Parallels between Linguistics and Biology

Ashish Vijay Tendulkar
IIT Madras
Chennai-600 036. India.
ashishvt@gmail.com

Sutanu Chakraborti
IIT Madras
Chennai-600 036. India.
sutanu@cse.iitm.ac.in

Abstract

In this paper we take a fresh look at parallels between linguistics and biology. We expect that this new line of thinking will propel cross fertilization of two disciplines and open up new research avenues.

1 Introduction

Protein structure prediction problem is a long standing open problem in Biology. The computational methods for structure prediction can be broadly classified into the following two types: (i) Ab-initio or de-novo methods seek to model physics and chemistry of protein folding from first principles. (ii) Knowledge based methods make use of existing protein structure and sequence information to predict the structure of the new protein. While protein folding takes place at a scale of millisecond in nature, the computer programs for the task take a large amount of time. Ab-initio methods take several hours to days and knowledge based methods takes several minutes to hours depending upon the complexity. We feel that the protein structure prediction methods struggle due to lack of understanding of the folding code from protein sequence. In larger context, we are interested in the following question: Can we treat biological sequences as strings generated from a specific but unknown language and find the rules of these languages? This is a deep question and hence we start with baby-steps by drawing parallels between Natural Language and Biological systems. David Searls has done interesting work in this direction and have written a number of articles about role of language in understanding Biological sequences(Searls, 2002). We intend to build on top of that work and explore further analogies between the two fields.

This is intended to be an idea paper that explores parallels between linguistics and biology that have the potential to cross fertilization two disciplines and open up new research avenues. The paper is intentionally made speculative at places to inspire out-of-the-box deliberations from researchers in both areas.

2 Analogies

In this section, we explore some pivotal ideas in linguistics (with a specific focus on Computational Linguistics) and systematically uncover analogous ideas in Biology.

2.1 Letters

The alphabet in a natural language is well specified. English language has 26 letters. The genes are made up of 4 basic elements called as nucleotide: adenine (A), thymine (T), cytosine (C) and guanine (G). During protein synthesis, genes are transcribed into messenger RNA (mRNA), which is made up of 4 basic elements: adenine (A), uracil (U), cytosine (C) and guanine (G). mRNA is translated to proteins that are made up of 20 amino acids denotes by the following letters: {A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y}.

2.2 Words

A word is an atomic unit of meaning in a language. When it comes to biological sequences, a fundamental problem is to identify words. Like English, the biological language seems to have a fixed alphabet when it comes to letters. However, unless we have a mechanism to identify atomic "functional" units, we cannot construct a vocabulary of biological words.

The first property of a word in NL is that it has a meaning; a word is a surrogate for something in the material or the abstract world. One central question is: how do we make machines understand meanings of words? Humans use dictionaries which explain meanings of complex words

Proceedings of the 2013 Workshop on Biomedical Natural Language Processing (BioNLP 2013), pages 120–123,
Sofia, Bulgaria, August 4-9 2013. ©2013 Association for Computational Linguistics

in terms of simple ones. For machines to use dictionaries, we have two problems. The first is, how do we communicate the meaning of simple words (like "red" or "sad")? The second is, to understand meanings of complex words out of simple ones, we would need the machine to understand English in the first place. The first problem has no easy solution; there are words whose meanings are expressed better in the form of images or when contrasted with other words ("orange" versus "yellow"). The second problem of defining words in terms of others can be addressed using a knowledge representation formalism like a semantic network. Some biological words have functions that cannot be easily expressed in terms of functions of other words. For the other words, we can define the function (semantics) of a biological word in terms of other biological words, leading to a dictionary or ontology of such words.

The second property of a word is its Part of Speech which dictates the suitability of words to tie up with each other to give rise to grammatical sentences. An analogy can be drawn to valency of atoms, which is primarily responsible in dictating which molecules are possible and which are not. Biological words may have Parts of speech that dictate their ability to group together to form higher level units like sentences, using the composition of functions which has its analog in compositional semantics. The third property of a word is its morphology, which is its structure or form. This refers to the sequence of letters in the words. There are systematic ways in which the form of a root word (like sing) can be changed to give birth to new words (like singing). Two primary processes are inflection and derivation. This can be related to mutations in Biology, where we obtain a new sequence or structure by mutating the existing sequences/structures.

3 Concepts

Effective Dimensionality: The Vector Space Model (VSM) is used frequently as a formalism in Information Retrieval. When used over a large collection of documents as in the web, VSM pictures the webpages as vectors in a high dimensional vector space, where each dimension corresponds to a word. Interestingly, thanks to strong clustering properties exhibited by documents, this high dimensional space is only sparsely populated by real world documents. As an example to illustrate this, we would not expect a webpage to simultaneously talk about Margaret Thatcher, Diego Maradona and Machine Learning. Thus, more often than not, the space defined by intersection of two or more words is empty. The webspace is like the night sky: mostly dark and few clusters sprinkled in between. In IR parlance, we say that the effective dimensionality of the space is much less than the true dimensionality, and this fact can be exploited cleverly to overcome "curse of dimensionality" and to speed up retrieval. It is worth noting that the world of biological sequences is not very different. Of all the sequences that can be potentially generated, only a few correspond to stable configurations.

Ramachandran plot is used to understand constraints in protein conformations (Ramachandran, 1963). It plots possible $\phi - \psi$ angle pairs in protein structures based on the van der Waal radii of amino acids. It demonstrates that the protein conformational space is sparse and is concentrated in clusters of a few $\phi - \psi$ regions.

3.1 Machine Translation

Genes and mRNAs can be viewed as strings generated from four letters (A,T,C,G for genes and A,U,C,G for mRNAs). Proteins can be viewed as strings generated from twenty amino acids. In addition proteins and mRNAs have corresponding structures for which we do not even know the alphabets. The genes are storing a blue-print for synthesizing proteins. Whenever the cell requires a specific protein, the protein synthesis takes place, in which first the genes encoding that protein are read and are transcribed into mRNA which are then translated to make proteins with relevant amino acids. This is similar to writing the same document in multiple languages so that it can be consumed by the people familiar to different languages. Here the protein sequence is encoded in genes and is communicated in form of mRNA during the synthesis process. Another example is sequence and structure representations of protein: Both of them carry the same information specified in different forms.

3.2 Evolution of Languages

Language evolves over time to cater to evolution in our communication goals. New concepts originate which warrant revisions to our vocabulary. The language of mathematics has evolved to make communication more precise. Sentence structures evolve, often to address the bottlenecks faced by native speakers and second language learners. English, for example, has gone out of fashion. Thus there is a survival goal very closely coupled to the environment in which a language thrives that dictates its evolution. The situation is not very different in biology.

Scientific community believes that the life on the Earth started with prokaryotes[1] and evolved into eukaryotes. Prokaryotes inhibited earth from approximately 3-4 Billion years ago. About 500 million years ago, plant and fungi colonized the Earth. The modern human came into existence since 250,000 years. At a genetic level, new genes were formed by means of insertion, deletion and mutation of certain nucleotide with other nucleotides.

3.3 Garden Path Sentences

English is replete with examples where a small change in a sentence leads to a significant change in its meaning. A case in point is the sentence "He eats shoots and leaves", whose meaning changes drastically when a comma is inserted between "eats" and "shoots". This leads to situations where the meaning of a sentence cannot be composed by a linear composition of the meanings of words. The situation is not very different in biology, where the function of a sequence can change when any one element in the sequence changed.

3.4 Text and background knowledge needed to understand it

Interaction between the "book" and the reader is essential to comprehension; so language understanding is not just sophisticated modeling of interaction between words, sentences and discourse. Similarly the book of life (the gene sequence) does not have everything that is needed to determine function; it needs to be read by the reader (played by the CD player). This phenomenon is similar to protein/ gene interaction. Proteins/genes possess binding sites, that is used to bind other proteins/genes to form a complex, which carry out the desired function in the biological process.

3.5 Complexity of Dataset

Several measures have been proposed in the context of Information Retrieval and Text Classification which aim at capturing the complexity of a dataset. In unsupervised domains, a high clustering tendency indicates a low complexity and a low clustering tendency corresponds to a situation where the objects are spread out more or less uniformly in space. The latter situation corresponds to high complexity. In supervised domains, a dataset is said to be complex if objects that are similar to each other have same category labels. Interestingly, these ideas may apply in arriving at estimates of structure complexity. In particular, weak structure function correspondences would correspond to high complexity.

3.6 Stop words (function words) and their role in syntax

Function words such as articles, prepositions play an important role in understanding natural languages. On the same note, function words exist in Biology and they play various important roles depending on the context. For example, Protein structures are made up of secondary structures. Around 70% of these structures are α-helix and β-strands which repeat in functionally unrelated proteins. Based on this criterion, α-helix and β-strands can be categorized as functional words. These secondary structures are important in forming protein structural frame on which functional sites can be mounted. At genomic level, as much as 97% of human genome does not code for proteins and hence termed as junk DNA. This is another instance of function word in Biology. Scientists are realizing off late some important functions of these junk DNA such as their role in alternative splicing.

3.7 Natural Language Generation

Natural Language Generation (NLG) is complementary to Natural Language Understanding (NLU), in that it aims at constructing natural language text from a variety of non-textual representations like maps, graphs, tables and temporal data. NLG can be used to automate routine tasks like generation of memos, letters or simulation reports. At the creative end of the spectrum, an ambitious goal of NLG would be to compose jokes, advertisements, stories and poetry. NLG is carried out in four steps: (i) macroplanning; (ii)microplanning; (iii) surface realization and (iv) presentation. Macroplanning step uses Rhetorical Structure Theory (RST), which defines relations between units of text. For example, the relation cause connects the two sentences: "The hotel was

[1] http://www.wikipedia.org

costly." and "We started looking for a cheaper option". Other such relations are purpose, motivation and enablement. The text is organized into two segments; the first is called nucleus, which carries the most important information, and the second satellites, which provide a flesh around the nucleus. It seems interesting to look for a parallel of RST in the biological context.

Analogously protein design or artificial life design is a form of NLG in Biology. Such artificial organisms and genes/proteins can carry out specific tasks such as fuel production, making medicines and combating global warming. For example, Craig Venter and colleagues created synthetic genome in the lab and has filed a patent for the first life form created by humanity. These tasks are very similar to NLG in terms of scale and complexity.

3.8 Hyperlinks

Hyperlinks connect two or more documents through links. There is an analogy in Biology for hyperlinks. Proteins contain sites to bind with other molecules such as proteins, DNA, metals or any other chemical compound. The binding sites are similar to hyperlinks and enable protein-protein interaction and protein-DNA interaction.

3.9 Ambiguity and Context

An NLP system must be able to effectively handle ambiguities. The news headline "Stolen Painting Found by Tree" has two possible interpretations, though an average reader has no trouble favoring one over the other. In many situations, the context is useful in disambiguation. For example, protein function can be specified unambiguously with the help of biological process and cellular location. In other words, protein functions in the context of biological process and within a particular cellular location. In the context of protein structure, highly similar subsequences take different substructures such as α-helix or β-strand depending on their spatial neighborhood. Moonlighting proteins carry out multiple functions and their exact function can be determined only based on the context.

Let us consider the following example: "Mary ordered a pizza. She left a tip before leaving the restaurant." To understand the above sentences, the reader must have knowledge of what people typically do when they visit restaurants. Statistically mined associations and linguistic knowledge are both inadequate in capturing meaning when the background knowledge is absent. Background knowledge about function and interacting partners about a protein help in determining its structures.

4 Conclusion

In this paper, we presented a number of parallels between Linguistics and Biology. We believe that this line of thought process will lead to previously unexplored research directions and bring in new insights in our understanding of biological systems. Linguistics on other hand can also benefit from a deeper understanding of analogies with biological systems.

Acknowledgments

AVT is supported by Innovative Young Biotechnologist Award (IYBA) by Department of Biotechnology, Government of India.

References

David B. Searls. 2002. The language of genes *Nature*, 420:211–217.

G. N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan. 1963. Stereochemistry of polypeptide chain configurations *Journal of Molecular Biology*, 7:95–99.