Foundations for a Syntactic Pattern-Recognition System for Genomic DNA Sequences: Languages, Automata, Interfaces, and Macromolecules

David B. Searls and G. Christian Overton[1]
SmithKline Beecham Pharmaceuticals; King of Prussia, PA 19406
610/270-4551, Fax: -5580, *searldb@sb.com*
[1]Department of Genetics; University of Pennsylvania; Philadelphia, PA 19104

Viewed as strings of symbols, biological macromolecules can be modelled as elements of formal languages. Generative grammars have been useful in molecular biology for purposes of syntactic pattern recognition, for example in the author's work on the GenLang pattern matching system, which is able to describe and detect patterns that are probably beyond the capability of a regular expression specification. More recently, grammars have been used to capture intramolecular interactions or long-distance dependencies between residues, such as those arising in folded structures. In the work of Haussler and colleagues, for example, stochastic context-free grammars have been used as a framework for "learning" folded RNA structures such as tRNAs, capturing both primary sequence information and secondary structural covariation. Such advances make the study of the formal status of the language of biological macromolecules highly relevant, and in particular the finding that DNA is beyond context-free has already created challenges in algorithm design.

Moreover, to date, such methods have not been able to capture relationships between strings in a collection, such as those that arise via intermolecular interactions, or evolutionary relationships implicit in alignments. Recently we have attempted to remedy this by showing (1) how formal grammars can be extended to describe interacting collections of molecules, such as hybridization products and, potentially, multimeric or physiological protein interactions, and (2) how simple automata can be used to model evolutionary relationships in such a way that complex model-based alignment algorithms can be automatically generated by means of visual programming. These results allow for a useful generalization of the language-theoretic methods now applied to single molecules.

In addition, we describe a new software package—bioWidget—for the rapid development and deployment of graphical user interfaces (GUIs) designed for the scientific visualization of molecular, cellular and genomics information. The overarching philosophy behind bioWidgets is componentry: that is, the creation of adaptable, reusable software, deployed in modules that are easily incorporated in a variety of applications and in such a way as to promote interaction between those applications. This is in sharp distinction to the common practice of developing dedicated applications. The bioWidgets project additionally focuses on the development of specific applications based on bioWidget componentry, including chromosomes, maps, and nucleic acid and peptide sequences.

The current set of bioWidgets has been implemented in Java with the goal in mind of delivering local applications and distributed applets via Intranet/Internet environments as required. The immediate focus is on developing interfaces for information stored in distributed heterogeneous databases such as GDB, GSDB, Entry, and ACeDB. The issues we are addressing are database access, reflecting database schemas in bioWidgets, and performance. We are also directing our efforts into creating a consortium of bioWidget developers and end-users. This organization will create standards for and encourage the development of bioWidget components. Primary participants in the consortium include Gerry Rubin (UC Berkeley) and Nat Goodman (Jackson Labs).

DOE Grant No. DE-FG02-92ER61371.

Relevant Publications

D.B. Searls, "String Variable Grammar: A Logic Grammar Formalism for DNA Sequences," Journal of Logic Programming 24 (1,2):73-102 (1995).

D.B. Searls, "Formal Grammars for Intermolecular Structure," First International Symposium on Intelligence in Neural and Biological Systems, 30-37 (1995).

D.B. Searls and K.P. Murphy, "Automata-Theoretic Models of Mutation and Alignment," Third International Conference on Intelligent Systems for Molecular Biology, 341-349 (1995).

D.B. Searls, "bioTk: Componentry for Genome Informatics Graphical User Interfaces," Gene 163 (2):GC1-16 (1995).

Analysis and Annotation of Nucleic Acid Sequence

David J. States, Ron Cytron, Pankaj Agarwal, and Hugh Chou
Institute for Biomedical Computing; Washington University; St. Louis, MO 63108
314/362-2134, Fax: -0234, *states@ibc.wustl.edu*
http://www.ibc.wustl.edu

Bayesian estimates for sequence similarity: There is an inherent relationship between the process of pairwise sequence alignment and the estimation of evolutionary distance. This relationship is explored and made explicit. Assuming an evolutionary model and given a specific pattern of observed base mismatches, the relative probabilities of evolution at each evolutionary distance are computed using a Bayesian framework. The mean or the median of this probability distribution provides a robust estimate of the central value. Bayesian estimates of the evolutionary distance incorporate arbitrary prior information about variable mutation rates both over time and along sequence position,