Syntactic Structures in Languages and Biology

David Horn
School of Physics and Astronomy, Tel Aviv University, Tel Aviv 69978, Israel

Both natural languages and cell biology make use of one-dimensional encryption. Their investigation calls for syntactic deciphering of the text and semantic understanding of the resulting structures. Here we discuss recently published algorithms that allow for such searches: ADIOS (Automatic DIstillation of Structure) that is successful in discovering syntactic structures in linguistic texts and its MEX (Motif EXtraction) component that can be used for uncovering motifs in DNA and protein sequences. The underlying principles of these syntactic algorithms and some of their results will be described.

Introduction

There exists an interesting analogy between three different kinds of languages: the human languages, the language of computation, and the language of cell biology. In all these cases, language is being realized by one-dimensional constructs. In human language it is speech, with time being this one-dimension. In computation it may be viewed as the tape of the Turing machine, i.e. the ordered code into which all computations can be transformed. In biology it is the order of nucleotides in the DNA sequence, and the resulting sequential orders of nucleotides in RNA and amino-acids in proteins.

Viewing the one-dimensional language (e.g. utterances of speech, or written language as composed of letters and symbols) as a code describing some reality, one needs to have rules describing how its words or sentences are connected to that reality (semantics) as well as other rules of how to put them together into a one-dimensional language representation (syntax). These two sets of rules comprise the logic of this language.

The distinction between the semantic and syntactic levels, brought forward in the linguistic study of (Chomsky, 1957), is clearly well suited for biology. Syntax stands for the rules specifying structures of chromosomes (e.g. separation into genes, segmentation of genes into exons and introns) or structures of proteins. Semantics refers to the roles of the different elements, e.g. in guiding the complicated dynamics from transcription, the birth of mRNA, to translation, the birth of the protein. When applied to proteins, syntax should refer to possible structures existing in their sequential description in terms of amino-acids. Semantics may refer to their secondary or tertiary geometrical forms, as well as to the functional roles of different elements in specifying interactions between the protein in question and its biological neighborhood.

In computation the rules are man-made. All computational devices work according to specified rules (software) that are carried out by some relevant hardware, and they are equivalent to a universal computational structure foreseen by (Turing, 1937) and known as the Turing machine. The latter has a code written on a single tape and the machine carries out operations by moving the tape and writing on it. The code leads to a computational result when the machine stops.
Is there an analog of the Turing machine in biological or human languages? It seems that in biology we are still at the beginning of the road. Less than a decade into the post-genomic era, there exist still many open questions on the way to comprehending the full logic from embryo to organism and from the genome of a cell to its protein manufacturing. Linguistic studies for the past fifty years have often concentrated on the concept of ideal machinery that is specific to the

human species. In the Chomsky approach this is the language faculty that endows us all with a Universal Grammar (Nowak et al., 2002). However, this idealistic picture is far from being formulated in well defined rules from which one can deduce actual human languages. In other words, the Turing machine of the human mind still remains an enigma.

The set of rules specifying the syntax of a language in known as its grammar. Although speaking a human language comes naturally, setting up its rules in a comprehensive manner is a formidable task. As human beings we learn through examples, sometimes aided by explicit rules, yet when confronted with a certain sentence and having to decide whether it is grammatically right or wrong, we often rely on intuition. Hopefully one day we will understand the genetic reasons for the special human faculties of speech and language, but we are not there yet. Even when we will know it, this will not explain Universal Grammar. The latter would require extracting logical grammatical rules from the underlying rules of neural circuitry, which seems an impossible task.

Chomsky (1957) has introduced a categorization scheme of formal languages by defining four types of grammars that differ from each other by their rules of generation. The simplest one is defined as Regular Grammar and can be generated by finite-state automata. The next one in hierarchy is the Context Free Grammar whose constructs can be generated by substitution rules of the type $S \rightarrow aSb$ where a and b belong to the lexicon, i.e. are `words' or, in general, `terminals' and S is a non-terminal. They are surpassed in complexity by Context Sensitive Grammars (CSGs, with rules such as $Sc \rightarrow aSb$) and by Unrestricted Grammars. The latter are equivalent to Turing machines. Computational complexity of grammars is defined by the time it takes to decide whether a given string belongs to a given language. This time grows linearly in length of the string in regular grammars, polynomial in length for CFGs, presumably exponential for CSGs and it is undecidable in the general case of unrestricted grammars. For a recent review see (Nowak et al., 2002).

Can biological sequences be characterized by these formal languages? A recent review by (Searls, 2002) points out the many parallels that were drawn in the literature between linguistic methods in general, and formal languages in particular, with different constructs in biological sequences. One should realize, at the outset that it may be much too naive to expect all the genome to fit into one well defined grammar. One may, however, expect that its different functional parts will be susceptible to some specific grammatical constructs or, at least, will be amenable to searches using specific grammatical machinery. For example, there exist interesting non-coding RNA constructs that seem to fit into CSG construction rules.

Motif extraction and grammar induction

Grammar induction is a complex pattern recognition problem, which is being studied by the Machine Learning community (Duda, Hart & Stork, 2001).
Recently (Solan et al., 2005) have introduced a new algorithm that looks for patterns (e.g. combinations of words) in the data and extracts syntactic rules by searching for patterns of patterns, allowing also for the occurrence of equivalence classes in this process.
Their ADIOS (for Automatic DIstillation of Structure) algorithm is being trained on some training set, composed of sentences of a given corpus. It loads these sentences on a graph, whose vertices are the different words of the corpus. The resulting syntactic structure that it finds is exemplified in Fig.1.In 1A we observe how the combination of words 'far away' is made into a pattern (number 67). In 1B, upon reiteration of the algorithm, this pattern becomes part of another pattern (number 101) that includes also an equivalence class (number 98). The way these tree-like structures have to be read is described by numbers that follow the parsing rules from top down

and from left to right. All elements of a pattern have to be chosen (in the given order) and only one of all elements of an equivalence class (denoted here by underlined numbers) has to be chosen. Further iterations of the algorithm eventually replace the original sentence by one pattern (the 'root pattern' in 1C). The original sentence can be read-off at the leaf-level of the tree, and is only one out of many sentences that the root-pattern can generate. This generalization of the model is being tested using standard measures. The novel sentences generated by ADIOS are tested for their grammaticality, thus calculating the precision of the method. In addition, the recall of the ADIOS machine is tested by running a test-set and finding out how many of its sentences can be accounted for by the trees that ADIOS has constructed. Fig. 1D is a standard CFG representation of the tree-structure of 1C.

The algorithm was tested (Solan et al., 2005) on many benchmarks with remarkable success. The first was a subset of the CHILDES collection (MacWhinney and Snow, 1985), which consists of transcribed speech produced by, or directed at, 3-year old children. The ADIOS-students were subjected to the grammaticality judgment test used in English as Second Language (ESL) and, in spite of the fact that the test is aimed at students who have had 6-7 years of studies, ADIOS has reached a success rate of 60%, which is considered adequate.

A second benchmark was the ATIS (Moore and Carroll, 2001) corpus of natural language sentences. This corpus contains 13,043 sentences. ADIOS machines were trained on 12,700 sentences and the rest were used for measuring recall. Human subjects have judged the sentences generated by the ADIOS-trained machines, to be grammatically correct to about the same level as those of the original corpus. There also exists an artificially-constructed ATIS CFG (357 terminals and 4592 rules) (Moore and Carroll, 2001). To learn and reproduce it well (Solan et al., 2005), one had to rely on a group of 150 ADIOS-learners, where each learner has been trained on the same corpus using a different order of the sentences.

An important component of ADIOS is the Motif Extraction algorithm MEX.
It is an unsupervised algorithm that extracts motifs from sets of sequences which may be regarded as strings of letters from some given alphabet.
Each sequence is represented as a path over a graph containing vertices representing the different elements of the alphabet. An example of such a graph is presented in Fig. 1. After uploading all sequences onto the graph, one counts the number of paths connecting vertices in order to define probabilities such as

$p(e_j|e_i)$=(number of paths proceeding from e_i to e_j)/ (total number of paths leaving e_i)

$p(e_k|e_j,e_i)$=(number of paths proceeding from e_i to e_j to e_k) / (number of paths proceeding from e_i to e_j)

for all vertices e_i of the graph. These data-driven probabilities allow for the definition of a position-dependent variable-order Markov model describing the data.
A significant motif that is extracted by MEX is a sub-path along the graph defined by probability based criteria that account for convergence of many paths into the beginning point of a motif, and divergence of many paths from the end-point of the motif, as exemplified in Fig. 1. Motifs are not constrained by length, and may overlap with one another. There are two parameters of MEX: η, specifying a decrease in probability measures that determine convergence and divergence, and α specifying their statistical significance. For more details see (Solan et al 2005) and http://adios.tau.ac.il.

In most linguistic ADIOS studies, the strings used by MEX were the sentences in the studied corpus, and the graph started out with vertices that were the words in the text. Further iterations of ADIOS were implemented by including within the graph vertices that corresponded to patterns, and vertices that corresponded to equivalence classes.
In the biological applications to be described below, MEX has been applied to strings of DNA, where the alphabet consists of four nucleotides, and strings of proteins, whose alphabet consists of 20 amino-acids.

Applications to Biology

In natural languages we are accustomed to think in terms of a single grammar. As far as ADIOS-learning is concerned, this may not be true; i.e. different ADIOS CFGs may result from corpora that differ in their contents. In Biology one may expect different processes to have different grammars associated with them. Examples were pointed out by (Searls, 2002).
Here we wish to discuss two examples of the usefulness of the novel apparatus of MEX: uncovering motifs in DNA and protein sequences.

Transcription factor binding sites

The DNA application concentrates on motif search in the promoter regions of genes. These regions lie upstream to the gene and serve to control its transcription into mRNA. Such control comes about by proteins, known as transcription factors (TFs), which bind to particular loci on the chromosome. The computational challenge is to find subsequences that serve as candidates for transcription factor binding sites (TFBSs). The conventional approach to solve such questions is to select genes that are known to be activated by a particular TF, and search subsequences that are over-represented on their promoters. MEX is an unsupervised algorithm and may be applied to the whole genome. This has the advantage that it can find novel strings of nucleotides (four nucleotides form the alphabet of DNA) unconstrained by present biological knowledge of TF binding properties.

Such a study for the yeast genome has recently been published by (Segal et al. 2007). MEX was applied to all 4483 promoters of 6335 genes revealing a large number of motifs. The latter were then subjected to screening by results of 40 different experiments in which expression analysis of mRNA has been performed for several time points in each experiment. For a motif to serve as the candidate for a TFBS it has been required that the set of genes whose promoters carry the motif in question display coherent activation patterns over the time-points of at least one experiment.
The results have been compared to present knowledge and a few interesting conclusions have emerged: 1. many motifs can be clustered into groups that cover previously known TFBS. 2. single nucleotide changes in cores of clusters (cores are common substrings of motifs in the cluster) may turn out to be very meaningful, leading to different types of expression behavior. 3. novel motifs and novel clusters were uncovered.

TFBS are the analog of words in promoter application. However other similarities to words in a human language disappear. First and foremost is the fact that words do not show up continuously along the string of letters that comprise the DNA. They are interspersed by numerous strings of nucleotides that do not seem to have any particular 'meaning' that we are aware of. Moreover, there exist different TFBS that serve as binding loci for the same TF. Thus a 'word' is not unique and many motifs may have a similar 'meaning'. Nonetheless the analogy with words holds as far as the MEX search is concerned: it has the capability of selecting motifs from strings of DNA that turn out to have important biological roles.

Specific peptides

The second biological application of MEX concerns the search for motifs on the amino-acid sequences (having an alphabet of 20 elements) of proteins. The biological question that is of interest is that of predicting the function of a protein from its sequence. This is conventionally done by relying on sequence similarity that is known to be a good predictor of functional similarity (Tian and Skolnick, 2003). Nonetheless it is of interest to find out whether there exist on the protein's sequence strings of amino-acids, also known as peptides, which can indicate what its function is.

The study by (Kunik et al., 2007) has answered this question affirmatively. It has applied MEX to a large set of proteins known as enzymes, derived from the Swiss-Prot database (release 48.3 in Oct 2005). The catalytic functions of enzymes are classified by the Enzyme Commission (EC) into a 4-dimensional hierarchy forming a tree that branches into 6 classes, many subclasses, etc. MEX was applied to over 50,000 enzymes. The resulting motifs were further filtered by the EC classification, leading to over 50,000 specific peptides (SPs), i.e. strings of amino-acids that appear only on enzymes within a particular branch of the EC hierarchy.

SPs are the analog of 'words' on the amino-acid sequences of enzymes. They can serve as means for functional classification with a high degree of accuracy. On the training set they have coverage of 87% for the full EC number (4^{th} level of the hierarchy) and 93% for the 1^{st} EC level. Their accuracy on a test set is of order 84%. These very high levels mean that SPs can serve as agents for data mining of enzymes. An example of appearances of SPs on the sequence of an enzyme and their use for classification purposes is provided in Fig. 2. This is a screenprint of the webtool http://adios.tau.ac.il/SPSearch.

The SP approach can be contrasted with other machine-learning approaches, such as classification by sequence similarity (Tian and Skolnick, 2003) and a method representing proteins in a space spanned by indices referring to properties of the amino-acids on their chains (Cai et al., 2003). Such a comparison, based on linear SVM (Vapnik, 1995; Scholkopf, 1997) training and testing on the class of oxidoreductases leads to the conclusion that MEX motifs do better than the other methods (Kunik et al 2007).

Another important fact about SPs is that many of them are responsible for the catalytic functions of enzymes. This can be established by searching for loci of active and binding sites on these proteins and finding out that a majority (65%) lies on SPs. Moreover, many SPs are observed to lie in three-dimensional pocket structures in which the active sites reside, even if these active sites are not located on these SPs.

Discussion

Taking a step backwards from the drawing-board to get a perspective of the biological scene, we should realize that cell functions are carried out by machines that involve many proteins as their building blocks. The proteins themselves are constructs of hundreds of amino-acids, but should perhaps be viewed as constructs of peptides connected by many amino-acids that are of secondary importance. Hence, if the amino-acids are the alphabet, the peptides should be the `words', the proteins are the `sentences' and the biological machines are the `paragraphs' that are responsible for the cell function.

But this is only one analogy - on the protein level. Other analogies may exist on the DNA level, as seen in our discussion of the transcription factor binding sites. DNA sequence data continues to present major challenges to the biological community. To put some of the challenge into formal linguistic context, one may say that the community is still trying to understand the syntax and to infer its semantic interpretation. There may be more interesting challenges for motif-extraction algorithms and/or grammar induction ones in these active research fields.

In any case, it is worthwhile emphasizing again, that hereditary biology possesses linear structures (chains of nucleotides in chromosomes, chains of amino-acids in proteins) that are suggestive of linguistic interpretations. Moreover, these objects interact with each other in some temporal order that guarantees the unfolding story of the development of an organism and its daily routine. Thus the syntax within the linear structures plays an important role in the creation of the most important biological reality, life itself. In the examples discussed here we emphasized the results of one particular approach, based on ADIOS and MEX in trying to address some of the challenges in these fields. A survey of many other approaches is presented by Searls (2002).

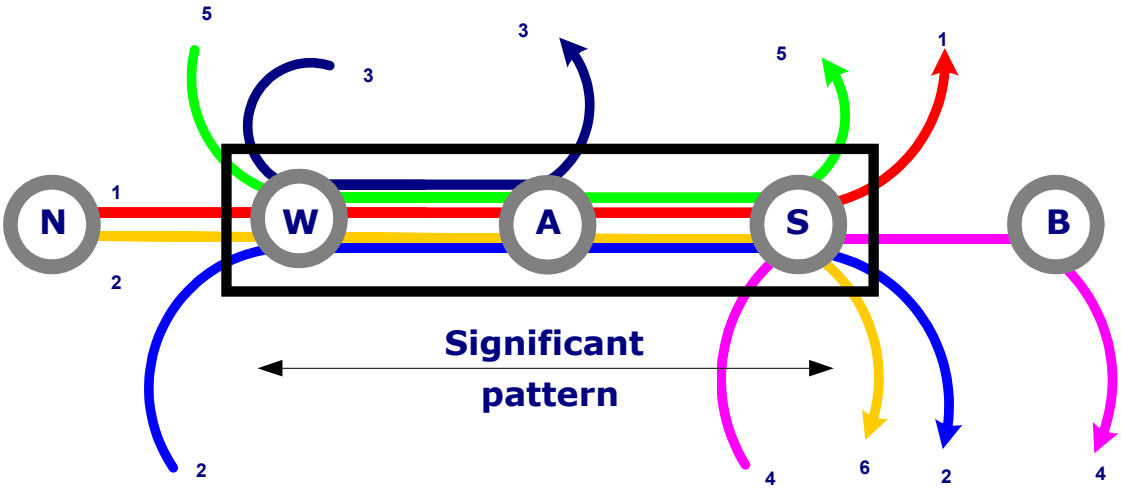

Figure 1

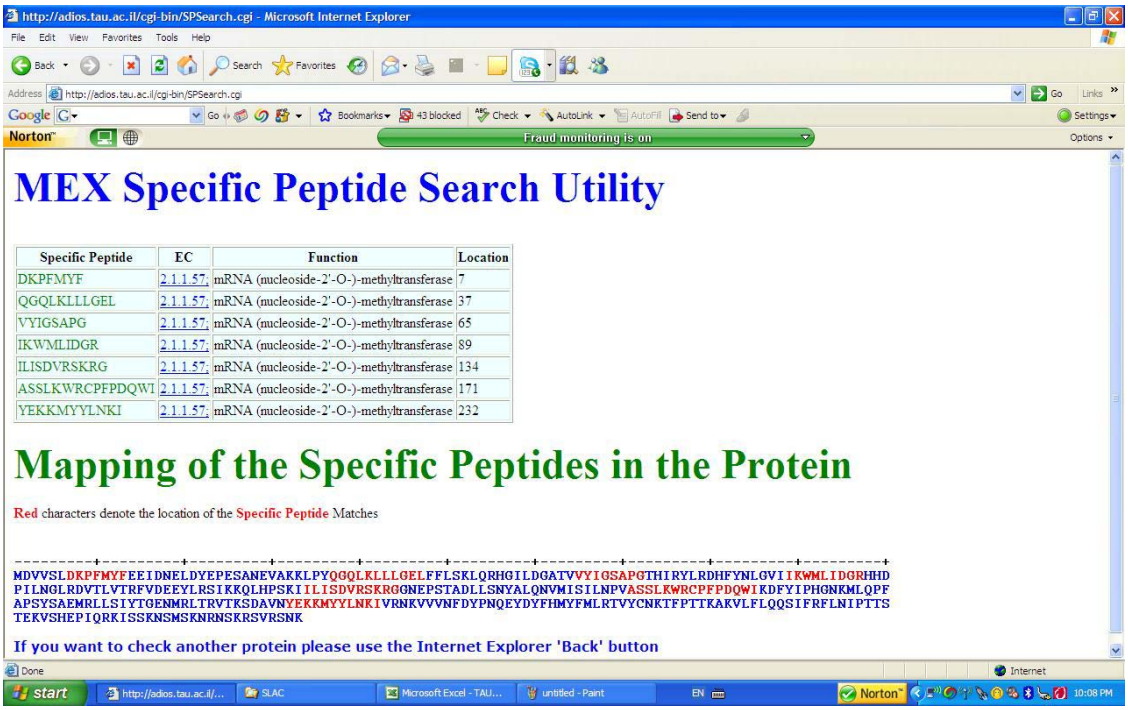

Figure 2

Figure legends

1. An example of applying MEX to an artificial linguistic problem with strings presenting sentences in which all words were concatenated into a long sequence. The purpose of such an exercise is to see if MEX retrieves correctly words. In this example the word ‘was’ appeared in sequences (sentences) 1, 2, 4, 5 and 6. Sequence 3 contained a word in which ‘a’ followed ‘w’ but

did not continue to form the word 'was'. Significance conditions are applied to the entry and the exit of the motif that is being tested. If they pass the set thresholds, the motifs are selected by MEX.

2. A screen-print of the webtool http://adios.tau.ac.il/SPsearch demonstrating the search of Specific Peptides on the sequence of an enzyme, and the association of EC assignments with these SPs.

References

Ben-Hur, A., Brutlag, D. (2004) Sequence motifs: highly predictive features of protein function. *Neural Information Processing Systems 200*4.

Cai, C.Z., Han, L. Y., Ji, Z. L., Chen, Y. Z. (2003) SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence. *Nuclear Acids Researc*h, 31, 3692- 3697.

Chomsky, N. (1957).*Syntactic Structure*s. Mouton & Co., The Hague.

Duda, R. O., Hart, P. E. & Stork, D. G. (2001), *Pattern Classification* (2 ed.), New York: John Wiley & Sons

Horn, D., Solan, Z., Ruppin, E., and Edelman, S.(2004). Unsupervised language acquisition: syntax from plain corpus. *Proc. of Newcastle Symposium on Human Language: cognitive, neuroscientific and dynamical systems perspectives*.

Kunik, V., Meroz, Y., Solan, Z., Sandbank, B., Weingart, U., Ruppin, E. and Horn, D. (2007) Functional representation of enzymes by specific peptides. PLOS Comp.Biol. 3(8): e167. doi:10.1371/journal.pcbi.0030167

MacWhinney, B. and Snow, C. (1985). The Child Language Exchange System. *Journal of Computational Linguistic*s, 12: 271–296.

Moore, B. and Carroll, J. (2001). Parser comparison –context-free grammar (CFG) data. Online at http://www.informatics.susx.ac.uk/research/nlp/carroll/cfg-resources/.

Nowak, N. A., Komarova, N. L. and Niyogi, P. (2002) Computational and evolutionary aspects of language. *Nature* 417,611-617

Scholkopf, B., (1997) *Support Vector Learning*. R. Oldenburg Verlag, Munich.

Segal, L., Lapidot, M., Solan, Z., Ruppin, E., Pilpel, Y. and Horn, D. (2007). Nucleotide variation of regulatory motifs may lead to distinct expression profiles. Bioinformatics Vol. 23 (ISMB/ECCB 2007), pages i440-i449.

Searls, D. (2002). The language of genes. *Nature*, 420:211–217.

Smith, T., Waterman, M., (1981) Identification of common molecular subsequences *J. of Mol. Biology* 147, 195-197.

Solan, Z., Horn, D., Ruppin, E., and Edelman, S. (2005). Unsupervised learning of natural languages. *Proc. Nat. Acad. Science, U*S, 102: 11629 – 11634.

Tian, W., Skolnick, J., (2003) How well is enzyme function conserved as a function of pairwise sequence identity? *J. Mol. Biol.*333, 863-882.

Turing, A.M.(1936-7).On Computable Numbers, With an Application to the Entscheidungsproblem. *Proc. London Math. Society*, 42, 230-265; correction ibid. 43, 544-546.

Vapnik, V. (1995) *The Nature of Statistical Learning Theory*. Springer Verlag, NY.