LA-UR-98-2091

Approved for public release; distribution is unlimited.

Title: Nonlinear Analysis of Biological Sequences

Author(s): David C. Torney, William Bruno, Vincent Detours, Xioyi He, Bette Korber, Catherine Macken, Alan Perelson, T-10
Emanuel Knill, CIC-3
Jan Rehacek, Bruce Sawhill, T-13

(see attached sheet for additional authors)

Submitted to: DOE OFFICE OF SCIENTIFIC AND TECHNICAL INFORMATION (OSTI)

Los Alamos
NATIONAL LABORATORY

DISCLAIMER

Nonlinear Analysis of Biological Sequences

Authors

David C. Torney , William Bruno, Vincent Detours, Xioyi He,
Bette Korber, Catherine Macken, Alan Perelson,
T-10

Emanuel Knill, CIC-3

Jan Rehacek , Bruce Sawhill,
T-13

Igor Aronson, National Academy of Science, Ukraine

Lars Arvastad, University of Sweden

Yaneer Bar-Yam, Boston University

David J. Blading, University of Reading

Baird H. Brandow, (self employed)

Joseph Felsenstein, University of Washington

Tracey Handel, DuPont Merck Pharmaceutical

David Hillis and Valery Petrov, University of Texas

Alfred W. Hubler, University of Illinois

Ljubinto Kondic, Jerome Percus, and Ora Percus
Courant Institute of Mathematical Science

Karina Ladizhansky, Ruth Lev Bar-Or, Lee Segel, and Victor Steinberg
Weizmann Institute of Science

James Crutchfield, David Haussler, Robert J. Mashl, and Jose Onuchic
University of California

Stephanie Forrest, University of New Mexico

Chip Lawrence, National Institute of Standards and Technology

Jia Li, University of Alabama

Gustavo Lopez, Universidad de Guadalajara

Bernard Matkowsky, Northwestern University

Daniel O'Connor, Veterans' Hospital, San Diego

Stephen J. Paddison, University of Calgary

Allon Percus, University of Paris

Alexander Schliep, University of Cologne

Michael Seul, AT&T Bell Laboratories

Edward Speigel, Columbia University

Gary Stormo, University of Colorado

Nonlinear Analysis of Biological Sequences

William Bruno, Vincent Detours, Xioyi He, Bette Korber, Emanuel Knill, Catherine Macken , Alan Perelson, Jan Rehacek , Bruce Sawhill, and David C. Torney*
Los Alamos National Laboratory

Igor Aronson
National Academy of Science, Ukraine

Lars Arvastad
University of Sweden

Yaneer Bar-Yam
Boston University

David J. Blading
University of Reading

Baird H. Brandow
(self employed)

Joseph Felsenstein
University of Washington

Tracey Handel
DuPont Merck Pharmaceutical

David Hillis and Valery Petrov
University of Texas

Alfred W. Hubler
University of Illinois

Ljubinto Kondic, Jerome Percus, and Ora Percus
Courant Institute of Mathematical Science

Karina Ladizhansky, Ruth Lev Bar-Or, Lee Segel, and Victor Steinberg
Weizmann Institute of Science

James Crutchfield, David Haussler, Robert J. Mashl, and Jose Onuchic
University of California

Stephanie Forrest
University of New Mexico

Chip Lawrence
National Institute of Standards and Technology

Jia Li
University of Alabama

Gustavo Lopez
Universidad de Guadalajara

Bernard Matkowsky
Northwestern University

Daniel O'Connor
Veterans' Hospital, San Diego

Stephen J. Paddison
University of Calgary

Allon Percus
University of Paris

Alexander Schliep
University of Cologne

Michael Seul
AT&T Bell Laboratories

Edward Speigel
Columbia University

Gary Stormo
University of Colorado

Abstract

This is the final report of a three-year, Laboratory Directed Research and Development (LDRD) project at the Los Alamos National Laboratory (LANL). The main objectives of this project involved deriving new capabilities for analyzing biological sequences. We focused on tabulating the statistical properties exhibited by Human coding DNA sequences and on techniques of inferring the phylogenetic relationships among protein sequences related by descent.

Background and Research Objectives

Biological sequences are inherently difficult to analyze. An important goal for the analysis of biological sequences is better insight into Human health. If you are lucky, a sequence of interest will closely resemble a sequence of known function, but one cannot count on this felicity and one must be prepared to look deeper into the "organization" of the sequence at hand. Similarly, it is useful to know about the different types of organization exhibited by various classes of sequences.

---

*Principal Investigator, e-mail: dct@lanl.gov

Prior to our work it was not known how to capture, from examples, all of the ways in which coding sequences distinguish themselves from other classes of DNA sequences. There are a large number of ways in which sequences can be organized, and in this project we developed techniques for surveying these possibilities.

There has always been a well-justified concern that there might be additional statistical properties of Human coding sequences than are manifest in, say, codon frequencies, or found by neural networks. The identification of these properties should improve the accuracy of gene prediction---a natural application for our results. Our techniques are also applicable whenever the objective is to learn the statistical properties of a class of sequences from example sequences. For example, computational modeling of many phenomena employs a source of "random numbers", numbers which are supposed to have joint properties approximating those of independent trials. The technique we developed allows the measurement of the departures of the sequence of "random numbers" from sequences that would be obtained from independent trials.

Based upon our results, it will be possible for any sample of biological sequences to be used as effectively as possible to learn the statistical properties of the corresponding class of sequences. This knowledge will enable the identification of related sequences in databases, even if they have little sequence similarity to the sequences in the sample.

Importance to LANL's Science and Technology Base and National R&D Needs

This project provides supporting research for the Human Genome Project. We have derived new techniques that are useful for the analysis of biological sequences. Aspects of the functions of novel biological sequences can be inferred from insights derived from our techniques. Many pharmaceutical companies are searching large collections of molecules to find candidates for new pharmaceuticals, and there is also a component of rational drug design that is furthered by our techniques of sequence analysis.

Scientific Approach and Accomplishments

Sequences are a special type of mathematical object, just like real numbers, or triangles, etc. In order to analyze them, it is beneficial to use combinatorial techniques---well suited to the analysis of arrangements. The use of cumulants, themselves combinatorial quantities, is recommended for the analysis of sequences. Cumulants are polynomials of the moments. Cumulants of two digits equal their covariance, and cumulants of more digits can reasonably be thought of as "higher-order" covariances.

We completed a full characterization of the stationary statistical properties of Human coding DNA sequences. This characterization was achieved using the sample cumulants of a dataset containing in-frame, nonredundant sequences, totaling over 500,000 bases.

To accomplish this result, the bases of DNA were encoded into binary, as follows: A with 11, C with 1-1, G with -11 and T with -1-1. This encoding maintains all of the information present in the sequences. With this encoding, it is reasonable to use cumulants on digit positions to characterize the statistical properties of a sample of sequences. For example, we illustrate the cumulants of four digits, where the first two digits and the last two digits correspond to bases from specified positions within codons. (Recall that codons are triples of bases for the genetic code, used to encode one of the 20 amino acids occurring in proteins). Figure 1a pertains to first bases of codons, Figure 1b pertains to middle bases of codons, and Figure 1c pertains to last bases of codons. Although the figures depict the cumulants with inter-base spacings of up to 100 bases, non-zero asymptotes obtain for larger spacings, a remarkable feature of DNA sequence data. The magnitude of the cumulant is largest for the cumulant of Figure 1c, involving two third bases, because these are the ``silent'' bases of the degenerate genetic code. The cumulants for two second bases include a small peak at about a 10 base separation, perhaps due to correlations of hydrophobic amino-acids in alpha helices. It should be noted that non-coding sequences could exhibit only one type of base-base cumulant because there are no codons, whereas coding sequences exhibit nine distinct types of cumulants---the three depicted plus the "cross-base" cumulants.

In Figure 2, we depict the covariance of the third digit of a codon with the same digit from another codon. (Codons, having three bases, have six digits. This is the third digit from the left). The structure present in this trace---its peaks and valleys---would be impossible to obtain from any simple statistical model, such as a Markov model, but they are easily incorporated in our model.

Coding sequences exhibit hundreds of distinctive types of cumulants. Fortunately, most of these involve only a few digit positions. For example, at large separations, the cumulants with six digits drawn from three codon bases have zero for their asymptotes. This parsimoniousness is the principal advantage of using cumulants to represent the statistical properties of coding sequences. In addition, the cumulants capture all of the statistical properties of coding sequences.

We derived an expression for the probabilities of all sequences of a given length as a function of the cumulants of the probability distribution. This can be used to derive probabilities that a sequence belongs to a class of sequences---the likelihood of a sequence arising within each class---given the example sequences.

Hidden Markov Models have been used effectively to fit biological sequence data. We have studied the identifiability of such models, i.e., determining which combinations of models and initial conditions are inequivalent. These results should be essential for deriving parameters of the models from sequence data, and should be useful for related objectives.

We have also developed a new, distance-based phylogeny reconstruction program. Molecular phylogenies (or evolutionary trees) have many uses in biology, but we are especially interested in using them in the statistical analysis of protein alignments to model protein structure and function. Existing methods are either hopelessly slow (likelihood methods, e.g. DNAML) or strikingly inaccurate (neighbor-joining, or NJ) for this application.

Our new approach is similar in spirit to the neighbor-joining and BIONJ methods, but aims to approximate the likelihood function as closely as possible without greatly sacrificing computational speed. We have found that much of the correlation structure in the problem occurs in simple-weighted least-squares expressions with reduced variances, reflecting only the "non-additive noise" in the distance fluctuations.

Our new Weighted Neighbor-Joining method is being implemented in a program called WEIGHBOR. WEIGHBOR is much faster (> 100 times for 6 taxa) than the fastest likelihood program ("FASTDNAML"), and currently has a computational complexity scaling as N^4, N being the number of sequences analyzed. An N^3 implementation, which would be comparable in speed to NJ, should be possible.

The performance highlights so far are:

(1) No detectable bias in the unresolved 4 taxon case. This is also true of the likelihood methods, whereas NJ shows a marked bias.

(2) Statistically significant improvement over NJ on resolved 4 taxon case. This is gratifying because a recent attempt by another group to improve on NJ, called BIONJ, is equivalent to NJ for 4 taxa.

(3) Notable improvements over NJ, BIONJ and FITCH for 5 and 6 taxa. Here are some 5 taxa results on 5000 simulated trees, showing the fraction of trees correctly recovered:

NJ	69.1%
BIONJ	77.0%
FITCH	77.8%
WEIGHBOR1.9	81.6%
FASTDNAML	81.8%

Recall that FASTDNAML is too slow for most applications. With 6 taxa WEIGHBOR beat the other distance methods by an even greater margin, but also fell behind FASTDNAML by a greater amount. We are currently working on a new version (WEIGHBOR2.2), which we expect will close some of that gap.

(4) Notable improvements over NJ, BIONJ, and FITCH on an 8 taxon tree that obeys the molecular clock model of evolution. Although real data do not obey this model, it is considered a reasonable approximation, and it tends to reduce the differences between the performances of the foregoing methods.

Publications

1. Bruno W.J., Rota, G-C. and Torney, D.C., "Probability Distributions on Binary Sequences", Proceedings of the National Academy of Sciences, submitted. LAUR 97-323.

2. Macken, C. and Schliep, A. "Equivalence of Hidden Markov Models", IEEE Journal on Information, submitted. LAUR 97-3300.

3. Arvestad, L. and Bruno, W.J., "Estimation of Reversible Substitution Matrices from Multiple Pairs of Sequences", J. Molecular Evolution, **46**, pp. 696-703 (1997). LAUR 96-3521.

4. Huynen, M., Gutell, R., and Konings, D., "Assessing the Reliability of RNA Folding using Statistical-Mechanics", J. Mol. Biol., **267**, 1104-1112 (1997).

5. Huynen, M. A., Diazlazcoz, Y., Bork, P., "Differential Genome Display", Trends in Genetics, **13**, 389-390 (1997).

6. Huynen, M. A., Stadler, P. E., and Fontana, W., "Smoothness within Ruggedness: The Role of Neutrality in Adaptation", Proc. Natl. Acad. Sci., U.S.A., **93**, 397-401 (1996).

7. Huynen, M. A., "Exploring Phenotype Space through Neutral Evolution", J. of Molecular Evolution, **43**, 165-169 (1996).

8. Huynen, M. A., Perelson, A., Viera, W. A., and Stadler, P. E., "Basepairing Probabilities in a Complete HIV-1 RNA", J. Computational Biology, **3**, 253-274 (1996).

9. He, X. and Dembo, M., "Modeling Chemoattractant-elicited Relocalization of Myosin Filaments in Dictyostelium", Biochemistry and Cell Biology, 421-429 (1995).

10. Huynen, M., et. al, "Rate of Killing of HIV-Infected T-Cells and Disease Progression", Science, **272,** 1962 (1996)

11. Huynen, M.A., et. al, "Smoothness Within Ruggedness: The Role of Neutrality in Adaptation", Proceedings of the National Academy of Sciences, **93**, 397 (1996)

12. Huynen, M.A. et. al, "Base Pair Probabilities in a Complete HIV-1 Genome", Journal of Computational Biology, **3**, 253 (1996). LAUR 95-1613.

13. Huynen, M., "Exploring Phenotype Space through Neutral Evolution", Journal of Molecular Evolution, **43**, 165 (1996). LAUR-95-3812.

14. Bruno, W.J., "Modeling Residue Usage in Aligned Protein Sequences via Maximum Likelihood", Molecular Biology and Evolution (in press). LAUR-96-1854.

15. Huynen, M., et. al, "Assessing the Reliability of RNA Folding Using Statistical Mechanics", Journal of Molecular Biology (submitted). LAUR-96-2384.

16. Torney, D.C., "Group Tests for Complex Biological Systems", Nature (submitted). LAUR-96-2428.

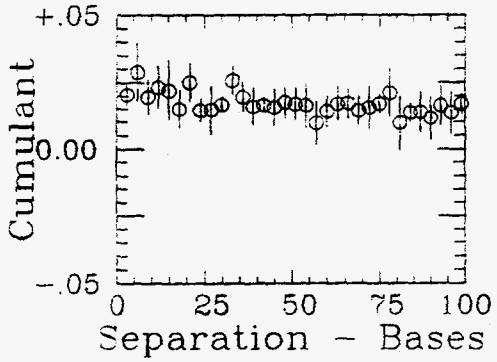
References

[1] Bruno, W.J., Rota, C-G. and Torney, D.C., "Probability Distributions on Binary Sequences", Proceedings of the National Academy of Sciences, submitted. LAUR 97-323.

[2] Macken, C. and Schliep, A., "Equivalence of Hidden Markov Models", IEEE Journal on Information Theory, submitted. LAUR 97-3300.

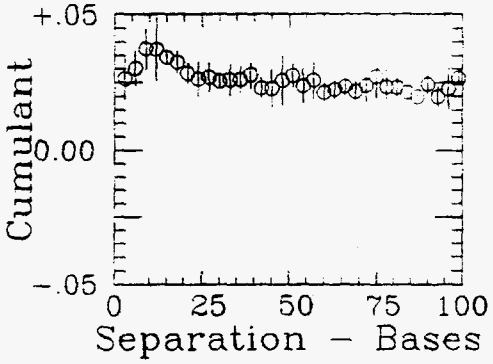
[3] Arvestad, L. and Bruno, W.J., "Estimation of Reversible Substitution Matrices from Multiple Pairs of Sequences", J. Molecular Evolution, **46**, 696-703 (1997). LAUR 96-3521.

First Base Position=1 Second Base Position=1



(Figure 1a)

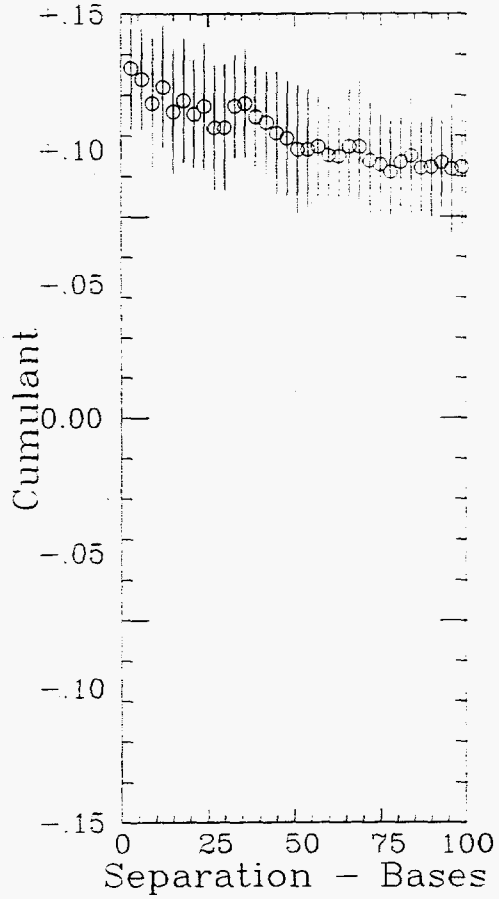
First Base Position=2 Second Base Position=2



(Figure 1b)

Figure 1. Cumulants of coding sequences for three selected pairs of bases are plotted in Figures 1a, 1b, and 1c, as functions of the spacing between the bases. The data were partitioned into five equal-sized blocks, enabling the determination of the average sample cumulants and standard deviations. The averages are marked with circles and one standard deviation, in both directions, is indicated by the vertical lines through the circles. Figure 1a pertains to first bases of codons, Figure 1b pertains to middle bases of codons, and Figure 1c pertains to last bases of codons.

First Base Position=3 Second Base Position=3



(Figure 1c)

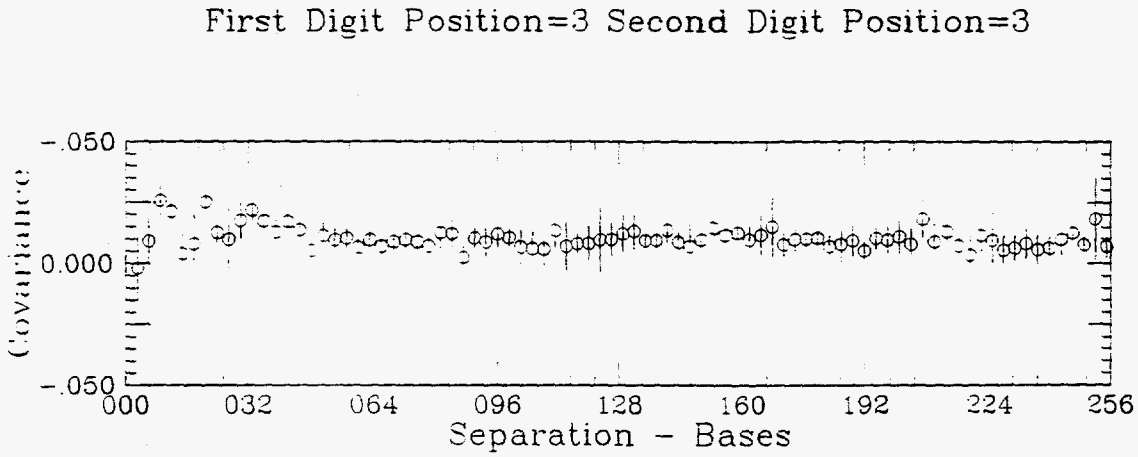


Figure 2. The cumulants of coding sequences for a selected pair of digits is plotted as a function of the spacing between the bases. These are both third digits from codons: third from the left. The data were partitioned into five equal-sized blocks, enabling the determination of the average sample cumulants and standard deviations. The averages are marked with circles and one standard deviation, in both directions, is indicated by the vertical lines through the circles.