Bilateral conception strategy for efficient design of Lattice models and well Structured Genomic data processing with the semantic web-based system

Narahari Narasimhaiah[1], Dr. R.PRAVEEN SAM [2]
[1] Research Scholar, Research and Development Center, Bharathiar University, Coimbatore, India.narasimhaiah.narahari@gmail.com
[2]Professor, Department of CSE, G.Pulla Reddy Engineering College, Kurnool, India.rpraveensam.cse@gprec.ac.in

Abstract
Protein folding design and genomic data processing methods are more powerful tools to investigate the fundamental principle of IoT based healthcare. In previous researches, protein folding has the problem of finding the lowest energy conformation and size complexity. Along with this, there is a security problem in genome sequence and complications in storing genome data due to the unstructured format. To tackle these issues a novel Bilateral conception strategy has been proposed in this paper. Initially, the problems of finding the lowest energy conformation and size complexity have been solved by a novel Reconcile pirouette mechanism-based quantum processor through decrease the quantum circuit from quadratic to quasi-linear. Then, the novel Adumbrating algorithm for secure commune based genomic data analysis deals with the security issue and provides the highest privacy for the genome data by using Single Nucleotide Polymorphisms. Finally, a novel Adroit Semantic web-based system is proposed for securely storing and retrieving genome data by Collateral Description frame format. Thus the novel bilateral conception strategy effectively solves the problem in protein folding and provides a high security and storage format for genome sequence.
Keywords: Protein folding, Genome sequence, Genomic data analysis, Single Nucleotide Polymorphism (SNP).

1 Introduction

The protein structure gathers essential knowledge about its biological activity and its therapeutic potential [1]. Protein structure knowledge unlocks useful biological information, ranging from the ability to predict protein-protein interactions [2] to the discovery of new drugs [3] and catalysts [4] based on structure. Unfortunately, studies are difficult and take considerable time and energy to determine protein structure [5, 6]. The TrEMBL database [7] contained 176 million protein sequences as of February 2020, while only 160,000 protein structures were deposited in the Protein Data Bank [8]. A reliable computational algorithm for the template-free prediction of the structure of a protein and its folding pathway from sequence data alone would allow millions of proteins to be annotated and could stimulate significant advances in biological science. However, a reliable and precise protein folding algorithm has remained elusive, despite steady advances in the past six decades [9-11]. There have been attempts to use quantum computing for protein structure prediction over the past decade. It is assumed that the biological

structure of a protein corresponds to the minimum of a free energy hypersurface, which is too large for any classical machine to investigate exhaustively for even tiny peptides [12].

In that Adiabatic quantum computing (AQC), an approach to leveraging the physics of a controlled quantum system that is considered to be of potential use in optimization issues (whether classical or quantum in nature) [13], is a form of quantum computation that may be appropriate to support. In order historically, BioCreative challenges have aimed to bring forward group tasks that contribute to the creation of text-mining systems that can be of practical benefit to database curators and users of biology textual data. Project selection involved the recognition in biomedical literature of biologically related organisms, such as genes, proteins, animals, diseases, and chemicals, as well as their interactions. These tasks have investigated important factors in usability and understanding of duration workflows [14], have concentrated on developing text mining systems that meet the requirements of users, and encourage standard developments for use, reuse, and integration issues [15]. In addition, the development of knowledge extraction systems for specific and emerging research areas has been addressed by group challenges in biomedical natural language processing, such as BioNLP and BioASQ [16], in line with current requirements.

Specifics on genetic mutations that decide the appropriateness of a cancer patient being given a certain drug/therapy can be included in the eligibility requirements of a clinical trial protocol. This knowledge on genetic mutations and the associated background has been considered an important source of evidence for compiling pharmacogenomics (PGx) [17]. Unfortunately, this data occurs in the eligibility requirement for the free-text format, which is difficult to parse using traditional text mining techniques. [18] Different methods have been explored in previous studies to capture and structure information in clinical trials, but most systems do not rely on capturing mutations specified in the eligibility criteria. A dictionary-oriented approach [19] to the identification of genes, drugs, and diseases, for example, is based on the fields of condition, action, and research definition. Machine learning strategies are based on eligibility requirements to classify genes and their categorical status (mutated or not), but they do not identify greater structural variations or unique variants. To provide efficient filtering of trials and promote the search for trials, eTACTS3 mines frequently [20] occurring tags from the free-text eligibility criteria. Since this method only retains the commonly occurring tags for high-level definitions, less repeated mention of mutations is likely to be overlooked. Other articles concentrate on collecting references to mutations from biomedical literature, but not from clinical research.

Despite the fact that from the above-mentioned works of literature and existing works it is clearly indicated that there is no work that has focused the major significance on protein folding and genomic data analysis on the basis of finding the lowest energy conformation of folding protein secondary structure elements in three dimensions with compact size, reduction of data breaches in collaborative genomic information analysis, structured format of gene mutation annotations and simplest manner to search for relevant trials. Hence, to tackle those issues, this paper develops a new strategy in the promising field of IoT based healthcare.

The contribution of the novel Bilateral Conception Strategy is,

- Solves the problem of finding a lattice protein's lowest energy conformation and reduce the size complexity.
- Provides protection to genomic sequence via SNPs.
- Provides a scalable framework for standard-based data representation, integration, and sharing also structured format for storing such data.
- The performance of the proposed work has been analyzed through MATLAB.

The structure of the paper as follows: Sect.2 discusses the related researches with protein folding and genome sequence analysis. Sect.3 describes the proposed methodologies. Sect.4 comprises the results and comparison of proposed methods. Sect.5 concludes this paper.

2 Literature Survey:

Cao et al [21]For decades, the topic of protein folding had been studied extensively, and hundreds of thousands of protein structures had been solved. Yet it is not fully known how proteins fold from a linear peptide chain to their special 3D structures. A "Confined Lowest Energy Fragment" (CLEF) hypothesis was suggested with key clues having emerged unexpectedly from the field of nanoscience. The CLEF hypothesis noted that a protein chain can be separated by a small number of main residues that shape key long-range interactions in CLEFs, the semi-independent folding units. Under the limitations of the main long-range interactions, the native structure of a CLEF is the lowest energy state, but the native structure of the whole protein does not need the lowest energy state, as indicated by Anfinsen's thermodynamic hypothesis. The CLEF hypothesis, essentially a two-step method, proposes a single CLEF mechanism for protein folding.In the first step, the positive enthalpy of CLEFs easily brought together certain residues for the main long-range interactions for native structures, forming intermediates corresponding to the so-called hydrophobic collapse. In the second step, to shape the native key long-range interactions, those collapsed key residues shuffle for the right combination. The CLEF hypothesis offered a simple solution to all paradoxes of protein folding and presented a "CLEF Age" or "Stone Age" for protein prebiotic growth.

Li et al [22] The consistency and reliability of health-related decisions taken by physicians in modern medicine had become an important feature of the accuracy of a prognostic prediction model. Unfortunately, adequate samples are often insufficient for individual entities. One mitigation is to spread data collection to several centers from a single organization to collectively increase the sample size. Confidential biomedical knowledge for research was shared, however, complicated problems were entailed. In multicenter privacy-preserving data mining scenarios, machine learning models such as random forests (RF), while they are widely used and achieved good performance for prognostic prediction; typically suffered worse performance compared to a centrally trained version. A multicenter random forest prognosis prediction model that allows federated clinical data mining from horizontally partitioned datasets is proposed in this report.

Xiang et al [23] The feed conversion ratio (FCR) is a significant efficient feature that has a major effect on pig industry income. Elucidating the FCR's genetic mechanisms could promoted the effectiveness of improving FCR through artificial selection. In this paper, a genome-wide association study (GWAS) was combined with transcriptome analysis of Yorkshire pigs (YY) in different tissues to classify key genes and signaling pathways significantly associated with FCR.GWAS had observed a total of 61 important single nucleotide polymorphisms (SNPs) in YY. All of these SNPs are located on the porcine chromosome (SSC) 5 and the quantitative trait locus (QTL) region for FCR was considered to be the protected region. Some genes distributed around these essential SNPs, including TPH2, FAR2, IRAK3, YARS2, GRIP1, FRS2, CNOT2, and TRHDE, had been considered as candidates for regulating FCR. TPH2 had the ability to control intestinal motility through a serotonergic synapse and an oxytocin signaling pathway, according to transcriptome research in the hypothalamus. Furthermore, GRIP1 is involved in a signaling pathway for glutamatergic and GABAergic, which controls FCR by influencing the appetite in pigs. In addition, through a thyroid hormone signaling pathway, GRIP1, FRS2, CNOT2, TRHDE controls the metabolism in different tissues.

Trębacz et al [24] Stratifying cancer patients based on their levels of gene expression helps diagnosis, prediction of survival, and preparation of treatment to be enhanced. Such data, however, is extremely highly dimensional because it contained expression values for more than 20000 genes per patient, and there is a low number of samples in the datasets. In order to deal with these settings, the paper proposed to integrate prior biological knowledge of ontological genes into the machine learning system for the task of classifying patients provided their data on gene expression. In order to guide a Graph Convolutional Network, the ontology embeddings were used where the semantic similarities were captured between the genes and thus sparsified the links of the network. The research demonstrated that this method offers an advantage of high-dimensional low-sample data for predicting clinical targets.

Quan et al [25]Genetics is ideally committed to the understanding and unveiling of pathogenesis and gene functions. The last decade had seen unprecedented progress in genetics, especially through Genome-Wide Association Studies (GWAS) and Phenome-Wide Association Studies in the genome-wide identification of disorder variants (PheWAS). Nevertheless, it is still a major challenge to use GWAS/PheWAS-derived data to elucidate pathogenesis. In this research, HotNet2, a genetic algorithm focused on heat diffusion systems, was used to measure the networks of disease genes obtained from GWAS and PheWAS, in an attempt to gain deeper insights at the molecular level into disease pathogenesis. At the level of biological networks, the system genetics algorithm HotNet2 can create genotype-phenotype ties effectively.HotNet2-calculated disease-gene associations with greater biomedical importance compared to original GWAS/PheWAS outcomes, thereby providing improved interpretations of genome-wide variant pathogenesis, and also offering new insights into gene functions.

In [21] conquer more intricacy for folding longer protein sequences [22] complicated size and difficulty to find the lowest energy conformation [23] exchanging genomic sequence information in a very abundant manner typically not feasible in genomic medicine remains a

difficult task. [24] it should evaluate the method on clinical targets derived from heterogeneous data sources [25] Systems lack the capability to provide sufficient mutation annotations, and it has to identify the disease networks more in line with biological realitythough there is enormous development is essential in the protein folding design and genomic data analysis in IoT based healthcare field.

3 Bilateral Conception strategy

As genome sequencing technologies and protein folding is a most essential biological process in the health care field. Over the past thirty years, lattice models have been used extensively to explore the concepts of protein folding and design. From a large number of potential conformations, these models can be used to evaluate the conformation of the lowest energy fold. However, prior researches has conquer more intricacy for folding longer protein sequences due to the complicated size and difficulty to find the lowest energy conformation of the relative location and orientation of protein secondary structure elements in three dimensions. In addition, one of the most competent technology as genome sequencing/ gene data processing, existing collaborative genomic data analysis processes enable all people concerned to share individual patient data and conduct all analysis locally or use a trusted server to hold all data for analysis on a single site (e.g. the Cancer Genome Collaboratory). Since both methods include exchanging genomic sequence information in a very abundant manner, which is typically not feasible due to data breaches, in that collaborative data analysis in genomic medicine remains a controversial topic. Along with this, research on genetic mutations and the associated background has been considered an important source of evidence for compiling pharmacogenomics (PGx). Unfortunately, this data occurs in a compliance requirement for the free-text format, which is difficult to parse using traditional text mining techniques. Thus, most of the systems lack the capability to provide sufficient mutation annotations for clinical trials, which makes it hard to capture and store due to unstructured format. Consequently, it difficult to search for relevant trials based on patients' mutation information. From the aforementioned concerns, this paper creates a novel strategy to tackle these issues in an emerging field of healthcare.

Fig.1 shows the block diagram of the proposed methodology. This paper proposed the novel Bilateral Conception strategy to efficiently tackles the above-mentioned issues with efficiently accomplished the immaculate outcomes of protein folding design and genomic data analysis. Initially, the work introduces the Reconcile pirouette mechanism-based quantum processor for the problem of finding a lattice protein's lowest energy conformation. Thus it performs through decreases in the quantum circuit from quadratic to quasi-linear function. Furthermore, in order to obtain findings with greater correlation to the real atomistic 3D structure of the protein, the paper generalizes to three spatial dimensions and a heuristic strategy for breaking large problem instances into smaller subproblems of protein structure, thus it reduced the size complexity. Subsequently, the work incorporates the Adumbrating algorithm for secure commune based genomic data analysis for collaborative or remote genomic computation in an

efficient and effective protected manner. That accomplished via recognizing the top small value of Single Nucleotide Polymorphisms (SNPs), which can be predetermined or can be defined during query processing. That small value is identified amongst a user-specified subset of SNPs across case/control samples.

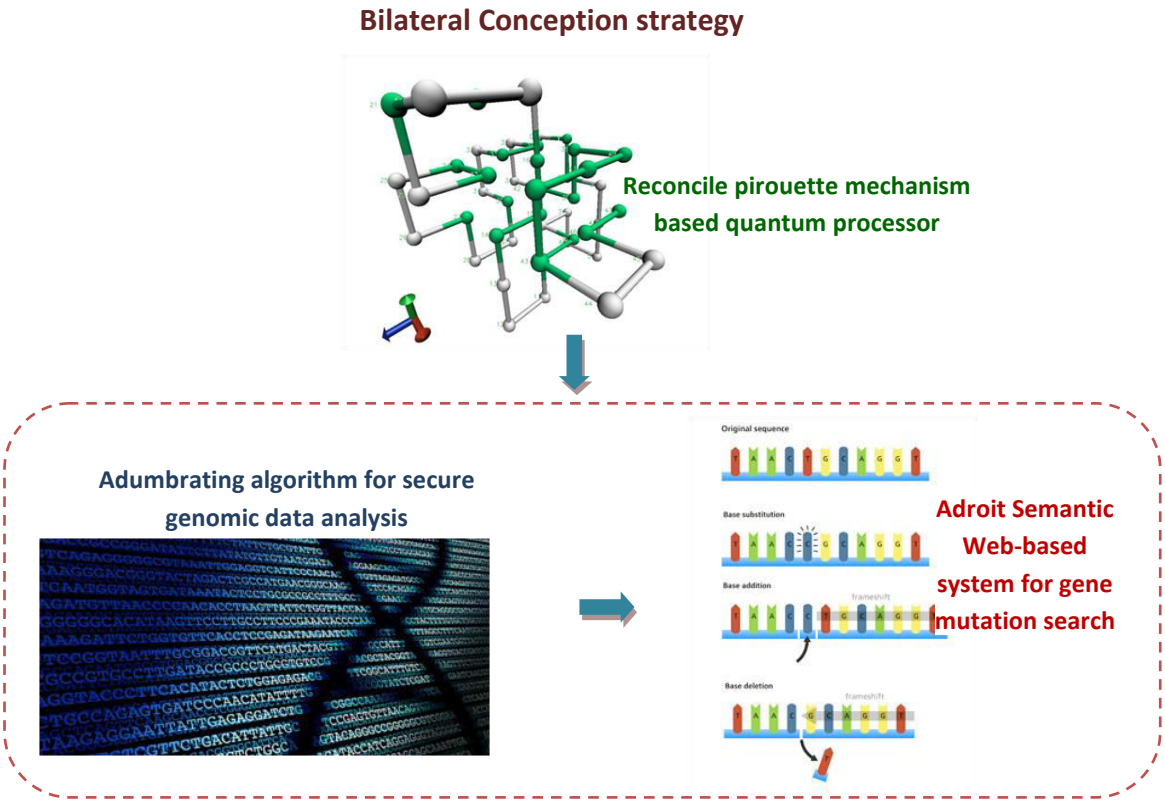

Fig. 1 Block diagram of proposed work

At last, to enlighten the captures and manages gene mutation evidence and related contextual information in a well-structured format, the work includes the new Adroit Semantic Web-based system. That comprises the exegetics and semantic web module, which provides a scalable framework for standard-based data representation, integration and sharing then it stores the extracted annotations in a structured representation using Collateral Description Framework. Subsequently, the proposed bilateral conception strategy overwhelmed the most significant issues in protein folding and genomic data processing in an acrobatic and cost-effective manner, which perceptibly helps the majority of amyloidogenic diseases.

3.1 Reconcile Pirouette Mechanism-Based Quantum Processor

Prior researches have conquered more intricacy for folding longer protein sequences due to the complicated size and difficulty to find the lowest energy conformation of the relative location and orientation of protein secondary structure elements in three dimensions. So, the novel Reconcile pirouette mechanism-based quantum processor is proposed to deal with these issues by decreasing the quantum circuit from quadratic to quasi-linear. For this, the injective

mapping is constructed between the set of all possible lattice protein folds and the set of binary strings, represented by a sequence of qubits in the machine. Thus the solution string can decode into lattice protein fold uniquely. Then it constructs the energy landscape for the Ising system such that the valid, lowest energy conformation of the lattice protein corresponds to the ground state of the system. For this purpose, this work uses pseudo-boolean expressions that are subsequently reduced to 2-local interactions implementable on the device.

The most compact way and globally defined directions called turns are used to encode the lattice proteins on the cubic lattice. Fig.2 shows a binary mapping that encodes each of the six spatial directions on a cubic lattice as 3 qubits, which requires a total of *Ω(l)*qubits. This proposed work requires only $3L - 8 \in O(l)$ qubits to encode the protein length of l. In previous researches, the k-body Hamiltonian can be reduced to a 2-body Hamiltonian with an equivalent ground state by introducing ancillaqubits as resource-efficient gadgets.But this proposed work is more efficient for encoding small problem instances.The turn circuit encoded Hamiltonian consists of two main terms,

$$H(q) = H_{ovlap}(q) + H_{pair}(q) \quad \ldots. (1)$$

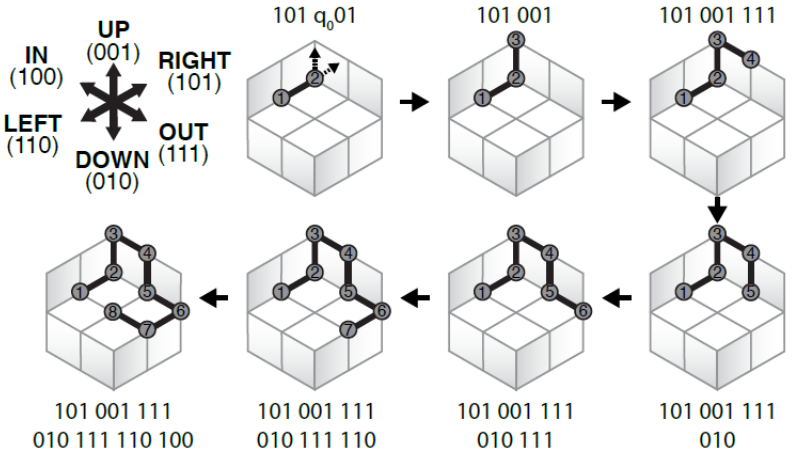

Fig.2 Binary mapping in cube lattice

For reducing the size complexity, the sum strings are constructed. By constructing some strings for every pair of amino acids, the position of every amino acid on the lattice can be traced,which will enable us to test for possible overlaps or interactions of residues. Half-adder circuits are used to construct the sum strings. The basic half adder circuit is shown in fig.3. but this proposed work constructs the better designed larger half adder that reduces the number of terms in the Hamiltonian. Since a binary representation of n bits uses at most ($\log_2$N) bits, it follows that [N-($\log_2$N)] bits are not required to represent the sum string. So that, the half adder at the upper right section of the circuit can be removed once their whole information is propagated to the lower section in fig.4. By not adding these empty bits using superfluous half-adders can avoid inflation of the overall Hamiltonian which would be due to each half-adder introducing new high-order terms. Thus the half adder results in the sum string S.

Fig.3 Half-adder circuit

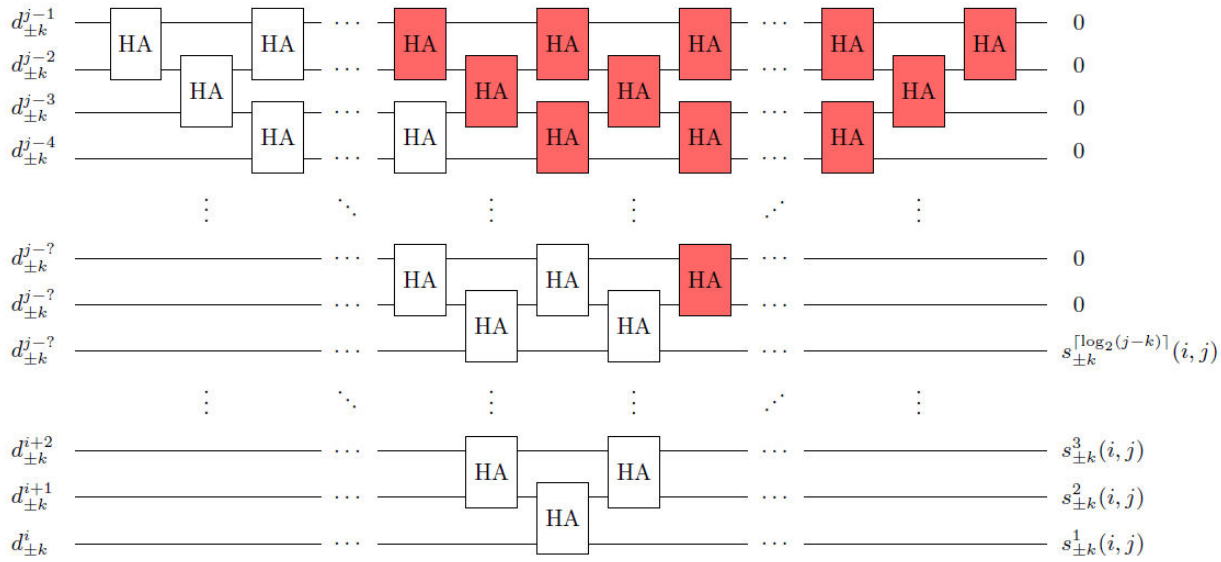

Fig.4 Circuit diagram for proposed work. In this circuit, the red-colored half-adders are superfluous and can be omitted which reduces the number of half adders from quadratic to quasilinear which reduces the circuit complexity.

From the above, a significant quadratic to quasilinear improvement in circuit complexity for sum strings is obtained. Now, this paper presents the analysis of the reduction in circuit complexity. The total number of half-adders $HA_{total}(N)$ in the sum string circuit of n bits is,

$$HA_{total} = 1 + 2 + \cdots + N = \frac{N(N+1)}{2} \quad \ldots. (2)$$

The circuit complexity in terms of a number of half-adders involved is then $O(N^2)$. However, since the sum of n bits only needs $[\log_2(N+1)]$ output bits, the number of bits N_{redun} not containing any information output is,

$$N_{redun} = N - \log_2(N+1) \quad \ldots. (3)$$

The number of associated half-adders that cannot propagate any information to output bits is equal to $HA_{total}(N_{redun})$, by isomorphism to the sum string circuit of N_{redun} binary variables. The necessary number of half-adders HA_{impro}in the improved circuit for the addition of n binary variables is then,

$$HA_{improv}(N) = HA(N) - HA_{total}(N_{redun}) \quad \ldots. (4)$$

$$= \frac{N^2+N}{2} - \frac{N_{redun}^2+N_{redun}}{2}$$

$$= \frac{N^2+N}{2} - \frac{(N-[\log_2(N+1)])^2+N-[\log_2(N+1)]}{2}$$

$$HA_{improv}(N) = \frac{2N[\log_2(N+1)]-[\log_2(N+1)]^2+[\log_2(N+1)]}{2} \quad \ldots. (5)$$

It follows that $HA_{improv}(N) \in O(N \log N)$. Thus the proposed mechanism provides a significant decrease in the circuit complexity, from quadratic to quasilinear. As a result, the size complexity in the protein fold has been highly reduced. Then, the privacy issues in the genome sequence are discussed in the forthcoming section.

3.2 Adumbrating Algorithm for Secure Commune Based Genomic Data Analysis

In the view of sharing genomic sequence data, existing approaches enable all people concerned to share individual patient data and conduct all analysis locally which is typically not feasible due to privacy issues, and collaborative data analysis remains to be a rarity in genomic medicine. To make this an efficient and protective manner, this paper introduces Adumbrating algorithm for Secure Commune (ASC) based genomic data analysis for collaborative or remote genomic computation which uses Intel's Software Guard Extensions(SGX) architecture. An SGX consists of one or more data owners, the untrusted cloud service provider (CSP), and a secure enclave or commune. First, the data owner establishes a secure channel with the enclave hosted by an untrusted CSP through the remote attestation process. Then, the data owner can securely upload data to the CSP. In SGX, all decrypted secrets can only be accessed by the authorized codes which also lie inside the enclave. Therefore, code and data cannot be accessed or modified by any software outside the secure enclave.

Block diagram for ASC based genomic data analysis illustrated in fig.5. ASC based genomic data analysis involves individual genomic data in the form of VCF files from one or more parties who would like to perform statistical tests on the entire data set.Each VCF file is marked as either case or control and is individually filtered, compressed, and encrypted, and is uploaded to an untrusted CSP. The specific analysis/querying offered by ASC to the users is, given a user-specified set of SNPs (which can be the entire set of SNPs in the human genome or a subset) and an integer k. ASC first processing the entire data set to establish a sketch within the enclave. On a given query, ASC identifies a super-set of potentially significant SNPs and re-accesses the relevant portions of the VCF files to identify the most significant k SNPs. For each SNP, a VCF file includes "ID", "TYPE", "CHROM", "POS", "REF", "ALT", "QUAL", and "FILTER" columns.Each VCF file needs to be encrypted, sent over some channel, loaded into the SGX enclave, decrypted, and finally processed, reducing the size of the data can improve the overall performance significantly.

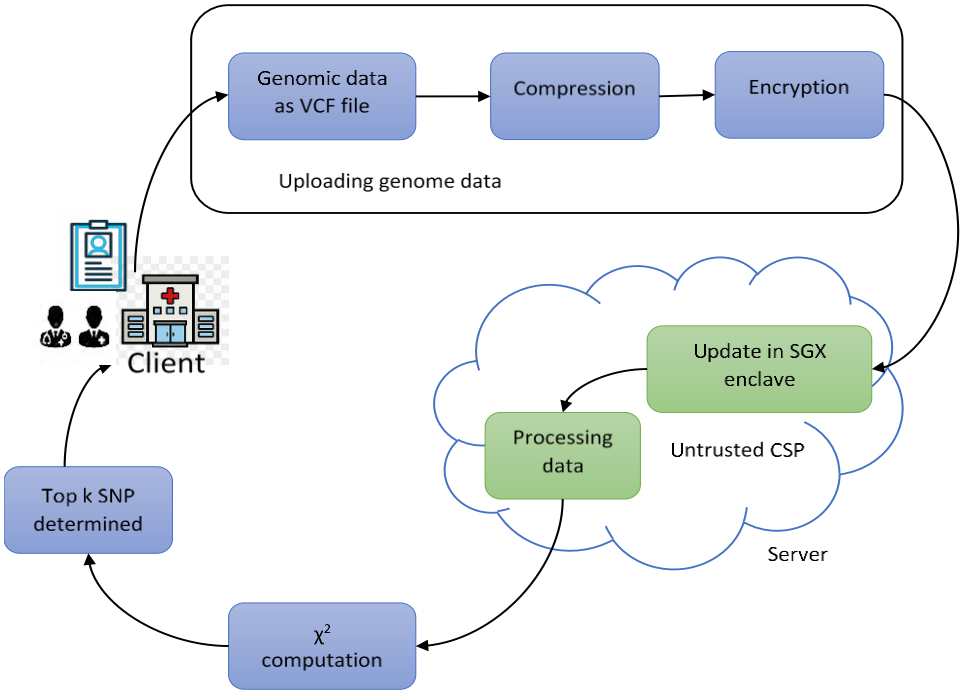

Fig. 5 Block diagram for ASC based genome data analysis

3.2.1 Identifying top Single Nucleotide Polymorphism (SNP)

Once the compression of genomic data has been performed, the parties can exchange data and messages required for the actual computation. The tasks are divided between the server which refers to the cloud service provider (CSP) equipped with Intel SGX supported hardware, and the client which refers to individual users (e.g. genome centers) that want to perform their collaborative analysis in a privacy-preserving and secure manner. In Intel SGX, the server begins by creating and initializing an SGX enclave and waits for service requests from the clients.Then, a client needs to initiate the remote attestation protocol to establish that the server is indeed recognized by Intel and is going to perform the desired computation.the server allocates sufficient resources to performs the computation.In this process, the majority of the SGX enclave memory is used to keep either the basic data structures in the form of hash tables maintaining allele counts or space-efficient data structures that are employed when the SGX memory cannot store all SNP IDs.Even though it offers a limited memory, Intel SGX secure enclave can still handle genomic data analysis on a limited scale (e.g. GWAS on a single human chromosome) through the use of standard (non-sketching) data structures.

As the server receives the compressed and encrypted data from the clients, the data structures residing in the SGX enclave are updated.Once all incoming data has been processed, the standard χ^2 statistic can be computed for each SNP entry within the hash table to determine the top-k SNPs with respect χ^2 to the statistic. For the case that the SNP IDs and their corresponding allele counts exceed the memory limit, ASC identifies some l SNPs (l > k) which

include the top-k SNPs with high probability and processes the data in another pass to filter out all but the most significant k SNPs. Finally, the results are sent back to the clients.

3.2.2 Computing top-k SNPs with respect to χ^2 statistics

Maintaining all SNPs in the human genome and their respective allelic counts exceeds the memory limit of the secure enclave. In other words, the working memory of the SGX enclave has n'= O(n) words (i.e. O (n (log m + log n)) bits). A straightforward solution, in this case, is to partition the n dimensions (SNP IDs) from the input v_j's into $\frac{n}{n'}$ blocks, and process each block (of n' dimensions) independently in the enclave memory. For a fixed k, the overall top-k dimensions (based on the standard χ^2 statistic) can then be obtained by sorting the top-k dimensions of all blocks.

The alternative solution is also implemented by ASC, which allows users to specify the value of k as well as a subset of SNPs of interest (among which the top-k dimensions need to be determined), even after all input data is processed. This solution maintains a summary of the potentially important dimensions and allows the user to identify the top-k dimensions with respect to the χ^2 values with high probability. ASC processes the input vectors v_j in a single pass: for each (v_j[i], sign(j)) it updates the appropriate entries of the sketch in an on-line fashion. The sketch basically maintains appropriate counts for l candidate dimensions (for a sufficiently large l > k) which include all top-k dimensions (with respect to the $\chi 2$ statistic) with high probability. Once the sketch is complete ASC accesses the relevant vectors v_j in a second pass to filter out the false positives among these candidates and identify only the most significant k dimensions.

Thus the effective protection is accomplished by recognizing the top small value (k) of SNP which is identified only amongst a user-specified subset of SNPs across case/control samples. After implementing the security of genomic data, the storage and the fast searching of data are implemented in the next section.

3.3 Adroit Semantic Web-Based System

Along with the privacy of genomic data, research on genetic mutations and the associated background has been considered an important source of evidence for compiling pharmacogenomics (PGx). Unfortunately, this data occurs in a compliance requirement for the free-text format, which is difficult to parse using traditional text mining techniques. Thus, most of the systems cannot provide sufficient mutation annotations for clinical trials, which makes it hard to capture and store due to the unstructured format. Consequently, it difficult to search for relevant trials based on patients' mutation information. To deal with these issues, this paper focuses on data integrity verification in the medical environment with clients' privacy protection using a novel Adroit semantic web-based system.

This proposed system having four phases,

(i) In the system setup phase, the Key Generation Center (KGC) sets the system public parameters and a master secret key.

(ii) In the registration phase, the KGC generates privacy keys for users and secret keys for auditing in the registration phase.

(iii) In the storage phase, users upload and update files to the cloud along with file warrants, authenticators, and tags.

(iv) In the integrity verification phase, Third Party Auditor (TPA) is entrusted by the data owner to verify corresponding data integrity.

(i) System setup phase:

A system taking a security parameter $\boldsymbol{K}$ as input, the KGC randomly selects two multiplicative cyclic groups G and $G_T$with prime order q, where g is a generator of G.After that, the KGC picks an integer$a \in_R Z_q^*$ at random and computes $g^1=g^a$where $g \in G$. Next, $v_0,v_1,\ldots v_l,u_1,\ldots u_s \in_R G$ is uniformly chosen at random. Thus, the system public parameter $PP=(g,g_1,g_2,v_0,v_1,\ldots v_l,u_1,\ldots u_s \in_R G,H_1,H_2,H_3,H_4)$, where, $H_1,\ldots H_4$ are the hash functions. Finally, the master secret key *msk*is set as $msk=g_2^a$with $g_2 \in G$ and keeps the *msk*in secret by the KGC.

(ii) Registration phase:

The KGC runs the *KeyGen*algorithm to yield a shared secret key for users with the *msk*and public parameter *PP*. The registration procedure consists of two phases:

(a) *PrivacyKeyGen*

(b) *SecretKeyGen*

(a) PrivacyKeyGen:

First, the KGC generates and distributes the corresponding private key for every user who may be a patient or a consultant in an e-healthy system. In detail, the KGC computes $Q_i$based on the user's identity as $Q_i=H1(ID_i)$. Then, KGC calculates the user private key is:

$$d_i = g_2^a.H_1(ID_i) \quad \ldots. (6)$$

For example, KGC independently yields a private key $d_A$for patient A, and a private key $d_B$for the attending physician B. Then, the KGC sends $d_i$to ID_i. After receiving the d_i, the user validates $ID_i$by calculating:

$$e(d_i,g) = e(g_2,g_1).e(H_1(ID_i),g) \quad \ldots. (7)$$

If the above equation is true, the user ID_iadopts the private key d_i; otherwise, the KGC fails to generate a valid private key.

(b)SecretKeyGen:

To protect the identity of patient A, patient A randomly chooses a number $r_A \in_R Z_q^*$, generate the pseudonym $P_A = r_A.Q_A$,and sends it instead of his or her actual identity to B. Then, A and B can calculate a session key K_{AB}, and this algorithm produces a secret key *KAB* for auditing. The specific algorithm is as follows:

$$K_{AB} = e(d_A,Q_B) = e(P_A,d_B) \quad \ldots. (8)$$

$$KAB = g_2^a.H_2(K_{AB}) \quad \ldots. (9)$$

(iii) Storage Phase:

The storage procedure contains the following three phases:

(a) *WarrantGen*
(b) *AuthenticatorGen*
(c) *TagGen*

(a) WarrantGen:

When the user uploads or updates new medical data, the corresponding file information will be updated. For confirming some additional information about the source, type, and consistency of the files outsourced to the cloud, the user generates a warrant Λ which includes the pseudonym of A, the identity hash value Q_i of attending physician B, and medical file information such as file type *filetype*, version number V_N, Time Stamp T_N, etc. For example, $\Lambda = P_A \| Q_B \| V_N \| T_N \| filetype$. Here, the N denotes the index of different medical files. Then, the following is calculated:

$$\overrightarrow{\Lambda} = (\zeta_1, \dots, \zeta_l) \leftarrow H_3(\Lambda) \quad \dots. (10)$$

The patient A picks a random number $t_\Lambda \in_R Z_q^*$,and generates an authorization:

$$\delta_\Lambda = (KAB.(v_0.\prod_{j=1}^{l} v_j^{\zeta_j})^{t_\Lambda}, g^{t_\Lambda}) \quad \dots. (11)$$

Finally, patient A sends the warrant pair$(\Lambda, \delta_\Lambda) = (\Lambda, (\alpha, \beta))$ to attending physician B. Then, the attending physician B validates the warrant pair by calculating:

$$e(\alpha, g) = e(g_2, g_1).e(H_2(K_{AB}), g).e(v_0 \prod_{j=1}^{l} v_j^{\zeta_j}, \beta) \quad \dots. (12)$$

If the above equation is true, the attending physician B accepts the authorization δ_Λ; otherwise, patient A fails to generate a valid warrant.

(b)AuthenticatorGen:

Given a medical file F to be outsourced, the user first splits F into n blocks, and each contains s sectors: $F \rightarrow \{\chi_{i,j}\}_{n \times s}$, where $\chi_{i,j} \in_R Z_q^*$. For each file F, choose a random number $t_\vartheta \in_R Z_q^*$, and for the i-th block, yield a block authenticator as follows:

$$\sigma_i = KAB.(H_4(\Lambda \parallel \text{FID} \parallel \text{i}).\prod_{j=1}^{s} u_j^{\chi_{i,j}})^{t_\vartheta} \quad \dots. (13)$$

(c)TagGen:

A random name *FID* is chosen for a file from Z_q^*, and s random elements $u_1, \dots, u_s \in G$. Set $\tau_0 = \Lambda \parallel FID \parallel n \parallel u_1 \parallel \cdots \parallel u_s \parallel g^{t_\Lambda} \parallel g^{t_\vartheta}$. Then, the user generates file tag τ based on $\tau_0$and $K_{AB}$to guarantee the integrity of each distinct file information.

$$\tau = \tau_0 \parallel S.Sign(\tau_0)_{K_{AB}} \quad \dots. (14)$$

Hereafter, the user sends the file tag τ to the TPA. Besides, $KP = e(H_2(K_{AB}), g)$can be pre-computed and sent to TPA. Besides, the processed fileF^*that comprises F, *FID*, Λ, δ_Λ, and $\sigma_i$is uploaded to the CS and can be stored in the collateral description structure and removed from the user's local side.

(iv) Integrity Verification Phase:

The integrity verification can be done through the auditing process which contains the following three phases:

(a) *Challenge*
(b) *Response*
(c) *Verification*

(a) Challenge:

First, the TPA confirms whether the file tag τ of outsourced file can pass the verification by retrieving τ from the CS and performing $S.Vrf(\tau_0, K_{AB})$. If the file tag τ of outsourced file cannot pass the verification, then the auditing task will not be executed, and the protocol aborts; otherwise, the TPA will analyze $\tau_0$to acquire the total number n of outsourced file blocks. The TPA picks a random nonempty subset $I \subseteq [1, n]$ and a number of values $s_i \in_R Z_q^*$ at random, for each $i \in I$. Then, the TPA distributes the challenge set $C = \{(i, s_i)_{i \in I}\}$and corresponding file identifier *FID* to the CS. After that, the TPA can compute $WP$for the final verification as

$$WP = e(H_4(\Lambda \parallel FID \parallel i), g^{t_\vartheta})^{\sum_{i \in I} s_i} \quad \ldots. (15)$$

(b) Response:

CS locates to the corresponding file F^*in the collateral description structure upon receiving a challenge C and its file identifier *FID* from the TPA. Then, the CS computes χ_j and σ as:

$$\chi_j = \sum_{i \in I} s_i \cdot \chi_{i,j} mod q, j \in [1, n] \quad \ldots..(16)$$

$$\sigma = \prod_{i \in I} \sigma_i^{s_i} \quad \ldots. (17)$$

After that, the CS sends to the TPA a proof P that consists of $\chi_i, \ldots, \chi_s, \sigma$ and corresponding authorization δ_Λ.

(c) Verification:

Once receiving the proof P, with public system parameter PP and file tag τ, the TPA first verifies the validity of $\delta_\Lambda$by demonstrating the equation (12), and then, verifies aggregate block authenticator σ as follows:

$$e(\sigma, g) = e(g_2, g_1)^{\sum_{i \in I} s_i} . KP^{\sum_{i \in I} s_i} . WP . e(\prod_{j=1}^{s} u_j^{\chi_j}, g^{t_\vartheta}) \quad \ldots. (18)$$

If the above equation is true, the challenged outsourced file in the cloud is verified as intact; otherwise, the challenged file is corrupted. In the above auditing process, TPA can also audit the details of the challenged file warrant. That is, the proof P, which will be fed back by CS, should contain more file details.

Thus the novel Adroit semantic web-based system compromising the issues as storing and verifying the genome annotations in a highly protective manner.

As a result of the novel Bilateral conception strategy, the novel Reconcile pirouette mechanism-based quantum processor, highly reduces the size complexity of protein folding. Then, the novel Adumbrating algorithm for secure commune-based genomic data analysis provides an effective security system for the genomic data. Furthermore, the novel Adroit semantic web-based system provides the well-structured format called collateral description framework for storing the genome data as well as this also provides the secured uploading of data in a web-based system.

4 Results and Discussion

This section provides a comprehensive description of the implementation result, performance analysis, and the comparison strategies of this proposed work.

4.1 Experimental Setup

This work has been implemented in MATLAB/SIMULINK in the working platform of MATLAB with the following system specification and the simulation results are discussed below

Platform: MATLAB 2018b

OS: Windows 8

Processor: Intel Core i5

RAM: 8GB RAM

4.2 Evaluation Metrics

The performance of the Bilateral conception strategy has been evaluated with the metrics such as accuracy, precision, and recall. Thus this section shows the performance analysis of the Bilateral conception strategy along with its evaluation metrics.

4.2.1 Accuracy

Accuracy is the ratio of the number of correct predictions of protein folding to the total samples. This proposed method attains a higher accuracy for a lower threshold value.

$$Accuracy = \frac{TruePositive+TrueNegative}{Allsamples} \quad \text{.... (19)}$$

4.2.2 Precision

Precision evaluates the fraction of correctly predicted instances or samples among the ones predicted as positives. Thus, the formula to calculate the precision is given by:

$$\square\square\square\square\square\square\square\square\square = \frac{\square\square\square\square\square\square\square\square\square\square\square\square\square}{\square\square\square\square\square\square\square\square\square\square\square\square\square+\square\square\square\square\square\square\square\square\square\square\square\square\square\square}$$

.... (20)

4.2.3 Recall

The recall is defined as the ratio of pertinent data that are recovered successfully.

$$\square\square\square\square\square\square = \frac{\square\square\square\square\square\square\square\square\square\square\square\square\square}{\square\square\square\square\square\square\square\square\square\square\square\square\square+\square\square\square\square\square\square\square\square\square\square\square\square\square\square}$$

.... (21)

4.3 Performance Evaluation

Inthis section, the performance of the proposed system has been analyzed graphically with the metrics. Fig.6 and table 1 illustrate that the variations in accuracy of the finding of lattice proteins with a threshold value.

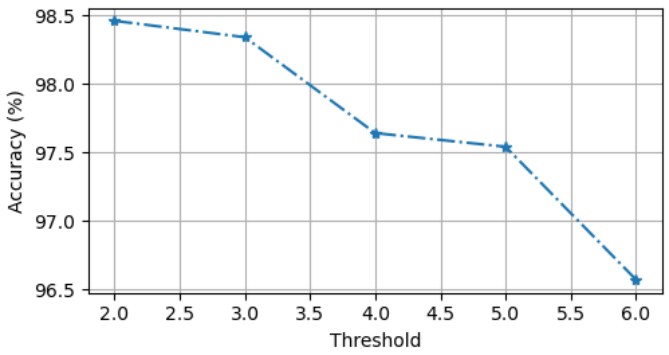

Fig. 6 Accuracy for various Threshold

Table 1 Accuracy for various threshold

Threshold	Accuracy(%)
2	98.46
3	98.34
4	97.64
5	97.54
6	96.57

Table 2 comprises the computation overhead of the proposed Adroit semantic web-based system. Primarily, the following notations are defined to represent the various operations in the specific algorithms of each phase. The symbols M, E, and H denote a multiplication operation, an exponentiation operation, and a hashing operation in *G*, respectively. In this paper, *H1*, *H2*, and *H3* are not distinguished and all can be expressed as H. Similarly, the symbols Mt and Ethave respectively expressed as a multiplication operation and an exponentiation operation in G_T. Aq and Mq are indicated as one addition operation and one multiplication operation inZ_q, respectively. And P represents a bilinear pairing evaluation operation.

Table 2 Computation overhead of Adroit semantic web-based system

Phases	**KGC**	**User (physician)**	**User (patient)**	**TPA**	**CS**
Setup	2E	/	/	/	/
KeyGen(a)	M + H	2P + H + Mt	2P + H + Mt	/	/
KeyGen(b)	/	P + H + M	P + 2P + Mq +M	/	/
Extract(a)	/	3P + 2H + 2Mt + lM	2E + H + (l+1)M	/	/
Extract(b)	/	E + H + (s +1)M	/	/	/
Audit(b)	/	/	/	/	n/I/Mq+n(/I/-1)Aq + (/I/M+/I/E
Audit(c)	/	/	/	(s+1)E + H +(I-1)A + 3P +3Et +(s+1)M + Mt	/

4.4 Comparison Strategies

In this section, the proposed methodology for genome analysis has been compared with the previously existing methods such as Bi-directional Best Hit (BBH) technique and DISPattern algorithm with two different genomes are NC_000962 and NC_002929.

Table 3 and Fig.7 illustrates that the Recall and Precision of BBH and DISPattern for the genomes NC_000962 and NC_002929 are compared with our proposed method. From the graph, one can say clearly that the recall and the precision of the proposed method have been highly enhanced. Fig.8 compares the number of the genome in the proposed method with the different genome sequences. It is clear that the genomes in the proposed method are highly reduced.

Fig.9,10 compares the number of errors in the proposed method with the existing techniques such as BBH and DISPattern for two different genomes NC_000962 and NC_002929 respectively.

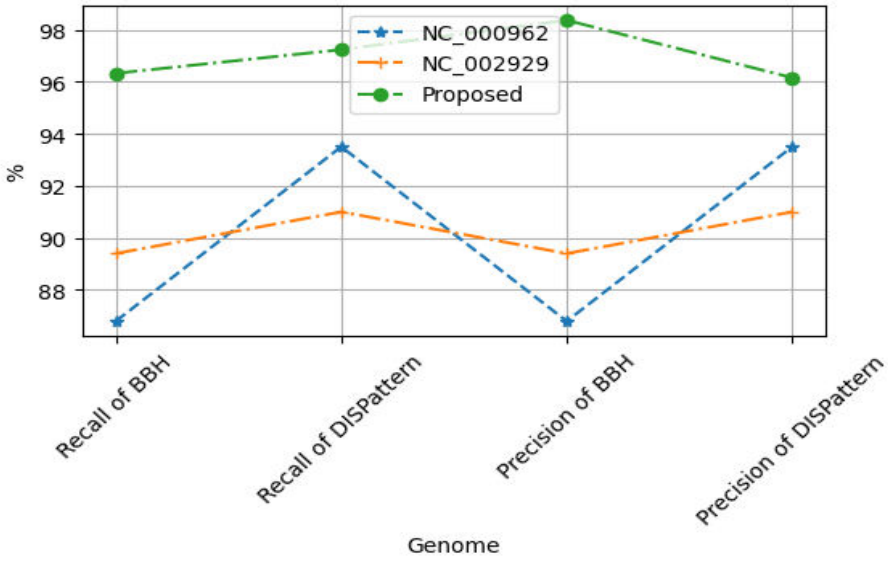

Fig.7 Comparison of recall and prediction of BBH and DISPattern with the proposed method

Table 3 Comparison with BBH and DISPattern

Genome	NC_000962	NC_002929	Proposed
No. of genes	2756	2723	2712
Recall of BBH	86.80%	89.40%	96.32
Recall of DISPattern	93.50%	91.00%	97.23
Precision of BBH	86.80%	89.40%	98.36
Precision of DISPattern	93.50%	91.00%	96.15
No. of errors by BBH	186	106	95

No. of errors by DISPattern	66	61	33

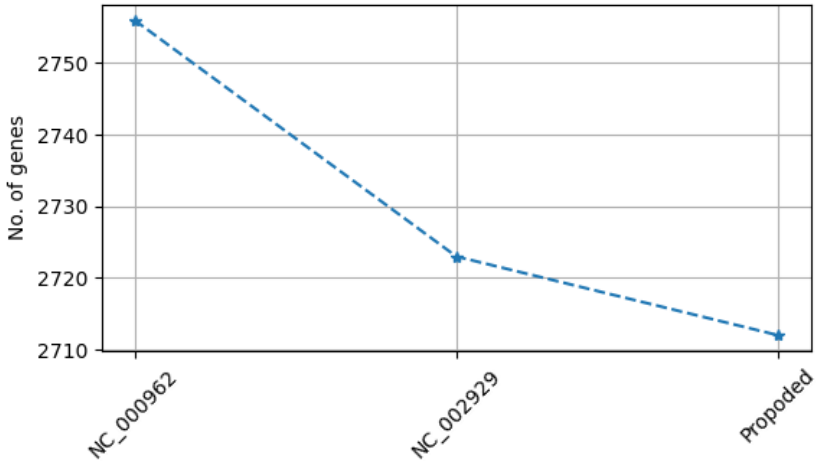

Fig. 8 Comparison of the number of genomes

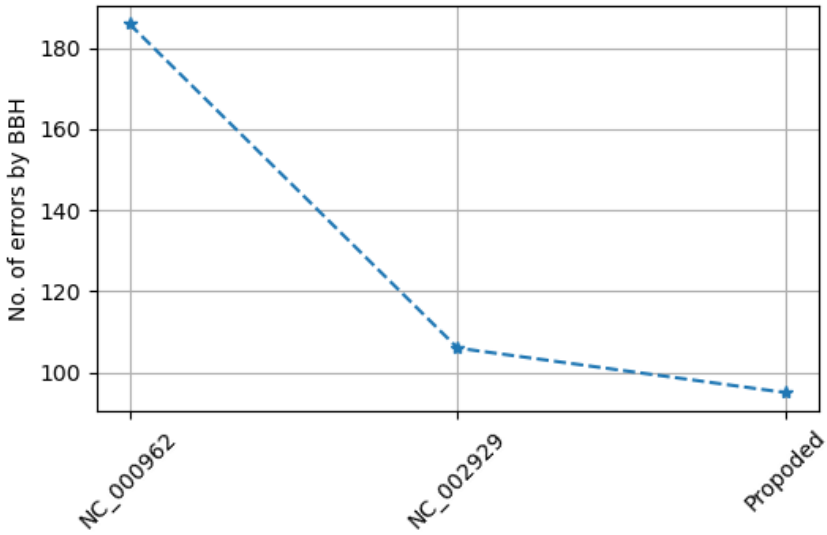

Fig.9 Comparison with BBH for no. of errors

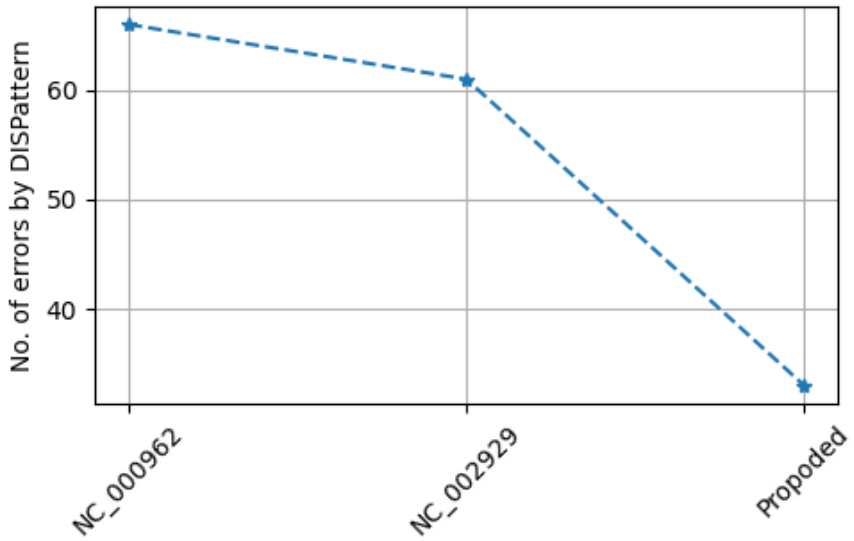

Fig.10 Comparison with DISPattern for no. of errors

Table 4 compares the characteristics of the proposed adroit semantic web-based system with the existing researches. From the table, one can easily identify that our proposed method has been complete all the major characteristics.

Table 4 Characteristic comparison of the proposed method with existing methods.

Schemes	Public verifiability	Certificate management simplification	Privacy protection	Dynamic operations
Worku et al.	√	×	√	√
Garg et al.	√	×	×	√
Daniel et al.	√	×	√	×
Zhao et al.	×	√	×	×
Jiang et al.	√	×	√	√
Proposed	√	√	√	√

Thus, the performance of the proposed methods has been verified and effectively compared with the existing techniques. As a result, the proposed methods have been performed well than the existing techniques.

5 Conclusion

As genome sequencing technologies and protein folding is a most essential biological process in the health care field. This paper provides the solution for the issues in those

technologies. The novel Reconcile pirouette mechanism-based quantum processor solves the problem of finding a lattice protein's lowest energy conformation by decrease the quantum circuit from quadratic to quasi-linear, thus it reduces the size complexity. Then the novel Adumbrating algorithm for secure commune-based genomic data analysis solves the issues such as data breaches and collaborative data analysis by introducing small value in SNP, thus provides the highest privacy for genome data. Finally, the novel Adroit semantic web-based system solves the issues in searching genome data by introducing collateral description frame format, thus the data stored in the structured format and provides ease of search. As a result, the proposed work effectively solves the major issues in protein folding and genome sequences.

References

[1]. Wang, X., Tang, H., Wang, S., Jiang, X., Wang, W., Bu, D., Wang, L., Jiang, Y. and Wang, C., 2018. iDASH secure genome analysis competition 2017.

[2]. Chen, F., Dow, M., Ding, S., Lu, Y., Jiang, X., Tang, H. and Wang, S., 2016. PREMIX: Privacy-preserving EstiMation of individual admixture. In *AMIA Annual Symposium Proceedings* (Vol. 2016, p. 1747). American Medical Informatics Association.

[3]. Shimizu, K., Nuida, K. and Rätsch, G., 2016. Efficient privacy-preserving string search and an application in genomics. *Bioinformatics*, *32*(11), pp.1652-1661.

[4]. Tubiana, J., Cocco, S. and Monasson, R., 2019. Learning compositional representations of interacting systems with restricted boltzmann machines: Comparative study of lattice proteins. *Neural computation*, *31*(8), pp.1671-1717.

[5]. Babej, T. and Fingerhuth, M., 2018. Coarse-grained lattice protein folding on a quantum annealer. *arXiv preprint arXiv:1811.00713*.

[6]. Fingerhuth, M. and Babej, T., 2018. A quantum alternating operator ansatz with hard and soft constraints for lattice protein folding. *arXiv preprint arXiv:1810.13411*.

[7]. Wang, Y., Li, X., Zang, D., Tan, G. and Sun, N., 2018, August. Accelerating fm-index search for genomic data processing. In *Proceedings of the 47th International Conference on Parallel Processing* (pp. 1-12).

[8]. Masseroli, M., Canakoglu, A., Pinoli, P., Kaitoua, A., Gulino, A., Horlova, O., Nanni, L., Bernasconi, A., Perna, S., Stamoulakatou, E. and Ceri, S., 2019. Processing of big heterogeneous genomic datasets for tertiary analysis of Next Generation Sequencing data. *Bioinformatics*, *35*(5), pp.729-736.

[9]. Rowe, W.P., 2019. When the levee breaks: a practical guide to sketching algorithms for processing the flood of genomic data. *Genome biology*, *20*(1), p.199.

[10]. Yang, J., 2019. Cloud computing for storing and analyzing petabytes of genomic data. *Journal of Industrial Information Integration*, *15*, pp.50-57.

[11]. Vaske, C.J., Sanborn, J.Z. and Benz, S.C., Five3 Genomics LLC, 2018. *Distributed system providing dynamic indexing and visualization of genomic data*. U.S. Patent 10,140,683.

[12]. Grishin, D., Obbad, K., Estep, P., Quinn, K., Zaranek, S.W., Zaranek, A.W., Vandewege, W., Clegg, T., César, N., Cifric, M. and Church, G., 2018. Accelerating genomic data generation and facilitating genomic data access using decentralization, privacy-preserving technologies and equitable compensation. *BlockchainHealthc Today*, *1*, pp.1-23.

[13]. Langmead, B. and Nellore, A., 2018. Cloud computing for genomic data analysis and collaboration. *Nature Reviews Genetics*, *19*(4), p.208.

[14]. Li, X., Tan, G., Wang, B. and Sun, N., 2018. High-performance genomic analysis framework with in-memory computing. *ACM SIGPLAN Notices*, *53*(1), pp.317-328.

[15]. Katayama, T., Kawashima, S., Okamoto, S., Moriya, Y., Chiba, H., Naito, Y., Fujisawa, T., Mori, H. and Takagi, T., 2019. TogoGenome/TogoStanza: modularized Semantic Web genome database. *Database*, *2019*.

[16]. Koehorst, J.J., van Dam, J.C., Saccenti, E., Martins dos Santos, V.A., Suarez-Diez, M. and Schaap, P.J., 2018. SAPP: functional genome annotation and analysis through a semantic framework using FAIR principles. *Bioinformatics*, *34*(8), pp.1401-1403.

[17]. Thomas, P.D., Hill, D.P., Mi, H., Osumi-Sutherland, D., Van Auken, K., Carbon, S., Balhoff, J.P., Albou, L.P., Good, B., Gaudet, P. and Lewis, S.E., 2019. Gene Ontology Causal Activity Modeling (GO-CAM) moves beyond GO annotations to structured descriptions of biological functions and systems. *Nature genetics*, *51*(10), pp.1429-1433.

[18]. Giglio, M., Tauber, R., Nadendla, S., Munro, J., Olley, D., Ball, S., Mitraka, E., Schriml, L.M., Gaudet, P., Hobbs, E.T. and Erill, I., 2019. ECO, the Evidence & Conclusion Ontology: community standard for evidence information. *Nucleic acids research*, *47*(D1), pp.D1186-D1194.

[19]. Acharya, S., Saha, S. and Pradhan, P., 2018. Novel symmetry-based gene-gene dissimilarity measures utilizing Gene Ontology: Application in gene clustering. *Gene*, *679*, pp.341-351.

[20]. Brionne, A., Juanchich, A. and Hennequet-Antier, C., 2019. ViSEAGO: a Bioconductor package for clustering biological functions using Gene Ontology and semantic similarity. *BioData Mining*, *12*(1), p.16.

[21]. Cao, A., 2020. The Last Secret of Protein Folding: The Real Relationship Between Long-Range Interactions and Local Structures. *The Protein Journal*, pp.1-12.

[22]. Li, J., Tian, Y., Zhu, Y., Zhou, T., Li, J., Ding, K. and Li, J., 2020. A multicenter random forest model for effective prognosis prediction in collaborative clinical research network. *Artificial Intelligence in Medicine*, *103*, p.101814.

[23]. Xiang, T., 2020. Integrating a Genome-Wide Association Study WithTranscriptome Analyses to Identify Candidate Genes and Pathways for Feed Conversion Ratio in Yorkshire Pigs.

[24]. Trębacz, M., Shams, Z., Jamnik, M., Scherer, P., Simidjievski, N., Terre, H.A. and Liò, P., 2020. Using ontology embeddings for structural inductive bias in gene expression data analysis. *arXiv preprint arXiv:2011.10998*.

[25]. Quan, Y., Zhang, Q.Y., Lv, B.M., Xu, R.F. and Zhang, H.Y., 2020. Genome-wide pathogenesis interpretation using a heat diffusion-based systems genetics method and implications for gene function annotation. *Molecular Genetics & Genomic Medicine*, *8*(10), p.e1456.

[26]. Yogesh Hole et al 2019 J. Phys.: Conf. Ser. 1362 012121