Biolinguistics: The Use of Analogies for Interdisciplinary Research

Betty Yee Man Cheng
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213
ymcheng@cs.cmu.edu

Jaime Carbonell
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213
jgc@cs.cmu.edu

Judith Klein-Seetharaman
School of Medicine
University of Pittsburgh
Pittsburgh, PA 15261
judithks@cs.cmu.edu

Modern research would be unthinkable without interdisciplinary approaches. For example, many of the biological disciplines provide proof by their names: biochemistry and biological chemistry, biophysics, biotechnology, biomaterials, biostatistics, bioinformatics, and computational biology, all encompass aspects of biology that are influenced by at least one other scientific discipline. How do converging technologies fit into this existing, rich interdisciplinary framework of modern biological research? Technologies are based on science that has found a particular application area, where the driving force is a practical outcome. Different application areas can have the same scientific principles and approaches in common and only differ by the details of their implementation. Thus, in theory, the combination of two scientific disciplines that form the basis of two technologies in different application areas should yield the same benefit as the combination of the two technologies. While this is true in principle, in practice the convergence of technologies provides complementary benefits to existing interdisciplinary research. The way by which technology in one application area is transferred to another application area is through ***analogy***. The use of analogies provides a link to the general public since analogies allow people to grasp concepts immediately by relating them to something they know. Thus, converging technologies have tremendous social impact because they provide a direct means of communicating interdisciplinary research using analogies.

One example is the Biological Language Modeling project at Carnegie Mellon University which converges Human Language Technologies with Biological Chemistry in an effort to provide novel approaches to the mapping of biological sequences to the structure, function and dynamics of proteins and ultimately, of biological systems. Complex biological systems are built from cells that have differentiated to perform specialized functions. This differentiation is achieved through a complicated network of interacting biological molecules. The sequences of proteins are encoded in their entirety in the genomes. The linear strings of amino acids contain in principle all the information needed to fold a protein into a 3-D shape capable of fulfilling its designated function. With the advent of whole-genome sequencing projects, we now have complete sequences of all the proteins that define the complex functions carried out by several organisms — this amounts to hundreds of thousands of sequences in bacteria and tens of thousands in humans [2]. Individual proteins and functions have been studied for decades at various levels, from atomic to macroscopic. More recently, a new field has evolved — proteomics, which looks at all the proteins in a cell simultaneously. This multitude of data provides new opportunities for the application of statistical methods to yield practical answers in terms of likelihood for a particular biological phenomenon to occur.

The availability of enormous amounts of data has also transformed linguistics. The analogy between biology and language is shown in Figure 1. In language, instead of genome sequences, raw text stored in databases, websites and libraries maps to the meaning of words, phrases, sentences and paragraphs as compared to protein structures and functions. After decoding, we can extract knowledge about a topic from the raw text. In language, success in this process has been demonstrated by the ability to retrieve, summarize and translate text. Examples include powerful speech recognition systems, fast web document search engines and computer-generated sentences that are preferred by human evaluators in some applications over sentences that humans build naturally. The transformation of linguistics through data availability has allowed convergence of linguistics with computer science and information technology. Thus, even though a deep fundamental understanding of language is still lacking, e.g. anaphora resolution, data availability has allowed us to obtain practical solutions that have a drastic impact on our lives.

In direct analogy, transformation of biological chemistry by data availability opens the door to convergence with computer science and information technology. This convergence between biological chemistry and computer science and information technology has already happened

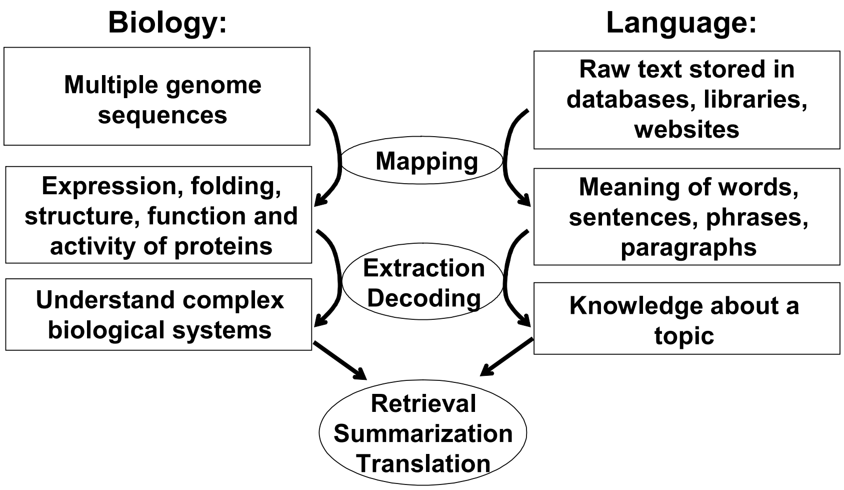
Biology:
Language:
Multiple genome sequences
Raw text stored in databases, libraries, websites
Mapping
Expression, folding, structure, function and activity of proteins
Meaning of words, sentences, phrases, paragraphs
Extraction
Decoding
Understand complex biological systems
Knowledge about a topic
Retrieval
Summarization
Translation

Figure 1. Concept of Analogy between Language and Biology.

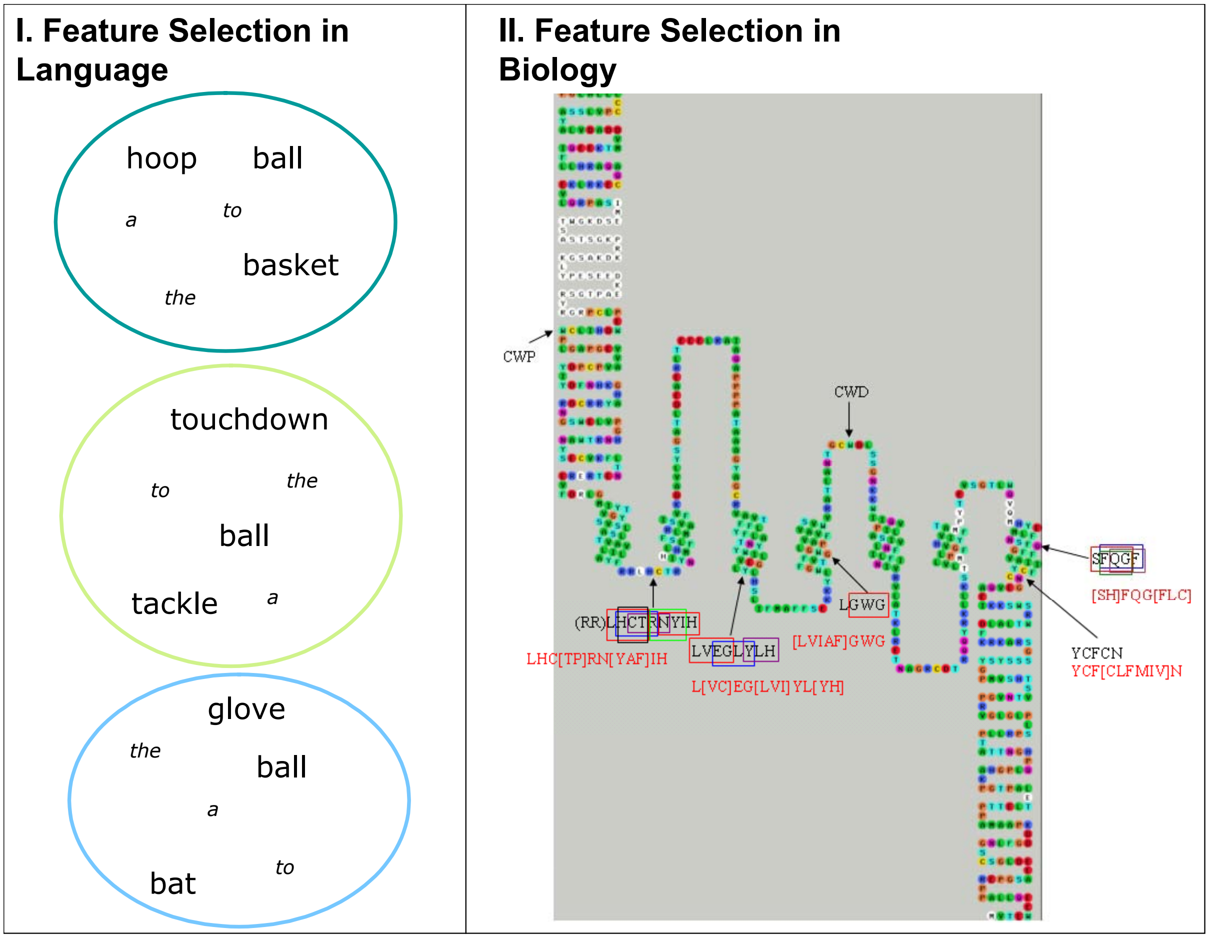

Figure 2. Example of the utility of details in language for biological questions. I. Feature selection in language. The bags of words in three related (ball games) but different documents are shown. Some of the words identify the relatedness of the documents ("ball"), while others identify the differences ("hoop – basket", "touchdown – tackle", "glove – bat"). Some are irrelevant for distinguishing these documents from each other and from unrelated ones ("a – to – the"). **II. Feature selection in biology.** Word equivalents used in protein sequence language are short stretches of amino acids, here shown mapped onto a secondary structure representation of a G protein coupled receptor. Only some amino acid positions (labeled) are useful in distinguishing different subtypes of G protein coupled receptors, while the helices (center) are common to all G protein coupled receptors. Other areas cannot be distinguished from any other protein.

— namely when the disciplines of bioinformatics and computational biology first emerged. Thus, biological chemistry is another application domain for statistical approaches, with the analogy to human languages providing clues to how such approaches could be implemented. For example, classification of protein sequences into families and subfamilies based on their functional and evolutionary relatedness is a typical problem in bioinformatics and has been studied extensively previously. Typical classification methods using k-Nearest Neighbor approaches, Hidden Markov Models, Support Vector Machines and Neural Networks have all been tried without any reference to language technologies, although they are used extensively in speech recognition and information retrieval. Previously, it had been concluded that it was necessary to apply classifiers as complex as Support Vector Machines to the protein classification problem to achieve high accuracy. However, application of the much simpler Naïve Bayes classifier to the same problem in conjunction with chi-square feature selection (Figure 2) — a combination that has been found to be a very successful combination in the language domain [3] — has been shown to outperform the complex Support Vector Machines in this task [1]. The reason for this success lies in the analogy between language and biology that impacts methodology details such as the inclusion of feature selection.

The above results demonstrate how the analogy between language and biology has encouraged the envisioning of new approaches to effectively utilize the vast amounts of data that characterize modern research. Thus, a novel approach to goals are to "help investigators and innovators "see the big pictures", to make non-obvious connections, have data at the point-of-decision, and to solve problems creatively" – the goals of the workshop.

REFERENCES

1. Cheng, B.Y., Carbonell, J.G., Klein-Seetharaman, J.. Document Classification of Protein Sequences. *BLM*, 2003.
2. Venter, J.C., et al.. The sequence of the human genome. *Science*, 291(5507): p. 1304-51, 2001.
3. Yang, Y. and Pedersen, J.T.. A comparative study on feature selection in text categorization. *ICML*, 1997.