PHYLOGENETICS AND STRUCTURAL BIOINFORMATICS: A REVIEW ON CURRENT APPLICATIONS OF ALGORITHMS AND TOOLS

Rahul Singh[1], Farzana Parveen[2], Preeti Singh[3]

[1, 2, 3] Assistant Professor, Department of Computer Science, CGVVM, Bhilai (India)

ABSTRACT

Bioinformatics is a field within which we have seen fruitful interplay between several different branches of science, including biology, mathematics and computer science. This paper is an attempt to explore two of its branches, phylogenetics and structural bioinformatics. Phylogenetics is a branch of Bioinformatics that deals with the genealogies which are illustrated as tree like diagrams, known as evolutionary or phylogenetic trees Structural bioinformatics is the branch of bioinformatics which pertains to the analysis and prediction of the 3D structures of large biological molecules such as proteins, DNA etc. This paper presents a review of the various Algorithms, Databases, Frameworks, Methods and Tools pertaining to the two aforementioned fields of Bioinformatics, which have been put forward in the recent past.

Keywords: Bioinformatics, Biology, Mathematics, Phylogenetics, Review, Statistics

I. INTRODUCTION

Bioinformatics is the utilization of computers and IT innovations to the management of huge datasets and collection of information related to the field of biology. Computers are utilized to accumulate, store, break down and incorporate biological and genetic information which can then be applied to gene-based drug discovery and development. The requirement for Bioinformatics capacities has been accelerated by the blast of openly accessible genomic data which is the outcome of the Human Genome Project. Bioinformatics is an interdisciplinary field science, which combines computer science, statistics, mathematics, and engineering which develops methods and software tools to analyze and interpret biological data. Bioinformatics has turned into a vital piece of numerous regions of science. In experimental molecular biology, bioinformatics techniques such as image and signal processing permit extraction of helpful results from a lot of crude information. In the field genetics and genomics, it helps in sequencing and explaining genomes and their witnessed changes. It assumes a part in the content mining of natural writing and the advancement of biological and gene ontologies to arrange and inquiry biological information. It additionally assumes a part in the investigation of gene and protein expression and regulation. Bioinformatics devices help in the correlation of hereditary and genomic information and all the more by and large in the comprehension of evolutionary aspects of molecular biology. At a more integrative level, it helps break down and index the biological pathways and networks that are an important part of systems biology. In structural biology, it helps in the reenactment and displaying of DNA, RNA, and protein structures and additionally molecular associations.

1.1 Phylogenetics

Phylogenetics is the recreation and investigation of phylogenetic (evolutionary) trees and networks taking into account acquired attributes. It is a thriving field confluence of mathematics, statistics, computer science and biology. The primary role of phylogenetic procedures is in evolutionary biology where they are utilized to infer historical relationships between species. However, the methods are also relevant to a diverse range of fields including epidemiology (study of disease transmission), ecology, medicine, as well as linguistics and cognitive psychology [1]. Phylogenetics is the scientific field that deals with describing and recreating the patterns of genetic relationships among species as well as among higher taxa. In simpler words It is the study of evolutionary relationships between groups of organisms.

Phylogenetic trees are a handy and advantageous form of graphical representation of the evolutionary history of life. These diagrams portray the connoted connections between organisms and the order of speciation events that led from earlier common ancestors to their diversified descendants. A phylogenetic tree has several parts. Nodes represent taxonomic units, such as an organism, a species, a population, a common ancestor, or even an entire genus or other higher taxonomic group. Branches connect nodes uniquely and represent genetic relationships. The specific pattern of branching determines the tree's topology. Scaled trees have branch lengths that are proportional to some important biological property, such as the number of amino acid changes between nodes on a protein phylogeny. Trees may also be rooted or unrooted. Rooted trees have a special node, known as the root, that represents a common ancestor of all taxa shown in the tree. Rooted trees are thus directional, since all taxa evolved from the root. Unrooted trees illustrate relationships only, without reference to common ancestors [2]. Below are two examples of phylogenetic trees.

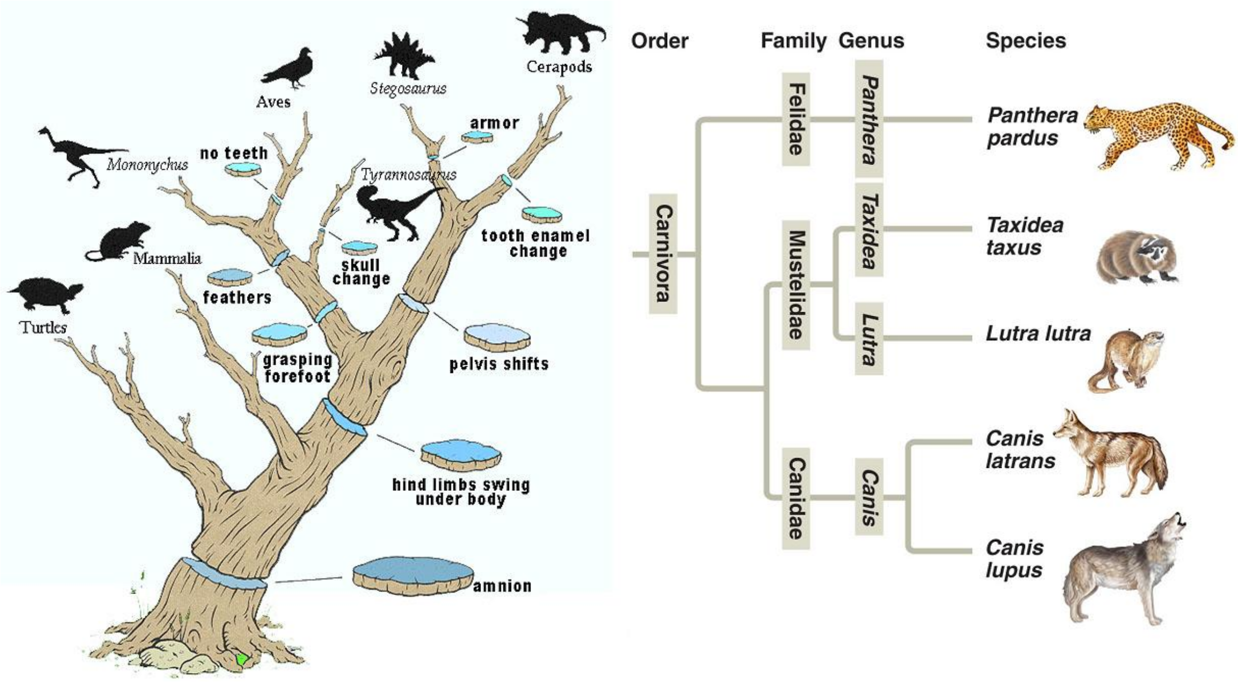

Figure 1: Examples of Phylogenetic Trees

1.2 Structural Bioinformatics

Structural bioinformatics or structural computational biology, extensively characterized, is a field at the convergence between Biology, Chemistry, Physics and Computer Science. Verifiably, the term 'structural bioinformatics' depicts information driven measurable, learning based examination of delegate non-repetitive troupes of structures to comprehend the factual conduct of the framework under scrutiny. Biologically, structural bioinformatics plans to comprehend the elements that impact and focus the capacity of organic macromolecules, the exchange between development, energy and thermodynamics, the determinants of specificity and selectivity in atomic communications, the dynamic parts of macromolecular structures and their impact on capacity and security and, at long last, the capacity to utilize all these for building, outline and biotechnology. Actually, a complete comprehension of organic procedures should unpreventably go through a comprehension of the elements impacting such procedures at the nuclear and some of the time even subatomic levels.

Structural bioinformatics, originally known as structural computational biology, predates other forms of bioinformatics. It can be argued that the seminal 1953 article by Watson and Crick is in fact a modeling paper and arguably the first structural bioinformatics paper [3]. The birth of the field can be credited to how computation was required to accurately refine the tRNA model predicted by Crick in building an actual model that was taller than him. Thus, computation has been an integral part of structural biology from its early days and has had an ever-increasing role in biochemistry and molecular biology with the passing of years [3]. Ilan Samish et al. in their paper titled "Achievements and challenges in structural bioinformatics and computational biophysics" discuss what they consider some of the most noteworthy achievements in structural bioinformatics over the past 10 years and explore the existing challenges in the field which can be tabulated as follows.

Table 1: Achievements and challenges in structural bioinformatics.

Achievements		Challenges	
1.	Data coverage and community resources	1.	Modeling large or multi-domain proteins and assemblies
2.	Computational power	2.	Biomolecules as dynamic objects
3.	Objective method assessment	3.	Modeling 3D RNA structures
4.	Correlated mutations and modeling protein structure	4.	Small differences may have drastic effects
5.	Chemical systems biology	5.	Integration with systems biology
6.	Small-molecule docking simulations	6.	Protein engineering and synthetic biology
		7.	Origins and evolution of protein structure
		8.	Protein folding
		9.	Accessibility and integration of data and methods

II. REVIEW

This section reviews the various algorithms, databases, frameworks and tools that have been put forward in recent past. These have been picked from two categories, namely: Phylogenetics and Structural Bioinformatics.

2.1 Phylogenetics

Analysis of Phylogenetics and Evolution (APE) is a package written in the R language for use in molecular evolution and phylogenetics. The description of APE has been given in a paper by Emmanuel Paradis and two others in 2004. It provides both utility functions for reading and writing data and manipulating phylogenetic trees, as well as several advanced methods for phylogenetic and evolutionary analysis (e.g. comparative and population genetic methods). APE takes advantage of the many R functions for statistics and graphics, and also provides a flexible framework for developing and implementing further statistical methods for the analysis of evolutionary processes [4].

In 2006 Paweł Górecki and Jerzy Tiuryn presented a model of reconciling unrooted gene tree with a rooted species tree, which is based on a concept of choosing rooting which has, minimal reconciliation cost. The analysis leads to the surprising property that all the minimal rootings have identical distributions of gene duplications and gene losses in the species tree. The paper implies that the concept of an optimal rooting is very robust, and thus biologically meaningful. Also, it has nice computational properties. A linear time and space algorithm has been presented for computing optimal rooting(s). This algorithm was used in two different ways to reconstruct the optimal species phylogeny of five known yeast genomes from approximately 4700 gene trees [5].

A paper titled “AWTY (are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics”, presented by Nylander, Wilgenbusch, Warren & Swofford in 2008 states “A key element to a successful Markov chain Monte Carlo (MCMC) inference is the programming and run performance of the Markov chain. However, the explicit use of quality assessments of the MCMC simulations—convergence diagnostics—in phylogenetics is still uncommon” [6]. The paper presents a simple tool that uses the output from MCMC simulations and visualizes a number of properties of primary interest in a Bayesian phylogenetic analysis, such as convergence rates of posterior split probabilities and branch lengths. Graphical exploration of the output from phylogenetic MCMC simulations gives intuitive and often crucial information on the success and reliability of the analysis. The tool presented, complements convergence diagnostics already available in other software packages primarily designed for other applications of MCMC. Importantly, the common practice of using trace-plots of a single parameter or summary statistic, such as the likelihood score of sampled trees, can be misleading for assessing the success of a phylogenetic MCMC simulation.

In 2012, Ponsecu, Huber & Paradis presented a paper describing APE 3.0, a new version of the R package introduced in [4]. The package had grown popular over the time due to its continuously increasing versatility and functionality. The paper discusses APE’s features & components, The one problem was that the data generated with modern genomic approaches can sometimes fail to give rise to sufficiently reliable distance estimates. One way to overcome this problem is to exclude such estimates from data analysis giving rise to an incomplete distance data set (as opposed to a complete one). So far their analysis has been out of reach for ape [7]. To remedy this, the authors have incorporated into APE 3.0, several methods from the literature for phylogenetic inference from incomplete distance matrices. They have also extended ape’s repertoire for phylogenetic inference from complete distances, added a new object class to efficiently encode sets of splits of taxa, and extended the functionality of some of its existing functions.

Phylogenetics, likelihood, evolution and complexity (PLEX) is a flexible and fast Bayesian Markov chain Monte Carlo software program for large-scale analysis of nucleotide and amino acid data using complex evolutionary models in a phylogenetic framework [8]. This application is described in a 2012 paper, by Jason de Koning, Gu, Castoe and Pollock. The program gains large speed improvements over standard approaches by implementing 'partial sampling of substitution histories', a data augmentation approach that can reduce data analysis times from months to minutes on large comparative datasets. A variety of nucleotide and amino acid substitution models are currently implemented, including non-reversible and site-heterogeneous mixture models. Due to efficient algorithms that scale well with data size and model complexity, PLEX can be used to make inferences from hundreds to thousands of sets of data in only minutes on a desktop computer. It also performs probabilistic ancestral sequence reconstruction. Future versions are expected to support detection of co-evolutionary interactions between sites, probabilistic tests of convergent evolution and rigorous testing of evolutionary hypotheses in a Bayesian framework [8].

Jhang J. et al., in a paper titled "A general species delimitation method with applications to phylogenetic placements", analyze current sequence-based methods to delimit species in the field of DNA taxonomy and highlight their drawbacks. They introduce the Poisson tree processes (PTP) model to infer putative species boundaries on a given phylogenetic input tree and integrate PTP with our evolutionary placement algorithm (EPA-PTP) to count the number of species in phylogenetic placements. The authors compare the newly proposed approaches with popular OTU-picking methods and the General Mixed Yule Coalescent (GMYC) model. For de novo species delimitation, the stand-alone PTP model generally outperforms GYMC as well as OTU-picking methods when evolutionary distances between species are small. PTP neither requires an ultrametric input tree nor a sequence similarity threshold as input. In the open reference species delimitation approach, EPA-PTP yields more accurate results than de novo species delimitation methods. Finally, EPA-PTP scales on large datasets because it relies on the parallel implementations of the EPA and RAxML, thereby allowing delimiting species in high-throughput sequencing data [9].

PyRAD is a pipeline to assemble de novo RADseq loci with the aim of optimizing coverage across phylogenetic datasets. It uses a wrapper around an alignment-clustering algorithm, which allows for indel variation within and between samples, as well as for incomplete overlap among reads (e.g. paired-end). This was presented by Deren Eaton in a paper in which he compares PyRAD with the program Stacks in their performance analyzing a simulated RADseq dataset that includes indel variation. Indels disrupt clustering of homologous loci in Stacks but not in PyRAD, such that the latter recovers more shared loci across disparate taxa. The author shows through reanalysis of an empirical RADseq dataset that indels are a common feature of such data, even at shallow phylogenetic scales. PyRAD uses parallel processing as well as an optional hierarchical clustering method, which allows it to rapidly assemble phylogenetic datasets with hundreds of sampled individuals [10].

Sand A. et al., present a library for computing the quartet and triplet distances between binary or general trees called "tqDist". It is a software package for computing the triplet and quartet distances between general rooted or unrooted trees, respectively. The program is based on algorithms with running time O (n log n) for the triplet distance calculation and O (d . n log n) for the quartet distance calculation, where n is the number of leaves in the trees and d is the degree of the tree with minimum degree. These are currently the fastest algorithms both in theory and in practice [11].

A paper titled “Efficient Bayesian inference under the structured coalescent”, was presented by Vaughan T.G. et al., in which they present a new MCMC sampler capable of sampling from posterior distributions over structured trees: timed phylogenetic trees in which lineages are associated with the distinct subpopulation in which they lie. The sampler includes a set of MCMC proposal functions that offer significant mixing improvements over a previously published method. Furthermore, its implementation as a BEAST 2 package ensures maximum flexibility with respect to model and prior specification. The authors demonstrate the usefulness of this new sampler by using it to infer migration rates and effective population sizes of H3N2 influenza between New Zealand, New York and Hong Kong from publicly available hemagglutinin (HA) gene sequences under the structured coalescent [12].

Weyenberg G. et al., focus on non-parametric estimation of phylogenetic tree distributions, the authors explain why such gene trees are considered to be biologically interesting as numerous other coexisting processes cause some genes to exhibit a history distinct from those of the majority of genes and the authors then propose and implement KDETREES, a non-parametric method for estimating distributions of phylogenetic trees, with the goal of identifying trees that are significantly different from the rest of the trees in the sample. The method compares favorably with a similar recently published method, featuring an improvement of one polynomial order of computational complexity (to quadratic in the number of trees analyzed), with simulation studies suggesting only a small penalty to classification accuracy. Application of KDETREES to a set of Apicomplexa genes identified several unreliable sequence alignments that had escaped previous detection, as well as a gene independently reported as a possible case of horizontal gene transfer. The authors also analyze a set of Epichloë genes, fungi symbiotic with grasses, successfully identifying a contrived instance of paralogy [13].

Phylogenetic estimates from published studies can be archived using general platforms like Dryad, TreeBASE or other similar alternatives. Such services fulfill a crucial role in ensuring transparency and reproducibility in phylogenetic research [14]. In any case, computerized tree information documents regularly oblige some altering (e.g. rerooting) to enhance the precision and reusability of the phylogenetic explanations. Besides, setting up the mapping between tip names utilized as a part of a tree and taxa in a solitary regular scientific classification significantly enhances the capacity of different specialists to reuse phylogenetic appraisals. As the procedure of curating a distributed phylogenetic appraisal is not slip free, holding a full record of the provenance of alters to a tree is significant for openness, permitting editors to get kudos for their work and making lapses acquainted amid curation less demanding with right. This paragraph can be considered as the essence of a paper titled “Phylesystem: a git-based data store for community-curated phylogenetic estimates”, presented by E.J. mcTavish et al. The authors report the advancement of programming framework to bolster the open curation of phylogenetic information by the group of scholars. The backend of the framework gives an interface to the standard database operations of making so as to make, perusing, upgrading and erasing records focuses on a git vault. The history's record of alters to a tree is safeguarded by git's variant control highlights. Facilitating this information store on GitHub (http://github.com/) gives open access to the information store utilizing apparatuses well known to numerous engineers. We have sent a server running the 'phylesystem-programming interface', which wraps the associations with git and GitHub. The Open Tree of Life task has additionally created and sent a JavaScript application that uses the phylesystem-programming interface and other web administrations to empower info and curation of distributed phylogenetic articulations.

In Ref. [15] Massingham & Goldman present a paper titled “EDIBLE: experimental design and information calculations in phylogenetics”. This is a comparatively old paper from the year 2000, which looks at the issue of evolutionary induction from molecular sequences. Albeit this is an issue of statistics, little consideration has been paid to inquiries of experimental design. An application, EDIBLE has been produced to perform likelihood calculations in light of Markov procedure models of nucleotide substitution unified with phylogenetic trees, and from these compute Fisher data measures under diverse trial scenarios. These computations can be utilized to answer inquiries of ideal exploratory outline in molecular phylogenetics.

Phylogenetic Analysis Library (PAL) is a collection of Java classes for utilization in molecular evolution and phylogenetics. This has been introduced by Drummond & Strimmer in [16], which is also a fourteen year old paper from 2001. It provides a modular environment for the rapid construction of both special purpose & general analysis programs. PAL version 1.1 consists of 145 public classes or interfaces in 13 packages, including classes for models of character evolution, maximum liklihood estimation, and the coalecent, with a total of more than 27000 lines of code. The PAL venture is set up as a synergistic undertaking to encourage commitments from different specialists.

2.2 Structural Bioinformatics

Dixit & Beveridge in [17], introduce an online tool (MDDNA) to study and model the fine auxiliary subtle elements of DNA on the premise of information separated from an arrangement of atomic motion (MD) directions of DNA successions including all the novel tetranucleotides. The dynamic web interface can be utilized to break down the first neighbor sequence setting impacts on the 10 exceptional dinucleotide steps of DNA. Usefulness is incorporated to assemble all particle models of any client characterized sequence in view of the MD results. The backend of this interface is a relational database which stores the conformational subtle elements of DNA got in 39 diverse MD reenactment directions involving all the 136 one of a kind tetranucleotide steps. Samples of the utilization of this information to anticipate DNA structures are incorporated.

Grünberg, Nilges & Lecknerpresent “Biskit”, which is a modular, object-oriented python library that gives instinctive classes to numerous run of the mill assignments of structural bioinformatics research. It facilitates the manipulation and analysis of macromolecular structures, protein complexes and molecular dynamics trajectories. At the same time, Biskit offers a software platform for the rapid integration of external programs and new algorithms into complex structural bioinformatics workflows [18]. Calculations are along these lines frequently designated to built up projects like Xplor, Amber, Hex, Prosa, Hmmer and Modeler; interfaces to further programming can be effortlessly included. Additionally, Biskit rearranges the parallelization of tedious counts by means of PVM (Parallel Virtual Machine).

A paper titled “RNAfbinv: an interactive Java application for fragment-based design of RNA sequences” has been presented by Weinbrand, Avihoo & Barash [19]. The paper focuses on RNA outline issues, where it is conceivable to accept that the client would be occupied with protecting a specific RNA optional structure theme, or section, for organic reasons. The protection could be in structure or arrangement, or both. Consequently, the backwards RNA collapsing issue could profit by considering section requirements. The authors have developed an intelligent Java application RNA fragment-based inverse that allows users to insert an RNA secondary structure in dot-bracket notation. It then performs grouping plan that fits in with the information's state auxiliary structure, the predefined thermodynamic soundness, the predetermined mutational power and the client chose section after shape deterioration. In this shape based design approach, specific RNA structural motifs with known biological functions are strictly enforced, while others can have more adaptability in their structure for protecting physical traits and extra requirements.

Pooya Zakeri et al. look at various approaches for the prediction of protein folds which are based on features extracted from protein sequences & machine leaning. Looking for an efficient technique through fusion of multiple kernels they highlight the limitations of linear combinations. The authors design several techniques to combine kernel matrices by taking more involved, geometry inspired means of these matrices instead of convex linear combinations, consider various sequence based protein features including information extracted directly from position-specific scoring matrices and local sequence alignment. The methods for classification on the SCOP PDB-40D benchmark dataset for protein fold recognition were evaluated. The best overall accuracy on the protein fold recognition test set obtained by the presented methods is approximately 86.7%. This is an improvement over the results of the best existing approach. Moreover, the computational model has been developed by incorporating the functional domain composition of proteins through a hybridization model. It is observed that by using the proposed hybridization model, the protein fold recognition accuracy is further improved to 89.30%. Furthermore, they investigate the performance of our approach on the protein remote homology detection problem by fusing multiple string kernes [20].

A paper titled, "Basin Hopping Graph: a computational framework to characterize RNA folding landscapes" was presented by Marcel Kucharík et al. RNA folding is a complicated kinetic process. The minimum free energy structure provides only a static view of the most stable conformational state of the system. It is insufficient to give detailed insights into the dynamic behavior of RNAs. A sufficiently sophisticated analysis of the folding free energy landscape, however, can provide the relevant information. Results: We introduce the Basin Hopping Graph (BHG) as a novel coarse-grained model of folding landscapes. Each vertex of the BHG is a local minimum, which represents the corresponding basin in the landscape. Its edges connect basins when the direct transitions between them are 'energetically favorable'. Edge weights endcode the corresponding saddle heights and thus measure the difficulties of these favorable transitions. BHGs can be approximated accurately and efficiently for RNA molecules well beyond the length range accessible to enumerative algorithms. The authors introduce the Basin Hopping Graph (BHG) as a novel coarse-grained model of folding landscapes. Each vertex of the BHG is a local minimum, which represents the corresponding basin in the landscape. Its edges connect basins when the direct transitions between them are 'energetically favorable'. Edge weights endcode the corresponding saddle heights and thus measure the difficulties of these favorable transitions. BHGs can be approximated accurately and efficiently for RNA molecules well beyond the length range accessible to enumerative algorithms [21].

The chemical structures of biomolecules, whether naturally occurring or synthetic, are composed of functionally important building blocks. Given a set of small molecules—for example, those known to bind a particular protein—computationally decomposing them into chemically meaningful fragments can help elucidate their functional properties, and may be useful for designing novel compounds with similar properties. Dario Ghersi and Mona Singh, in [22], introduce molBLOCKS, a suite of programs for breaking down sets of small molecules into fragments according to a predefined set of chemical rules, clustering the resulting fragments, and uncovering statistically enriched fragments. Among other applications, our software is supposed to be a great aid in large-scale chemical analysis of ligands binding specific targets of interest.

Antibodies are currently the most important class of biopharmaceuticals. Development of antibody-based drugs depends on costly and time-consuming screening campaigns. Computational techniques such as antibody–antigen docking hold the potential to facilitate the screening process by rapidly providing a list of initial poses that approximate the native complex [23]. Konrad Krawczyk et al. have developed a new method to identify the epitope region on the antigen, given the structures of the antibody and the antigen—EpiPred. The method combines conformational matching of the antibody–antigen structures and a specific antibody–antigen score. The method has been tested on both a large non-redundant set of antibody–antigen complexes and on homology models of the antibodies and/or the unbound antigen structure. On a non-redundanttest set, the epitope prediction method achieves 44% recall at 14% precision against 23% recall at 14% precision for a background random distribution. The epitope predictions are utilized to rescore the global docking results of two rigid-body docking algorithms: ZDOCK and ClusPro. In both cases including the epitope, prediction increases the number of near-native poses found among the top decoys [23].

Correlations between sequence evolution and structural dynamics are of utmost importance in understanding the molecular mechanisms of function and their evolution. Ahmet Bakan et al. have integrated Evol, a new package for fast and efficient comparative analysis of evolutionary patterns and conformational dynamics, into ProDy, a computational toolbox designed for inferring protein dynamics from experimental and theoretical data. Using information-theoretic approaches, Evol coanalyzes conservation and coevolution profiles extracted from multiple sequence alignments of protein families with their inferred dynamics [24].

Energy landscapes provide a valuable means for studying the folding dynamics of short RNA molecules in detail by modeling all possible structures and their transitions. Higher abstraction levels based on a macro-state decomposition of the landscape enable the study of larger systems; however, they are still restricted by huge memory requirements of exact approaches [25]. In an original paper titled "Memory-efficient RNA energy landscape exploration", Martin Mann et al. present a highly parallelizable local enumeration scheme that enables the computation of exact macro-state transition models with highly reduced memory requirements. The approach is evaluated on RNA secondary structure landscapes using a gradient basin definition for macro-states. Furthermore, the authors demonstrate the need for exact transition models by comparing two barrier-based approaches, and perform a detailed investigation of gradient basins in RNA energy landscapes [25].

The transport of ligands, ions or solvent molecules into proteins with buried binding sites or through the membrane is enabled by protein tunnels and channels. Barbora Kozlikova and 14 others present “CAVER Analyst”, which is a software tool for calculation, analysis and real-time visualization of access tunnels and channels in static and dynamic protein structures. It provides an intuitive graphic user interface for setting up the calculation and interactive exploration of identified tunnels/channels and their characteristics [26].

Reference [27] is a paper that introduces Cloud4Psi, which is a cloud based tool for searching 3D protein structure similarities. Popular methods for 3D protein structure similarity searching, especially those that generate high-quality alignments such as Combinatorial Extension (CE) and Flexible structure Alignment by Chaining Aligned fragment pairs allowing Twists (FATCAT) are still time consuming. As a consequence, performing similarity searching against large repositories of structural data requires increased computational resources that are not always available. Cloud computing provides huge amounts of computational power that can be provisioned on a pay-as-you-go basis. Dariusz Mrozek and others have developed the cloud based system that allows scaling of the similarity searching process vertically and horizontally. Cloud4Psi (Cloud for Protein Similarity) was tested in the Microsoft Azure cloud environment and provided good, almost linearly proportional acceleration when scaled out onto many computational units.

Miguel Vázquez, Alfonso Valencia, and Tirso Pons presented a paper titled "Structure-PPi: a module for the annotation of cancer-related single-nucleotide variants at protein–protein interfaces" in March 2015. The understanding of cancer related single-nucleotide variations (SNVs) considering the protein highlights they influence, for example, known functional sites, protein–protein interfaces, or connection with officially clarified changes, may supplement the annotation of hereditary variations in the investigation of NGS information. Current apparatuses that explain changes miss the mark on a few viewpoints, including the capacity to utilize protein structure data or the understanding of transformations in protein buildings. The authors introduce the Structure–PPi framework for the extensive examination of coding SNVs in view of 3D protein structures of protein complexes. The 3D store utilized, Interactome3D, incorporates trial and demonstrated structures for proteins and protein–protein edifices. Structure–PPi comments SNVs with components extricated from UniProt, InterPro, APPRIS, dbNSFP and COSMIC databases. The authors show the handiness of Structure–PPi with the understanding of 1027122 non-synonymous SNVs from COSMIC and the 1000G Project that gives an accumulation of approximately 172700 SNVs mapped onto the protein 3D structure of 8726 human proteins (43.2% of the 20 214 SwissProt-curated proteins in UniProtKB release 2014_06) and protein–protein interfaces with potential functional implications [28].

III. CONCLUSION

Bioinformatics is a relatively new field of studies which has taken form after the merger of various scientific streams such as biology, mathematics, statistics and computer science. Bioinformatics is a vast field of research and consists of many branches such as Databases & Ontologies, Genome Analysis, Phylogenetics, Sequence Analysis and Text Mining etc. It is flourishing with time & has gained significant popularity due to the successful endeavors like the Human Genome Project. More and more scholars are pursuing the field & proposing new algorithms, methods, techniques and tools to advance the field. This paper focuses on Phylogenetics and Structural Bioinformatics, and aims at reviewing many of the newly proposed works pertaining to them. There are hundreds of new applications & theories that are put forward every year but not all of them can be included in one paper lest it becomes too large to confine to the terms of the conference. More than two dozens of recently published papers have been reviewed in an attempt to acquaint the authors as well as the readers to these eminent and enchanting branches of bioinformatics.

REFERENCES

[1] Charles Semple and Mike Steel, *Phylogenetics* (Oxford University Press, New York, 2003).

[2] Douglas Theobald, "29+ Evidences for Macroevolution: Phylogenetics Primer", http://www.talkorigins.org/faqs/comdesc/phylo.html, accessed September 2015.

[3] Ilan Samish, Philip E. Bourne and Rafael J. Najmanovich, "Achievements and challenges in structural bioinformatics and computational biophysics", Bioinformatics (2014) 31 (1): 146-150, doi:10.1093/bioinformatics/btu769

[4] Paradis E. et al., “APE: Analyses of Phylogenetics and Evolution in R language”, OUP journal of Bioinformatics (2004) Vol. 20 no. 2, pages 289–290, doi: 10.1093/bioinformatics/btg412

[5] Paweł Górecki and Jerzy Tiuryn, "Inferring phylogeny from whole genomes", OUP journal of Bioinformatics, Vol. 23 ECCB (2006), pages e116–e122, doi:10.1093/bioinformatics/btl296

[6] Nylander J.A.A. et al., "AWTY (are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics", Bioinformatics (2008) 24 (4): 581-583 first published online August 30, 2007 doi:10.1093/bioinformatics/btm388

[7] Popescu A. et al., “APE 3.0: New tools for distance-based phylogenetics and evolutionary analysis in R”, Vol. 28 no. 11 2012, pages 1536–1537, doi:10.1093/bioinformatics/bts184

[8] Jason de Koning A.P. et al., "Phylogenetics, likelihood, evolution and complexity", Bioinformatics (2012) Vol. 28 no. 22 , pages 2989–2990, doi:10.1093/bioinformatics/bts555

[9] Jiajie Zhang, Paschalia Kapli, Pavlos Pavlidis, and Alexandros Stamatakis, "A general species delimitation method with applications to phylogenetic placements", Bioinformatics (2013) 29 (22): 2869-2876, doi:10.1093/bioinformatics/btt499

[10] Deren A. R. Eaton, "PyRAD: assembly of de novo RADseq loci for phylogenetic analyses", Bioinformatics (2014) 30 (13): 1844-1849, doi:10.1093/bioinformatics/btu121

[11] Andreas Sand, Morten K. Holt, Jens Johansen, Gerth Stølting Brodal, Thomas Mailund, and Christian N. S. Pedersen, "tqDist: a library for computing the quartet and triplet distances between binary or general trees", Bioinformatics (2014) 30 (14): 2079-2080, doi:10.1093/bioinformatics/btu157

[12] Timothy G. Vaughan, Denise Kühnert, Alex Popinga, David Welch, and Alexei J. Drummond, "Efficient Bayesian inference under the structured coalescent", Bioinformatics (2014) 30 (16): 2272-2279, doi:10.1093/bioinformatics/btu201

[13] Grady Weyenberg, Peter M. Huggins, Christopher L. Schardl, Daniel K. Howe, and Ruriko Yoshida, "KDETREES: non-parametric estimation of phylogenetic tree distributions", Bioinformatics (2014) 30 (16): 2280-2287, doi:10.1093/bioinformatics/btu258

[14] Emily Jane McTavish, Cody E. Hinchliff, James F. Allman, Joseph W. Brown, Karen A. Cranston, Mark T. Holder, Jonathan A. Rees, and Stephen A. Smith "Phylesystem: a git-based data store for community-curated phylogenetic estimates", Bioinformatics (2015) 31 (17): 2794-2800, doi:10.1093/bioinformatics/btv276

[15] Massingham T., and Goldman N., "EDIBLE: experimental design and information calculations in phylogenetics, OUP Journal of Bioinformatics (2000) Vol. 16 no. 03, pg:294-295

[16] Drummond A. and Strimmer K., "PAL: an object-oriented programming library for molecular evolution and phylogenetics", OUP Journal of Bioinformatics (2001) Vol.17 no.7, pages: 662-663

[17] Surjit B. Dixit and David L. Beveridge, "Structural bioinformatics of DNA: a web-based tool for the analysis of molecular dynamics results and structure prediction", Bioinformatics (2006) 22 (8): 1007-1009, doi:10.1093/bioinformatics/btl059

[18] Raik Grünberg, Michael Nilges, and Johan Leckner, "Biskit—A software platform for structural bioinformatics", Bioinformatics (2007) 23 (6): 769-770, doi:10.1093/bioinformatics/btl655

[19] Lina Weinbrand, Assaf Avihoo, and Danny Barash, "RNAfbinv: an interactive Java application for fragment-based design of RNA sequences", Bioinformatics (2013) 29 (22): 2938-2940, doi:10.1093/bioinformatics/btt494

[20] Pooya Zakeri, Ben Jeuris, Raf Vandebril, and Yves Moreau, "Protein fold recognition using geometric kernel data fusion", Bioinformatics (2014) 30 (13): 1850-1857, doi:10.1093/bioinformatics/btu118

[21] Marcel Kucharík, Ivo L. Hofacker, Peter F. Stadler, and Jing Qin, "Basin Hopping Graph: a computational framework to characterize RNA folding landscapes", Bioinformatics (2014) 30 (14): 2009-2017, doi:10.1093/bioinformatics/btu156

[22] Dario Ghersi and Mona Singh, "molBLOCKS: decomposing small molecule sets and uncovering enriched fragments", Bioinformatics (2014) 30 (14): 2081-2083, doi:10.1093/bioinformatics/btu173

[23] Konrad Krawczyk, Xiaofeng Liu, Terry Baker, Jiye Shi, and Charlotte M. Deane, "Improving B-cell epitope prediction and its application to global antibody-antigen docking", Bioinformatics (2014) 30 (16): 2288-2294, doi:10.1093/bioinformatics/btu190

[24] Ahmet Bakan, Anindita Dutta, Wenzhi Mao, Ying Liu, Chakra Chennubhotla, Timothy R. Lezon, and Ivet Bahar, "Evol and ProDy for bridging protein sequence evolution and structural dynamics", Bioinformatics (2014) 30 (18): 2681-2683, doi:10.1093/bioinformatics/btu336

[25] Martin Mann, Marcel Kucharík, Christoph Flamm, and Michael T. Wolfinger, "Memory-efficient RNA energy landscape exploration", Bioinformatics (2014) 30 (18): 2584-2591, doi:10.1093/bioinformatics/btu337

[26] Kozlikova B. et al., "CAVER Analyst 1.0: graphic tool for interactive visualization and analysis of tunnels and channels in protein structures", Bioinformatics (2014) 30 (18): 2684-2685, doi:10.1093/bioinformatics/btu364

[27] Dariusz Mrozek, Bożena Małysiak-Mrozek, and Artur Kłapciński, "Cloud4Psi: cloud computing for 3D protein structure similarity searching", Bioinformatics (2014) 30 (19): 2822-2825, doi:10.1093/bioinformatics/btu389

[28] Miguel Vázquez, Alfonso Valencia, and Tirso Pons, "Structure-PPi: a module for the annotation of cancer-related single-nucleotide variants at protein–protein interfaces", Bioinformatics (2015) 31 (14): 2397-2399, doi:10.1093/bioinformatics/btv142