologs, Algorithms, Alignment, consensus.

INTRODUCTION:

The post genomic era presents many new challenges in the field of bioinformatics. Massive deliveries of complex and extremely variable datasets are produced by high-throughput experimental technologies (Next generation sequencers). At the end of the 2012 the 1000 Genomes Project (www.1000genomes.org) will be providing sequence information of approximately ~21,000 genomes of different organisms. This flood of complex, heterogeneous and inherently uncharacterized datasets has contributed to the development of novel algorithms, methodologies for organizing and exploring the knowledge from data sets.

Functional profiling of sequences is very important to characterise large datasets generated from different sequencing methods. While, alignments are accurate for

sequences of high sequence similarity, they become unreliable as sequences approach the 'twilight zone'. In twilight zone **(fig-I),** presently available sequence alignment tools fail to explain the existing relationship between structure and sequence homology. Structure alignment is a better option to explain homology for distant pairs (ex: lipoxygenase family protein sequences). This is due to the distantly related homologous sequences with differences in the complex data structure. However presently, sequence alignment is the only choice in the majority of cases where structural information is not available. This situation demands development of computational algorithms that can produce and analyze alignments better and to explain relationship in structural homology in distinctly related sequences

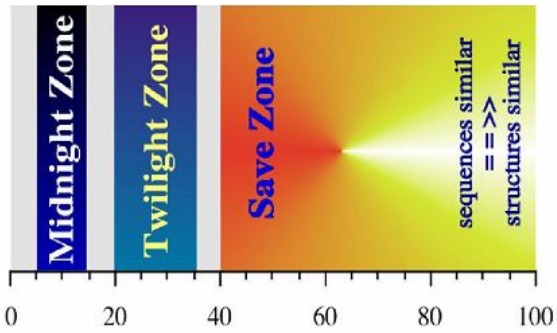

Fig-1: Showing three zones based on the %sequence similarity. Sequence similarity between sequences above 40% is safe zone where sequence similar equals to structural similar but it below 40% the sequence similar may or may equal to structural similar (twilight zone & mid night zone).

RECENT ADVANCES IN SEQUENCE ANALYSIS TOOLS:

Computational biologists throughout the world are actively involved in high end research such as integration of biological data and development of software and models for the evolution of various bio molecules using genome wide analyses [1-2-3-4-5], neuronal simulations and systems biology. Rapid advancement of sequencing technologies resulted in data on an unparalleled scale. A central challenge for the analysis of this data is sequence alignment, whereby sequence reads must be compared to a reference. Globally, over the past two years various alignment algorithms have been subsequently developed. The development of advanced computational strategies to maximally extract significant information from massive nucleotide data has become a major focus of the bioinformatics community.

In recent times, various new tools were developed for the functional analysis of sequences by considering several statistical parameters to reduce the errors in the complex biological sequence data (protein, DNA and RNA). 'Peptide Mine' [6] web server developed by Institute of Microbial Technology Chandigarh (IMTC) for the identification of peptides based on specific functional patterns present in the sequence of an interacting protein. 'CBS-Pred' prediction module [7], predicts carbohydrate-binding sites using single sequence or server-generated PSSM. Several pre-computed

structural and functional properties of complexes were also considered in the database for rapid and accurate data analysis. For the first time a sequence-based approach[8] has been developed to predict NAD binding proteins and their interacting residues, in the absence of any prior structural information which helps in the understanding of NAD+ dependent mechanisms in the cell. 'GWFASTA server' [13] assist the FASTA user to extend the similarity searches against partially and/or completely sequenced genomes by allowing the submission of more than one sequence as a single query for a FASTA search. It also provides integrated post-processing of FASTA output, including compositional analysis of proteins, multiple sequences alignment, and phylogenetic analysis (IMTC). 'ACUA (Automated Codon Usage Tool)' [9] has been developed to perform high throughput sequence analysis aiding statistical profiling of codon usage. The results of ACUA are presented in a spreadsheet with all perquisite codon usage data required for statistical analysis, displayed in a graphical interface. 'IMEx' [10] uses simple string-matching algorithm with sliding window approach to screen DNA sequences for microsatellites and reports the motif, copy number, genomic location, nearby genes, mutational events and many other features useful for in-depth studies. IMEx is more sensitive, efficient and useful than the available widely used tools. IMEx is available in stand-alone program as well as in the form of a web-server. 'Plasmo2D tool' [11] makes use of a Virtual 2-DE generated by plotting all of the proteins from the Plasmodium database on a pI versus molecular weight scale character based similarity cannot provide insight into the structural aspects of a protein. 'The Spectral Similarity Score (SSS)' [12] is an extension program developed on conventional similarity approach that aid in the analysis of various biological sequences and structural variations in proteins.

'Feature Architecture Comparison Tool (FACT)' [13] to search for functionally equivalent proteins. FACT uses the similarity between feature architectures of two proteins. FACT can identify functional equivalents that share no significant sequence similarity. 'proTF' [14] is constructed to serve as a comprehensive data resource and phylogenomics analysis platform for prokaryotic TFs which becomes a valuable resource for prokaryotic transcriptional regulatory network in the post-genomic era. 'BMGE (Block Mapping and Gathering with Entropy)' [15], designed to select regions in a multiple sequence alignment suited for phylogenetic inference. 'GeneOrder4.0' [16] enables the visualization of synteny by plotting protein similarity scores between two genomes along with visual annotation of "hypothetical" proteins from older archived genomes based on more recent annotations. 'iMotifs tool' [17] is a graphical motif analysis environment that allows visualization of annotated sequence motifs and scored motif hits in sequences. It also offers motif inference with the sensitive Nested MICA algorithm, as well as over representation and pair wise motif matching capabilities. PRG platform [18] represents an important starting point to conduct various experimental tasks leading to establishment of a relationship between genomic and phenotypic information. The 'galign' alignment tool [19] uses a simple algorithm to compare parsed sequence reads to parsed reference genome sequences. Ibarcode web based tool is used to identify haplotypes within a species along with species divergences.

However, the tools and resources to analyze present day's noisy data produced form different sequencing technologies are becoming complicated and also are not suitable for working on a large integrated diverge data-sets, in establishing a relationship between functional homology in protein sequences with low percentage of similarity.

CONCLUSION:

Conditions demand the development of an integrative platform for functional analysis of sequences which will be providing the solutions for:

- Determining the function of predicted genes more accurately than other existing function prediction tools.
- Identifying the best homologs for given protein query sequence based on sequence as well as structural similarity (even though the sequence similarity is low).
- In drug discovery research, accurate functional prediction determining new drug targets for various diseases.
- Various vast heterogeneous biological data sets in annotating the important functional role.

REFERENCES:

[1] Ansari HR et al., 2010. "Identification of NAD interacting residues in proteins. BMC Bioinformatics". Mar 30; 11:160.

[2] Bhattacharya A et al., 2000. "Identification of parasitic genes by computational methods. Parasitol Today". Mar; 16(3):127-31.

[3] Bai J etal., 2010." proTF: a comprehensive data and phylogenomics resource for prokaryotic transcription factors". Bioinformatics. Oct 1; 26(19):2493-5. Epub 2010 Jul.

[4] Cantacessi C et al., 2010,"A practical, bioinformatic workflow system for large data sets generated by next-generation sequencing". Nucleic Acids Res. Sep 1;38 (17):e171. Epub 2010 Aug 3.

[5] Criscuolo A., Gribaldo S., 2010. "BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments". BMC Evol Biol. Jul 13; 10:210.

[6] Fernando SA et al., 2004, K "THGS: a web-based database of Transmembrane Helices in Genome Sequences". Nucleic Acids Res. Jan 1;32 (Database issue):D125-8.

[7] Gupta K et al., 2005. "Detailed protein sequence alignment based on Spectral Similarity Score (SSS). BMC Bioinformatics". Apr 23;6:105.

[8] Issac B., Raghava GP., 2007." GWFASTA: server for FASTA search in eukaryotic and microbial genomes". Biotechniques. Sep; 33(3):548-50, 552, 554-6.

[9] Khachane A et al., 2005."Plasmo2D": an ancillary proteomic tool to aid identification of proteins from Plasmodium falciparum". J Proteome Res. Nov-Dec;4(6):2369-74.
[10] Koestler T., von Haeseler A., Ebersberger I., 2010." FACT: functional annotation transfer between proteins with similar feature architectures". BMC Bioinformatics. Aug 9; 11:417.
[11] Mahadevan P., Seto D., 2010 "Rapid pair-wise synteny analysis of large bacterial genomes using web-based GeneOrder4.0". BMC Res Notes. Feb 23; 3:41.
[12] Malik A et al., 2010 "A Database of Known and Modelled Carbohydrate-Binding Protein Structures with Sequence-Based Prediction Tools". Adv Bioinformatics. 436036. Epub 2010 Jun 29.
[13] Mudunuri SB., Nagarajaram HA., 2007." IMEx: Imperfect Microsatellite Extractor. Bioinformatics". May 15; 23(10):1181-7. Epub 2007 Mar 22.
[14] Ondov BD et al., 2010 "An alignment algorithm for bisulfite sequencing using the Applied Biosystems SOLiD System". Bioinformatics. Aug 1; 26(15):1901-2. Epub 2010 Jun 18.
[15] Pellegrini M et al., 2010. "An efficient heuristic for finding fuzzy tandem repeats". Bioinformatics.Jun 15;26 (12):i358-62.
[16] Piipari M et al., 2010." iMotifs: an integrated sequence motif visualization and analysis environment. Bioinformatics". Mar 15; 26(6):843-4. Epub 2010 Jan 26.
[17] Rouault H et al., 2010" Genome-wide identification of cis-regulatory motifs and modules underlying gene coregulation using statistics and phylogeny". Proc Natl Acad Sci U S A. Aug17; 107(33):14615-20. Epub 2010 Jul 29.
[18] Sanseverino W., 2010. "PRGdb: a bioinformatics platform for plant resistance gene analysis. Nucleic Acids Res". Jan; 38(Database issue):D814-21. Epub 2009 Nov 11.
[19] Shaham S., 200. "galign: a tool for rapid genome polymorphism discovery". PLoS One. 2009 Sep 25; 4(9):e7188.
[20] Shameer K et al., 2010. " PeptideMine – A webserver for the design of peptides for protein-peptide binding studies derived from protein-protein interactomes". BMC Bioinformatics. Sep 22; 11: 473.
[21] Singer GA., Hajibabaei M., 2009." iBarcode.org: web-based molecular biodiversity analysis. BMC Bioinformatics". Jun 16; 10 Suppl 6:S14.
[22] Vetrivel U et al., 2007."A software tool for automated codon usage analysis. Bioinformation". Oct 6; 2(2): 62-3.