Natural language processing, big data, bioinformatics and biology

Emdad Khan

Abstract – As we know, the most complex machine in this world is human being, especially, our brain. Understanding the human biological system and how human brain really works (even partially) are top research areas being addressed by many researchers around the world. This effort has been expedited significantly since the completion of the human genome project. With the rapid growth of biological data, this field has become even more multi-disciplinary that includes Big Data, Bioinformatics, Biology and Natural Language Processing (NLP). The intersection of NLP is interesting and important as NLP can contribute from multiple angles and help solve various problems in Big Data, Biology, Bioinformatics and more.

In this paper, we propose Semantic Engine using Brain-Like Approach (SEBLA) and associated NLP & Natural Language Understanding (NLU) based approach to address the key problems of Big Data in Bioinformatics and Biology. Our approach (SEBLA-NLU) resembles human Brain-Like and Brain-Inspired algorithms as humans can significantly compress the data by representing with a few words or sentences using the semantics of the information while preserving the core meaning. Thus, it very effectively converts data to knowledge and also compresses it; and hence addresses the key Big Data problems in an effective way. We describe how SEBLA-NLU can be used to handle both unstructured and structured Big Data for addressing complex problems including summarization and analytics. We also describe how SEBLA-NLU can help understand the DNA (including non-coding DNA) and hence biological systems/processes (e.g. Gene Expression, Gene Function, Protein Function, Protein Interactions and Protein Scaffolding). We also discuss how SEBLA-NLU can help the modeling aspect of biological systems / processes.

Keywords **----** Bioinformatics; Biology; Big Data; Unstructured Data; Natural Language Processing (NLP); Natural Language Understanding; Semantics; Semantic Engine; Intelligent Agent; Predictive Analysis; Business Intelligence; Biological Systems Modeling; Knowledge Discovery; DNA; Gene; Gene Function; Protein Function; non-Coding DNA.

Emdad Khan is with the College of Computer & Information Science, Imam University , P.O. BOX 5702, Riyadh , Saudi Arabia. He is also with InternetSpeech, Inc, San Jose, CA, USA (Phone: 408-532-9630, fax: 408-274-8151, email: emdad@internetspeech.com)

I. INTRODUCTION

THE advent of DNA sequencing methods has greatly accelerated biological and medical research and discovery. The DNA sequencing cost has come down significantly along with the time to complete it. Knowledge of DNA sequences has become indispensable for basic biological research, and in numerous applied fields such as diagnostic, biotechnology, forensic biology, and biological systematics. The rapid speed of sequencing attained with modern DNA sequencing technology has been instrumental in the sequencing of complete DNA sequences, or genomes of numerous types and species of life, including the human genome and other complete DNA sequences of many animals, plants, and microbial species [1].

What are the next key problems in biological systems / processes? There are 3 broad classes, namely, **analyzing & understanding biological systems, modeling biological systems / processes, and analytics.** Key problems under the **"**analyzing & understanding biological systems" are: understanding Gene Expression, Expression Pathway, Gene Function, Protein Function, Protein Interactions, Protein Scaffolding, Metabolism and the like. Modeling biological processes / systems is the key to address the issues under the first category. In fact, use of computational modeling is at the heart of systems biology. Although significant advancements have been made in modeling biological systems, it has long way to go. Today, there is no reliable and complete way to model a genetic network (e.g. Circadian clocks that provide endogenous cellular rhythms of approximately 24 hours that control many physiological processes), cells, organs, diseases (e.g. diabetes, cancer,..) and biological systems ([2], [3]). Solving the analytics problem in an automated way is the key as there are vast amount of literatures which is also growing very rapidly. Processing such literatures even at the initial stage of categorizing or grouping would provide a significant help. And, of course, summarization and drawing inferences in an automated or semi-automated way would be of great help in advancing the research.

Bioinformatics helps understanding of biological systems using computer science, especially to understand how information is **represented and transmitted** in biological systems. Bioinformatics is the key to help understand genomics, proteomics, biological processes, system biology, complex diseases (e.g. diabetes, cancer), drug discovery and more. ***Many aspects of computer science become handy including databases & database management, search engines, data visualization, NLU/NLP algorithms, machine learning, data mining, pattern matching, modeling and simulation.***

Due to the very large data size, the issues of Big Data come into play strongly in addressing most of the problems associated with biological systems. Big data in medical research is transforming research from hypothesis-driven to data-driven. Efficient analysis and interpretation of big medical data can open up new avenues to explore, new questions to ask, and new ways to answer, leading to better understanding of diseases and development of better and personalized diagnostics and therapeutics [4]. Thus, addressing big data problems in bioinformatics (and in biology) plays a critical role in turning data into meaningful biological applications and knowledge; thus help addressing all above mentioned three major categories of problems at hand, as well as help advancing the research.

In this paper, we discuss the use of Natural Language Processing (NLP), Natural Language Understanding (NLU) & associated semantics (i.e. SEBLA-NLU) to address above mentioned Big Data based key problems. Our main focus is on the 1st and 3rd categories of problems i.e. **analyzing & understanding biological systems and doing automated analytics using NLU/NLP & Intelligent Agents.** Section II describes Natural Language and Biology. Section III describes "NLP, Biology and Bioinformatics". Section IV describes "Solving Unstructured Big Data" and Section V describes **"**Solving Structured Big Data" in Bioinformatics / Biology. Section VI discusses how NLP (along with semantics) is used as a key element to help understand Biological systems. Section VII discusses "Understanding non-Coding DNA" and Section VIII discusses "NLP and Modeling Biological Systems". Section IX focuses on "Future Works" and Section X provides "Conclusions".

II. NATURAL LANGUAGE AND BIOLOGY

The relationship between Natural Language and biology comes from the fact that biological systems use biological alphabets, words and possibly sentences (e.g. in DNA, RNA and protein) which is similar to our natural language. So, there is possibly some good relationship between the two. In fact researchers found in early 1990s that key natural language features are also present in biological language ([17], [18]). One such feature is the Zipf's law [18]. In Zipf analysis, one calculates the histogram that gives the total number of occurrences of each word in a text. If all the words in the text are arranged in a ranked order, from most frequent to least frequent, then such a histogram is found to be linear on double logarithmic paper with unity slope with all human languages studied. This has been found to be the case in DNA [17] but with a slope of about 0.36 for non-coding DNA and 0.2 for coding-DNA (based on sequences of mammalian origin from GenBank).

The 2nd common features of human languages is redundancy: letters or even words can be omitted, changed or reordered without the text becoming non-decipherable. The notion of redundancy was quantified in the classic work of Shannon [19], who introduced the concept of entropy from which the redundancy can be computed. Such redundancy has also been found by many researchers in DNA, with more redundancy in the non-coding region [17].

The 3rd common feature in human languages is the long range correlation between words and sentences (e.g. discourse – coherence, co-reference). Correlation has been found in the non-coding DNA with large distance e.g. 1000 base pairs apart [17].

All these show that research in biological language using natural language ideas is a very promising area. Understanding biological language can help us really understand many open problems in biology.

However, the key issue not advanced much is the "semantics or meaning" which is an *essential component when we try a computing machine to understand our natural language*. The above findings are based on statistical methods which are great but when it comes to semantics, existing methods do not offer a good solution [12]. While traditional approaches to Natural Language Understanding (NLU) have been applied over the past 50 years and had some good successes mainly in a small domain, results show insignificant advancement, in general, and NLU remains a complex open problem. NLU complexity is mainly related to **semantics**: abstraction, representation, real meaning, and computational complexity.

In ([12], [14]), we have presented a Semantic Engine using Brain-Like Approach (SEBLA) to effectively address the semantic issue. As mentioned in Section I, our focus is to use SEBLA-NLU approach to address the 3 broad classes of biological problem, namely, **analyzing & understanding biological systems, modeling biological systems / processes, and analytics.**

It is important to note that the "Analytics" aspect has become more important with the large growth in biological data in various databases worldwide. Collecting, grouping, processing, analyzing, summarizing and drawing inferences of such data in an automated way is the key to help advance research in biology, bioinformatics, biomedical and other related areas.

III. NATURAL LANGUAGE PROCESSING (NLP), BIOLOGY AND BIOINFORMATICS

To handle Big Data in biology, bioinformatics, biomedical informatics (and Big Data in general), we would need some automated method as it is not possible for human to manually try to process, understand and derive new inferences from such large amount of data. Big Data consists of unstructured (free text data) and structured data (e.g. data in a database). Unstructured data dominates the data world. It is estimated that over 80% data in computers and Internet are unstructured [6]. In case of bioinformatics, the structured data is also very large - e.g. data in MEDLINE and GenBank. Computers are very good in processing structured data. This is mainly because computers are still mathematical devices, especially, fast number crunchers. When it comes to unstructured data, we are dealing with the meaning or semantics and associated context; and humans are very good at that [7]. Semantics is also very key to improve the usage of structured data – in finding relations, extracting new information and connecting / using structured data with unstructured data [8]. Thus, Natural Language Processing (NLP) and associated semantics become very useful in addressing Big data problems in bioinformatics and biology. In fact, use of NLP in biology has been increasing rapidly. A very good description of how NLP is used for Information Management in biology and bioinformatics is provided in [9]. In [10], Semantic MEDLINE integrates information retrieval, advanced natural language processing, automatic summarization, and visualization into a single Web portal. Semantic MEDLINE can make an impact on biomedicine by supporting scientific discovery and the timely translation of insights from basic research into advances in clinical practice and patient care.

It is important to note that although existing NLP approaches have made good progress and simplified the automation process somewhat, they still have not solved the problem of computers' inability to deal with tacit and context-based information. At present, we can conclude that text analysis technology may be better at data reduction than actual data analysis. As already explained, human brain is very good in addressing these problems. In case of bioinformatics, existing methods mainly do information management (information retrieval and information extraction). The capabilities to reliably finding relationships between genes / proteins, generating specific predictions that pertain to gene function, predicting essential genes, and finding correct interactions are limited. E.g. co-occurrence of gene and protein names in abstracts implies a biological relationship. But in many cases co-occurrences are not indicative of interaction. Negation is one trivial reason (e.g. A was found not to interact with B [9]). Use of controlled vocabulary in today's ontology is another key limitation. E.g. an author may refer to "type II diabetes mellitus" but an ontology concept may consider this as "diabetes, type II, mellitus" which usually cause major difficulty for a software used to search texts (not a big issue for humans though).

The key point is that we would need to use better semantics and NLU capabilities in dealing with both unstructured and structured data to more reliably and efficiently address such issues. In [8], we proposed to use Semantic Engine using Brain-Like Approach (SEBLA) to convert data to knowledge and also to compress it; thus addressing the Big Data problems in an effective way. SEBLA provides "Natural Semantics" i.e. semantics similar to what humans use (see Section IV for more details). Due to the natural semantics capability of SEBLA, more complex cases can be addressed e.g. in understanding **biological problems** (like Gene Expression, Gene Function, and Protein Scaffolding) and help modeling biological processes / systems (Sections VIII and IX).

Below is a brief description of how NLP with better semantic capability can address various problems including Business Analytics (BI), Information Management, Understanding Biological Systems and Modeling Biological Systems.

3.1 **Analytics**

Analytics, in general, is a process to analyze large data, discover meaningful patterns and then draw some inferences as well as do summarization. It is usually done for business intelligence (BI). But, the same concept can be applied in biology and bioinformatics to do Research Intelligence (RI) i.e. similar to Business Intelligence. In addition to using NLP for information management to retrieve and extract important information, we also need to do summarization and draw some good inferences from large biological data. This also includes filling some structured data tables (e.g. tables in a database) using relevant data from vast amount of text data. Understanding key research issues, research trends etc are important to advance the research more effectively. The same can be applied to medical, biomedical, biological and bioinformatics business intelligence.

The need for Analytics is increased significantly due the rapid data growth in Bioinformatics and Biology (and many other areas as well). For example, U.S. healthcare industry

alone had generated 150 exabytes (2^18) of data by 2011. Using such large data sets - so called **big data** - has become a critical issue providing both **challenges and opportunities**. There are multiple problems with big data including storage, search, transfer, sharing, analysis, processing, viewing, deriving meaning / semantics, and drawing inference / converting data to knowledge. Hence, the need to solve these key problems related to Big Data in a practical and effective way is becoming very important.

Converting Big Data to "Knowledge" is becoming increasingly important to get real benefits from Big Data. It is claimed that U.S. healthcare industry alone can save $450 billion a year with the help of advanced analytics. Our SEBLA-NLU approach can effectively address the key problems of big data in bioinformatics and biology.

3.2 Information / Knowledge Retrieval, Extraction and Integration from various sources

There are various sources for genomics and proteomics information. In general, such sources use different styles, formats even though most use common ontology like in Genome Ontology (GO). Correctly retrieving, extracting and integrating information from such sources is the key to better analyze, understand and derive new information. This mainly belongs to information management (i.e. information retrieval, extraction and associated alignment). NLP has made great progress in this area, especially, exploring and managing biomedical literature [9]. The flood of sequence information produce by the rapid advances in genomics and proteomics is a key driver in bringing the use of NLP to bioinformatics. The fact that so many texts and sequences are available now electronically, it is clear that NLP becomes an obvious choice of extracting key information from such vast sources.

From information management standpoint, NLP has 3 aspects: information retrieval, information extraction and semantics. Information retrieval refers to the recovery of documents from databases related to user's query (e.g. use of PubMed to find documents about a topic). Search from the Internet and databases can be grouped under Information Retrieval. The goal is to find the most related information to the query. This is probably the most common use of NLP today. Existing information retrieval methods are mainly based on string matching.

Information extraction is the process of retrieving some meaning from a text – for example, finding protein-protein interaction from MEDLINE. String based extraction is not useful to extract meaning, hence technologies like ontologies, parsing (syntactic and semantic) and regular expressions are needed.

Semantics (i.e. the meaning of words and sentences) is the critical element for information extraction. It is also an important element for much better information retrieval. Semantic search can provide much more relevant and much concise search results. However, semantics based on existing methods (e.g. ontologies) may not produce key information for many cases as just structural relationships between words do not convey the core meaning in many cases (refer to Sections IV and V for more details). As mentioned, natural semantics based semantic engine SEBLA can improve information extraction and retrieval in a major way.

3.3 Understanding Biological system

Information retrieval, extraction and integration from various sources using vast amount of data (Section 3.2) is very important to automatically process Big Data, and help understanding of biological systems by the researchers mainly from a **higher level**. Analytics described in Section 3.1 will also help this process.

However, we believe, NLP and NLU with semantics (especially using our proposed SEBLA-NLU approach) can be used to better understand the biological systems and processes at deeper levels – e.g. to understand Gene Function, Gene Expression, Genetic Messages, Expression Pathways, Protein Function, Protein Scaffolding, and Metabolism. This is because biological systems use **biological alphabets** in Genes and Proteins and there is a strong relationship of biological language with natural language as described in Section II. Thus, finding special sequences of such alphabets and words, their relations and drawing some good inferences are keys to understand biological systems.

The key is to use the semantics to understand the meaning of biological language. For example, we clearly understand the coding part of DNA today using the 3-letter word (Codon), its transcription to mRNA (and other RNAs like tRNA, MicroRNA and the like) and finally forming the proteins. However, we do not have clear understanding how Gene Expression works, how Expression Pathway works, what is the specific function of a gene, how genes interact and the like.

With latest research data, these are controlled and managed by the non-Coding part of the DNA – non-Coding part within a gene (introns) as well as non-Coding part in the other part of the DNA (non-gene areas).

While experimental methods are very important in understanding biological systems, it is not possible or not practical to do experiment for everything. Understanding the biological language (in addition to the use of experiments and

other methods) can possibly make this process not only possible and complete but also more practical and effective.

The key, of course, is to determine the biological words (i.e. the DNA words), understand and develop their semantics and hence understand the biological (i.e. the DNA) language.

3.4 Developing Semantics in Biological systems

There is a big caveat for the concept described above in Section 3.3. Biological words (i.e. the DNA words) are not like our natural language words for which we know the complete meaning. Only biological systems know the real meanings and vocabulary of such words.

In the coding region, we know at least one meaning of each codon (the corresponding amino acid), and we also know that the word length of a codon is fixed to 3 letters. But in the non-Coding region, neither the lengths of the words nor their meanings are known (see Section VII for more).

Thus, in understanding and developing semantics, starting with the coding region is a natural choice as it would be easier. However, there are also complexities in the coding region - even if, in general, each codon maps to one amino acid, there are cases where a codon has dual meaning (e.g. CUG may code for serine and leucine [20]). Researchers recently found similar dual meanings for several codons. The meaning can also overlap from other factors – e.g. the same codon can mean an amino for mRNA but a different meaning for tRNA.

The semantics of a codon will include all possible meanings. Similarly, the semantics of a sentence would include multiple meanings as appropriate depending on the words used in the sentence (see Section IV for the corresponding concept in SEBLA).

We believe, we should be able to use SEBLA's natural semantics approach to develop semantics of biological words and then apply it to understand biological systems and processes. This process will also involve the existing available knowledge for such words, analyses and many experiments.

It is important to note that only about 2% of total bases in a gene are used to code proteins. We do not know what exactly the remaining 98% of the gene are doing. As per [16], only about 1% of the three billion letters directly codes for proteins - of the rest, about 25% make up genes and their regulatory elements. The function of the remaining letters is still unclear. Some of it may be redundant information left over from our evolutionary past. Existing methods usually involve comparing new sequences with existing one, discovering structure and function by homology (the existence of shared ancestry between a pair of structures, or genes, in different species) rather than through a true understanding of the biological principles underlying structure and function. We believe such problems can be addressed using SEBLA-NLU principles after developing the semantics in biological systems. If successful, this would also help better understanding of the evolution process.

3.5 Modeling Biological Systems

Modeling biological processes / systems is the key to better understand such processes / systems, do deeper analyses, discover new information and draw valuable inferences. Thus, Modeling is a key component to help understand biological systems as described in Section 3.3.

This will significantly help advance the research, drug discovery, personalized medicine and more. SEBLA-NLU can also play a major role in modeling biological systems as briefly described in Section VIII.

IV. SEMANTICS AND NLU TO ADDRESS UNSTRUCTURED BIG DATA PROBLEMS

The key problems associated with unstructured data are related to the semantics of words, sentences and paragraphs. As mentioned, human brain uses semantics and natural language understanding (NLU) to very efficiently use unstructured data. Below, first we briefly describe a Semantic Engine ([11], [12]) using Brain-Like algorithms (SEBLA). Then we show how SEBLA can handle Big Data in bioinformatics.

4.1 Semantic Engine Using Brain-Like Approach (SEBLA)

While NLP / NLU are widely used, their success so far have been mainly in a small domain. For large domain and from semantic standpoint, NLU remains a complex open problem. NLU complexity is mainly related to **semantics**: abstraction, representation, real meaning, and computational complexity. We argue that while existing approaches are great in solving some specific problems, they do not seem to address key Natural Language problems in a practical and natural way. In [14], we proposed a Semantic Engine using **Brain-Like approach (SEBLA)** that uses Brain-Like algorithms to solve the key NLU problem (i.e. the semantic problem) as well as its sub-problems.

The main theme of our approach in SEBLA is to use each word as object with all important features, most importantly the semantics. In our human natural language based communication, we understand the meaning of every word even when it is standalone i.e. without any context. Sometimes a word may have multiple meanings which get resolved with the context in a sentence. The next main theme is to use the

semantics of each word to develop the meaning of a sentence as we do in our natural language understanding as human. Similarly, the semantics of sentences are used to derive the semantics or meaning of a paragraph. The 3rd main theme is to use natural semantics as opposed to existing "mechanical semantics" of Predicate logic or Ontology or the like.

A SEBLA based NLU system is able to:

1. Paraphrase an input text.
2. Translate the text into another language.
3. Answer questions about the content of the text.
4. Draw inferences from the text.

As an example, consider the following sentence:

"Maharani serves vegetarian food."

Semantics represented by existing methods, e.g. Predicate Logic, is

Serves(Maharani, Vegetarian Food) and
Restaurant(Maharani)

Now, if we ask
"is vegetarian dishes served at Maharani?"

the system will not be able to answer correctly unless we also define a semantics for "Vegetarian Dish" or define that "food" is same as "dish" etc. This means, almost everything would need to be clearly defined (which is what is best described by "mechanical semantics"). But with SEBLA based NLU, the answer for the above question will be "Yes" without adding any special semantics for "Vegetarian Dish".

The "mechanical semantics" nature becomes more prominent when we use more complex predicates e.g. when we use universal and existential quantifies, and/or add constructs to represent time.

It is important to note that ML (Maximum Likelihood) based performance commonly used in prediction (e.g. when one types words in a search field on a search engine it shows the next word(s) automatically) will be improved with natural semantics. Currently, mainly ML (and sometimes other techniques including existing semantics methods) is used for prediction. By using proposed more natural semantics, the meaning of the typed words will be more clear; thus helping better prediction of the next word(s). It will also help using natural sentences in the search field than special word combinations, e.g. when using advanced search.

Although above example shows the issue of existing semantics using a Question & Answer type system, the same applies for almost all cases including information retrieval, search and information extraction.

4.2 Using SEBLA to Handle Unstructured Big Data

To handle unstructured Big Data, an Intelligent Agent (IA) is used that utilizes semantics of SEBLA and NLU in various ways depending on the task. The Big Data tasks from biological context can be broadly classified as:

a. Information Retrieval (IR) / Search
b. Information Extraction
c. Question & Answer
d. Summarization
e. Converting data to information to knowledge to intelligence

[*Note: as mentioned above, semantics and NLU/NLP are also important to understand and model biological systems – these aspects are described in Sections VIII and IX*]

Note that all these do significant data compression that helps other key features of Big Data including storage, processing, and visualizing. E.g. in IR, instead of retrieving all information using string search, SEBLA will reject all information that is not related semantically i.e. it will retrieve information that are related semantically.

For the key tasks of IA, let's consider the case of a Q & A System. The key tasks for this case are:

1. Understand user's request and break it into key component parts.
2. Act on all the component parts, find requested answers by accessing appropriate sources (including database tables).
3. Assemble a concise answer, and then present it in a nice way.

The IA itself also uses SEBLA's natural semantic engine to make correct decisions by avoiding "mechanical semantics", as commonly used in existing systems. Such an IA for Q & A system (IAQA) is shown in Fig. 1.

The term "**rendering**" ([12], [13]) needs some explanation. As we know, the Internet was designed with visual access in a relatively large display screen (like a 8.5 inch x 11 inch page) in mind. Thus, all the content are laid out on any website and webpage in a manner that attract our eyes in a large screen. Retrieving the desired content (which is much smaller in size than the total content on a webpage or website) from a typical webpage / website and displaying that (or playing in audio) into a much smaller screen (like in a cell phone or PDA) is a

very challenging task. This process of retrieving and converting most desired content from a large source of content into a much smaller but desired content is called **"rendering".** Clearly, rendering is mainly related to Internet Browsing on a small device. A Q & A system uses rendering to get an initial answer and then further refines it with semantics. Rendering includes form rendering, retrieving appropriate data when a form is submitted, and retrieving multi-media data. A Q & A system also uses rendering to get appropriate data from various websites, via web services and other query methods.

V. SEMANTICS AND NLU TO ADDRESS LARGE STRUCTURED DATA

Structured data are much smaller in size compared to unstructured data and computers can handle structured data well. Thus, it may appear that the need to more efficiently address structured data is not that critical. While this perception is partially true, the need to more efficiently address structured data is also very important. The key reasons are:

a. Structured data are already very large for bioinformatics / biology and also growing very fast. Conventional algorithms are not sufficient for many cases.
b. Many Big Data applications, although are dominated by unstructured data still needs structured data (e.g. analysis report in a BI)
c. Meanings of structured data are critical to process them effectively and efficiently.

Thus, most of the issues related to unstructured data are also equally applicable for structured data. Accordingly, semantics and NLU can be efficiently applied for structured data.

Let's take an example of relationships between various data fields in various tables in a database (e.g. MEDLINE, GenBank). Today's approach using database programming (e.g. using a set of SQL queries and some associated conclusions) becomes difficult when relationship size and data size grow. Besides, such relationships are defined "mechanically" sometimes using "mechanical semantics" as explained for unstructured data.

In contrast, let's consider that data table headings have natural words or sentences. Using the semantics of such words or sentences, it would be much easier to express such relationships. Moreover, semantics will enable to define many complex relationships that cannot be defined currently. Via appropriate data-mining & other techniques and the use of semantics, a significant data compression will also be possible.

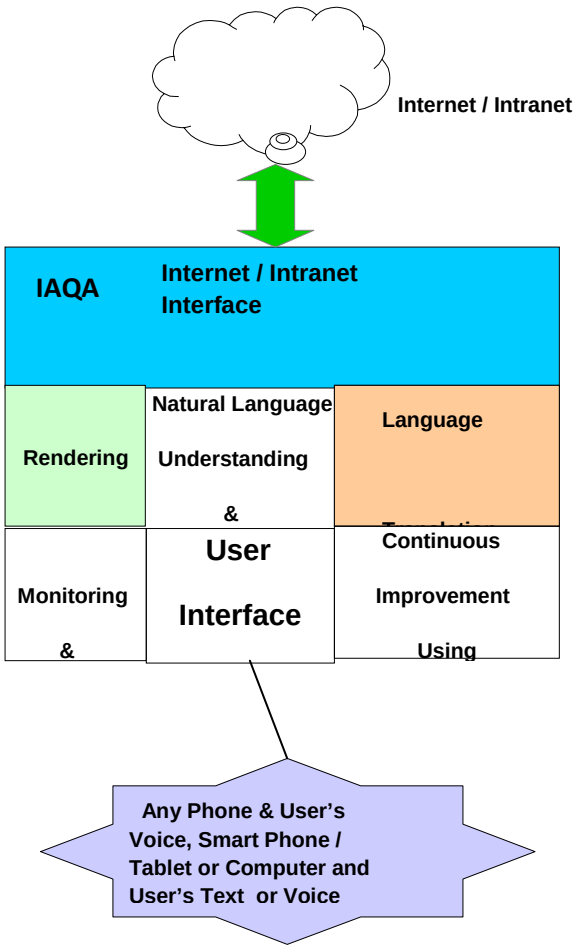

Fig. 1 IAQA: Intelligent Agent for a Question & Answer (Q & A) System.

VI. NLP IS A KEY ELEMENT TO HELP UNDERSTAND BIOLOGICAL SYSTEMS

The use of NLP to help understand biological systems and processes is already described in Section 3.3. There are two broad categories:

a. Use Big Data to understand at a higher level. This is basically automatic use of Big Data inferences by the researchers. Due to the nature of Big Data and the information that can be inferred, this can be a great contributor to researchers to better understand biological systems.
b. Applying NLP / NLU concept to biological language consisting of biological alphabets in genes (A, T, C, G), proteins (ALA, ARG, ASP,….), words and sentences.

Use of Big Data to better understand biological systems at a higher level is explained in Section 3.1. Application of NLP / NLU concept (especially SEBLA-NLU) to help understand biological systems / processes at deeper levels is mentioned in Section 3.3. However, it needs more explanation. At the first level, basics of NLP (e.g. Regular Expressions, String processing, String search, and pattern analysis) can be used to retrieve new information from large data set. This will help in finding similar genes, finding closest neighbor of a new gene, what specific patterns in gene sequence results 3-D shape of proteins and the like.

The next level is determining the real meaning of the genetic words and sentences (sequence of words) using the semantics. Semantics is also needed to understand the correlation between DNA words, especially at many base-pair distance apart. This will help us to really understand the genetic messages, how biological subsystems and systems work. It will help us to understand the general complete biological process (equation (1)) i.e.

Genetic Information -> Molecular Structure -> Biochemical Function -> Biological Behavior ……. (1)

It will also possibly help to understand the major part of the gene (about 74%, [16]) that is not understood yet. However, as discussed in Section 3.4, we would need to develop the semantics first which may be a daunting task. But it is surely worth pursuing.

VII. UNDERSTANDING NON-CODING DNA

Initially, a large proportion of noncoding DNA had no known biological function and was therefore sometimes referred to as "**junk DNA**", particularly in the lay press. However, it has been known for decades that many noncoding sequences are functional. These include genes for functional RNA molecules and sequences such as origins of replication, centromeres, and telomeres.

Some noncoding DNA is transcribed into functional noncoding RNA molecules (e.g. transfer RNA, ribosomal RNA, and regulatory RNAs), while others are not transcribed or give rise to RNA transcripts of unknown function. The amount of noncoding DNA varies greatly among species. For example, over 98% of the human genome is noncoding DNA, while only about 2% of a typical bacterial genome is noncoding DNA.

Some non-Coding DNA may have no biological function for the organism, such as endogenous retroviruses. However, many types of noncoding DNA sequences do have important biological functions as mentioned above. The Encyclopedia of DNA Elements (ENCODE) project suggested in September 2012 that over 80% of DNA in the human genome "serves some purpose, biochemically speaking". In fact a recent finding by EMBO (European Molecular Biological Organization) says

"The data from our experiments show that genome-wide changes in the expression levels of small non-coding RNAs in the first exons of protein-coding genes are associated with breast cancer" [21].

The key question is "how to understand the non-coding DNA"? Non-Coding region use the same DNA letters (i.e. A, T, C, G). The word size varies from 2 letters to 8 letters as per some recent studies.

A sample, DNA sequence in non-Coding region is

GCAAAACCGCGATTATCATGCTTC

To understand the words and the sentences, we would need to determine the semantics of each word (as we would need to do for 3-letter codon in the coding region – see Sections 3.3 & 3.4) as well as sentences. However, we would need to determine the length of each word and its meaning. This step will be more difficult than for words and sentences in the coding region. Because of the coupling of the non-coding region with the coding region (as non-coding region also controls the functions of the coding region), determining the semantics will be even more difficult. We would need to use more extensively the existing available knowledge for such words, analyses and many experiments.

We believe, we should be able to use SEBLA's natural semantics approach to help develop / refine semantics of the non-Coding DNA words and sentences once we can develop semantics of the basic words using the approach mentioned above.

VIII. MODELING BIOLOGICAL SYSTEMS

Modeling biological systems is very challenging for various reasons including:

a. Large range of spatial scales
b. Large range of temporal scales
c. A lack of separation between response to external stimuli versus internal programs
d. Multiple functionalities of constraints
e. Multiple levels of signal processing
f. Incomplete evolutionary record
g. Wide range of sensitivities to perturbations
h. Genotypic variations

There are also challenges from experimental aspects – e.g. reproducibility, spatial resolution, temporal resolution, cross-validation, combinatorial perturbations, accuracy.

From a knowledge perspective, there are the following 4 central problems [3]:

1. Find an appropriate level of abstraction for a given analytic problem
2. Find a common basis to relate knowledge gained using different experimental techniques on the same system
3. Find a common basis to relate knowledge gained from the same experiment on different model systems
4. Incorporate knowledge incrementally as new data is analyzed

There are also computational limitations due to combinatorial explosion, unsolvable equations, and incomplete models.

Hence, modeling biological systems is in fact very challenging. Although existing modeling approaches have been used successfully for many cases, many problems cannot be modeled at all (e.g. Circadian Oscillators – formal logic or traditional analytical tools of molecular biology or macroscopic descriptors such as differential equations cannot model Circadian Oscillators), and many can be modeled only partially with some approximations (e.g. Chemical master equation).

Existing key equations (e.g. CME - chemical Master equation) use many variables, discrete and stochastic approaches, which are good BUT such equations are not solvable – neither analytically nor numerical way (mainly due to combinatorial explosion). Hence, several approximations (e.g. using ensemble averages, assuming unimolecular reactions and assuming no fluctuations) are made so that existing methods such as deterministic ODE (Ordinary Differential Equations) or SDE (Stochastic Differential equations) can be successfully used to solve such problems usually in a small scale.

The key point is that such approximations usually lose a lot of key information and do not / may not represent the real biological systems/processes.

Of course, existing algorithms and methods are great as they have been successfully used in solving many real world problems. However, we got to keep in mind that such algorithms were developed mainly to solve *non-life* based physical world problems, especially, focusing in Physics and Chemistry. Although at the *molecular/atomic level quantum mechanics* plays important role, rich quantum mechanics have been developed and there are many similarities with Physics and Biology at the molecular/atomic level, *biological cells and systems are still far different from non-biological physical systems, as these have life and life processing elements.* For example, protein molecules in a cell work like highly Intelligent Agents - they can automatically understand and determine which specific molecular / chemical reactions to start & process (i.e. what specific reactions will take place between N molecules out if M molecules – the choice is usually very large - within certain time), control the rate of such equations and much more [22]. Such molecules understand the messages coming from DNA sentences (including the messages from non-Coding DNAs). Our existing computational model (e.g. Turing Machine) does not change the speed of computation based on the input from the tape, nor it changes the number of states based on the input.

Moreover, much of the logic of interactions in a living system is implicit. Whenever possible, nature leaves the interactions to the chemical properties of the molecule themselves and to the highly serendipitous way in which these properties have been exploited during evolution as nature has plundered its treasure chest of old genes to recruit new functions [3]. Of course, genetic code is used to get the main information but such information get translated into something that proteins can use to do their functions – they may create self-assembled programs.

Thus, to correctly model such molecules, we would need to clearly understand their behavior and the DNA messages they use. We need to develop the theory of interactions between proteins. The same is true to model the cells, tissues, organs and biological systems which are, of course, more complex. Thus, we have a long way to go.

Hence new paradigms are needed to better model biological systems and processes.

One such paradigm is to understand the biological language (i.e. the DNA language) by using the semantics of biological words and sentences as described in Sections 3.3 & 3.4. We believe our proposed SEBLA-NLU approach can contribute a lot in this process (see Section IX for more).

IX. FUTURE WORKS

We plan to develop a complete BI (business Intelligence) / RI (Research Intelligence System) using SEBLA based NLU (SEBLA-NLU). We also plan to develop semantics of biological basic words and sentences (in collaboration with others) by using the knowledge of how biological systems work (as much as we know today), associated Big Data and new experiments. We will then apply such findings and NLU

(a) to better understand how the biological systems and processes work via the semantics that will be developed.

(b) we will also try to model biological systems using the understanding developed via semantics.

(a) and (b) will help each other to further refine, and better understand as well as to better model biological systems and processes.

X. CONCLUSIONS

We have emphasized that use of the concept of our natural language along with associated semantics is the key to understand biological systems and processes as biological systems use biological alphabets (DNA alphabets), words and sentences (sequence of words) similar to our natural language.

The concept of our natural language fits well with biological language (the DNA language) from all key aspects, namely, Zipf analysis, redundancy and discourse – coherence and co-reference (from statistical standpoint). We believe this will be true for semantics (and hence meaning) of biological language. An important point to note is that we know the semantics / meaning of words and sentences in our natural language, but we do not know such semantics for the words and sentences used in biological systems and processes.

We believe we can develop the semantics of biological basic words and sentences (in both coding and non-Coding DNA regions) by using the knowledge of how biological systems work (as much as we know today), associated Biological / Bioinformatics Big Data and new experiments. We can then apply such findings and NLU approach to better understand how the biological systems and processes work via the use of semantics.

We have presented Semantic Engine using Brain-Like Approach (**SEBLA**) and associated Natural Language Understanding (**NLU**) based approach (SEBLA-NLU) to help develop the semantics of biological words (DNA words) and understand the biological language. SEBLA-NLU also addresses the key problems of Big Data in bioinformatics and biology. We have used human Brain-Like and Brain-Inspired algorithms as humans can significantly compress the data, preserve core meaning, extract latent information, and convert information to knowledge and intelligence. Thus, Brain-Like approach very effectively converts data to knowledge and also compresses it; and hence addresses the key Big Data problems in an effective way. We have presented how SEBLA-NLU is used to handle both unstructured and structured data for addressing complex problems including **analytics**, understanding **biological systems/processes** (e.g. Gene Expression, Gene Function, Protein Function, Protein Scaffolding and metabolism), and more completely, reliably and effectively **modeling biological systems/ processes**.

Our efforts to build the semantics of biological words and sentences, if successful, would enable us to not only understand how biological systems / processes really work but also to understand the evolution and other hidden functions / processes as the functions of about 74% of bases in a gene would be understood.

REFERENCES

[1] Wikipedia – "DNA Sequencing" – http://en.wikipedia.org/wiki/DNA_sequencing.

[2] R. Schwartz, "Biological Modeling and Simulation", ISBN 978-0-262-19584-3, MIT Press, 2008.

[3] Z. Azallasi et al, "System Modeling in Cellular Biology", ISBN 978-0-262-19584-5, MIT Press, 2008.

[4] Big Data Initiative by U.S. President Obama - http://www.whitehouse.gov/blog/2013/04/23/big-data-big-deal-biomedical-research.

[5] C. Eaton et al, "Understanding Big Data: Analytics for enterprise class Hadoop and Streaming Data", http://public.dhe.ibm.com/common/ssi/ecm/en/iml14296usen/IML14296USEN.PDF

[6] Wikipedia – "Big Data" - http://en.wikipedia.org/wiki/Big_data

[7] P. Ryan et al, "The Problem of Analyzing Unstructured Data", Grant Thoronton, 2009, http://www.grantthornton.ie/db/Attachments/Publications/Forensic_&_inve/Grant%20Thornton%20-The%20problem%20of%20analysing%20unstructured%20data.pdf

[8] E. Khan, "Addressing Big Data Problems using Semantics and Natural Language Understanding", 12th WSEAS International \ Conference on TELECOMMUNICATIONSand INFORMATICS (TELE-INFO '13) in Baltimore, MD, USA, September 17-19, 2013.

[9] M. Yandell et al, "Genomics and Natural Language Processing", Nature Reviews (Genetics), Vol. 3, Aug 2002.

[10] H. Kilicoglu et al, "Semantic MEDLINE: A Web Application for Managing the Results of PubMed Searches", Journal of Information Services and Use, IOS Press, Vol. 31, #1-2, Aug 11, 2011.

[11] E. Khan, "Processing Big Data with Natural Semantics and Natural Language Understanding using Brain-Like Approach", INTERNATIONAL JOURNAL of COMPUTERS AND COMMUNICATIONS, (NAUN & UNIVERSITY PRESS), January 2014.

[12] E. Khan, " Intelligent Internet: Natural Language and Question & Answer based Interaction", INTERNATIONAL JOURNAL of COMPUTERS AND COMMUNICATIONS, (NAUN & UNIVERSITY PRESS) Oct. 2013.

[13] Internet for Everyone - Reshaping the Global Economy by Bridging the Digital Divide", Book - ISBN 978-1-4620-4251-7 (SC ISBN) 978-1-4620-4250-0 (HC ISBN), Aug 2011.

[14] Khan, E., (2011): Natural Language Understanding Using Brain-Like Approach: Word Object and Word Semantic Based Approaches help Sentence Level Understanding. A Patent Filed in US in 2011.

[15] D. Brutlag et al, "Understanding Human Genome", Scientific American: Introduction to Molecular Medicine, 1994.

[16] "DNA Molecule: How Much DNA Codes for Protein?" http://www.dnalc.org/resources/3d/09-how-much-dna-codes-for-protein.html, April2, 2010.

[17] R. N. Mantegna et al, "Linguistic Features of DNA Sequences", Physical Review Letters, Vol. 73, No. 23, December, 1994.

[18] G. K. Zipf, "Human Behavior and the Principle of Least Effort", Addison-Wesley Press, Cambridge, MA, 1949.

[19] C. E. Shannon, Bell System Tech. Journal, 27-379 (1948); 30, 50 (1951).

[20] Crit Rev Biochem Mol Biol. 2010 Aug;45(4):257-65. doi: 10.3109/10409231003786094. Dual functions of codons in the genetic code. Lobanov AV[1], Turanov AA, Hatfield DL, Gladyshev VN.

[21] EMBO (European Molecular Biological Organization) on Feb 19, 2014 - http://www.sciencecodex.com/small_noncoding_rnas_could_be_warning_signs_of_cancer-128030 .

[22] Molecular Biology of the Cell, 4th Edition, http://www.ncbi.nlm.nih.gov/books/NBK26911/