

Challenges in Computational Biology: What Lies Beyond BLAST?

M. Vidyasagar

Cecil & Ida Green Professor, and
Head, Bioengineering Department
The University of Texas at Dallas
M.Vidyasagar@utdallas.edu

Indian Institute of Technology, Hyderabad
01 September 2010



Outline

- 1 The Role of Sequence Alignment in Biology
 - Statistical vs. Symbolic Similarity
- 2 Probability Theory Preliminaries
 - Relative Entropy (Kullback-Leibler Divergence)
 - Expected Values, Conjugate Distribution
- 3 The BLAST Algorithm
 - Problem Formulation
 - Main Results of BLAST Theory
- 4 Proposed Research
 - Problem Formulation and Expected Results
 - Limitations of Proposed Approach

Introduction

BLAST (Basic Local Alignment Search Tool) is among the most widely used methods in biology.

The paper by Altschul et al. in 1990 is *the most widely cited scientific paper* during the past two decades.

The papers by Dembo, Karlin and Zeitouni that provide the probability theoretic foundations are also widely referenced.

What lies ahead? That is the topic of this talk.



References for BLAST

- S. F. Altschul, W. Gish, W. Q. Miller, E. W. Myers and D. J. Lipman, “Basic local alignment search tool,” *J. Mol. Biol.*, 215, 403-410, 1990.
- A. Dembo, S. Karlin and O. Zeitouni, “Critical phenomena for sequence matching with scoring,” *Annals of Prob.*, 22(4), 1993-2021, 1994.
- A. Dembo, S. Karlin and O. Zeitouni, “Limit distribution of maximal non-aligned two-sequence segmental score,” *Annals of Prob.*, 22(4), 2022-2039, 1994.
- S. Karlin and A. Dembo, “Limit distributions of maximal segmental score among Markov-dependent partial sums,” *Adv. Appl. Prob.*, 24(1), 113-140, 1992.



Other References

- S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequences of two proteins," *J. Mol. Biol.*, 48(3), 443-453, 1970.
- T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences," *J. Mol. Biol.*, 147, 195-197, 1981.
- M. Vidyasagar, S. S. Mande, C. V. S. K. Reddy and V. Raja Rao, "The 4M (mixed memory Markov model) algorithm for finding genes in prokaryotic genomes," *IEEE Transactions on Circuits and Systems*, **CAS-55(1)**, 26-37, January 2008 (Special Issue on Systems Biology).

A Caution

Remember Einstein's maxim: *Things should be made as simple as possible – but not simpler!*



Caution: Advanced mathematics ahead!

Role of Alignment in Computational Biology

Many problems in computational biology can be formulated as 'aligning' two sequences defined over a common finite alphabet.

Most commonly: The alphabet is either the four-symbol set of nucleotides $\{A, C, G, T\}$ or the twenty-symbol set of amino acids.

The 'genome' can be thought of as just a very long (3×10^9) string over $\{A, C, G, T\}$, while the 'primary structure' of every protein is a longish string (300 to 1,500 or even 10,000) over the amino acid alphabet.



Typical Applications of Sequence Alignment

Typical applications:

- Determining which parts of the genome correspond to coding regions
- Determining the function of genes/proteins from those of 'similar' genes/proteins
- Grouping genes/proteins into 'similar families'

And so on.



Outline

- 1 The Role of Sequence Alignment in Biology
 - Statistical vs. Symbolic Similarity
- 2 Probability Theory Preliminaries
 - Relative Entropy (Kullback-Leibler Divergence)
 - Expected Values, Conjugate Distribution
- 3 The BLAST Algorithm
 - Problem Formulation
 - Main Results of BLAST Theory
- 4 Proposed Research
 - Problem Formulation and Expected Results
 - Limitations of Proposed Approach

Different Kinds of 'Similarity'

There are in essence two kinds of 'similarity': Statistical similarity and symbolic similarity.

Two strings of symbols are 'statistically similar' if they have similar statistics.

Two strings of symbols are 'symbolically similar' if they more or less match symbol for symbol.

An Example of Statistical Similarity

Example: Given two sets of coin tosses, say

THTTHTTTHHHTTHTTTHTHHHTHTHHHTHTTTHTHT,

THTHHHTTTHTHTHTHHHTHTHTHHHTHTHTHHHTTT,

we can ask: Did the same coin produce them?

Point: Even if we repeat the same experiment, we won't get the same sequence of outcomes. At best we can compare *the frequency of occurrence* of various symbols.

If one sequence of coin tosses had 57% of H and the other one had 55% of H , we can say that 'with high confidence' that the same coin produced both outcomes. (Statement can be made very precise.)



An Example of Symbolic Similarity

The two nucleotide sequences

ACACTGT, TAGACGGAGCTTCAC

are 'symbolically similar' because they can be imperfectly aligned using gaps:

			A	C	-	-	A	C	-	T	G	T					
T	A	G	A	C	G	G	A	G	C	T	-	T	A	A	C		

Obvious: Symbolic similarity implies statistical similarity, but not the converse!

Role of Statistical Similarity

Tests for statistical similarity are useful when that is all we can expect.

Example: Discriminating between coding and non-coding regions of a genome.

Especially in prokaryotes, the 'statistics' of coding regions are very similar from one to the other, and very dissimilar to those of non-coding regions.

Coding regions exhibit three-periodicity, whereas non-coding regions don't.

These features can be used to annotate genomes starting with very little information.

Statistical similarity is not the focus of this talk!



Deterministic Algorithms for Symbolic Similarity

When we wish to infer the function of a gene from that of another 'similar' gene, or the 3-D structure, active sites etc. of a protein from that of another 'similar' protein, *we should look for symbolic similarity – statistical similarity is not enough!*

The well-known Needleman-Wunsch (1970) and Smith-Waterman (1981) algorithms provide *deterministic* algorithms for optimally aligning two sequences using dynamic programming.

Limitation: If both sequences have length l , the running time is $O(l^2)$. If we wish to align k sequences, the running time is $O(2^k l^2)$.

Conclusion: Deterministic algorithms don't 'scale'.

We need statistical algorithms – *Enter BLAST!*



Outline

- 1 The Role of Sequence Alignment in Biology
 - Statistical vs. Symbolic Similarity
- 2 Probability Theory Preliminaries
 - Relative Entropy (Kullback-Leibler Divergence)
 - Expected Values, Conjugate Distribution
- 3 The BLAST Algorithm
 - Problem Formulation
 - Main Results of BLAST Theory
- 4 Proposed Research
 - Problem Formulation and Expected Results
 - Limitations of Proposed Approach

Basic Notation

Suppose $\mathbb{A} = \{a_1, \dots, a_n\}$, $\mathbb{B} = \{b_1, \dots, b_m\}$ are finite sets (alphabets).

A **probability distribution** ϕ over \mathbb{A} is an n -dimensional nonnegative vector whose components add up to one.

If \mathcal{X} is a r.v. over \mathbb{A} with distribution ϕ , we can think of $\phi_i = \Pr\{\mathcal{X} = a_i\}$.

Similarly for ψ , a distribution over \mathbb{B} .



Product Distributions

If \mathcal{X}, \mathcal{Y} are r.v.'s over \mathbb{A}, \mathbb{B} , then their **joint distribution** is denoted by μ , which has nm components.

If \mathcal{X}, \mathcal{Y} are *independent* r.v.'s, then

$$\mu_{ij} = \phi_i \cdot \psi_j, \quad \forall i, j.$$

In other words,

$$\Pr\{(\mathcal{X}, \mathcal{Y}) = (a_i, b_j)\} = \Pr\{\mathcal{X} = a_i\} \cdot \Pr\{\mathcal{Y} = b_j\}.$$

In this case we write $\theta = \phi \times \psi$, and call μ the **product distribution** of ϕ and ψ .



Relative Entropy (Kullback-Leibler Divergence)

If μ, ν are distributions over a *common* alphabet of cardinality k , the quantity

$$H(\nu \parallel \mu) := \sum_{i=1}^k \nu_i \log \frac{\nu_i}{\mu_i}$$

is called the **relative entropy** or the **Kullback-Leibler divergence** between μ, ν .



Outline

- 1 The Role of Sequence Alignment in Biology
 - Statistical vs. Symbolic Similarity
- 2 Probability Theory Preliminaries
 - Relative Entropy (Kullback-Leibler Divergence)
 - Expected Values, Conjugate Distribution
- 3 The BLAST Algorithm
 - Problem Formulation
 - Main Results of BLAST Theory
- 4 Proposed Research
 - Problem Formulation and Expected Results
 - Limitations of Proposed Approach

Expected Value

If \mathcal{Z} is a r.v. on some set $\mathbb{C} = \{z_1, \dots, z_k\}$ with distribution $\boldsymbol{\mu}$.
and $f : \mathbb{C} \rightarrow \mathbb{R}$, then the **expected value** of $f(\mathcal{Z})$ is defined by

$$E[f(\mathcal{Z}), \boldsymbol{\mu}] = \sum_{i=1}^k f(z_i) \mu_i.$$



Conjugate Distribution

Suppose f assumes both positive and negative values as a function of \mathcal{Z} . Then there exists a *unique* number λ^* , whose sign is opposite to that of $E[f(\mathcal{Z}), \mu]$, such that

$$E[\exp(\lambda^* f(\mathcal{Z})), \mu] = 1.$$

If we define

$$\theta_i = \exp(\lambda^* f(z_i)) \mu_i, i = 1, \dots, k,$$

then θ is also a probability distribution. It is called the **conjugate distribution** of μ .

Important Point: Suppose $\mathbb{C} = \mathbb{A} \times \mathbb{B}$ and $\mu = \phi \times \psi$. Still, the conjugate distribution of μ *need not be* a product distribution.



Outline

- 1 The Role of Sequence Alignment in Biology
 - Statistical vs. Symbolic Similarity
- 2 Probability Theory Preliminaries
 - Relative Entropy (Kullback-Leibler Divergence)
 - Expected Values, Conjugate Distribution
- 3 The BLAST Algorithm
 - Problem Formulation
 - Main Results of BLAST Theory
- 4 Proposed Research
 - Problem Formulation and Expected Results
 - Limitations of Proposed Approach

The Basic Set-Up

Suppose have two finite alphabets $\mathbb{A} = \{a_1, \dots, a_n\}$, $\mathbb{B} = \{b_1, \dots, b_m\}$, e.g. the four-symbol nucleotide set, the twenty-symbol amino-acid set.

We are given two sequences $x_1, \dots, x_l \in \mathbb{A}$, $y_1, \dots, y_l \in \mathbb{B}$ and we want to know how well they 'match'.

To check for a 'match' we use a local 'scoring function' $F : \mathbb{A} \times \mathbb{B} \rightarrow \mathbb{R}$. So for each symbol $a_i \in \mathbb{A}$, $b_j \in \mathbb{B}$, $F(a_i, b_j)$ gives the score of the match.

We *could* look at the 'total' score $\sum_{k=1}^l F(x_k, y_k)$. Instead, we look at so-called 'maximal segmental scores'.

Segmental Scores

Given the sequences $\{x_1^l\}$, $\{y_1^l\}$, choose an integer $L \leq l$, and starting indices i for x and j for y – which need not be the same!

The quantity

$$\sum_{k=1}^L F(x_{i+k}, y_{i+k})$$

is a *segmental score* on a segment of length L with matching starting points. The quantity

$$\sum_{k=1}^L F(x_{i+k}, y_{j+k})$$

is a segmental score on a segment of length L with possibly mismatched starting points.

Maximal Segmental Scores

Two finds of scores can be examined, depending on whether we want the starting points of the two segments to be aligned or not.

$$R_l := \max_{L \geq 0, 0 \leq i \leq l-L} \sum_{k=1}^L F(x_{i+k}, y_{i+k}),$$

$$M_l := \max_{L \geq 0, 0 \leq i, j \leq l-L} \sum_{k=1}^L F(x_{i+k}, y_{j+k}).$$

R_l is the maximal segmental score *if we insist that the two starting points should match*, whereas M_l is the maximal segmental score *if we don't insist that the two starting points should match*.



Questions to be Addressed

- 1 What is the *expected value* of the maximal segmental scores M_l or R_l ?
- 2 Let L_l denote the length of a maximal scoring segment. What is the *expected value* of L_l ?
- 3 What is the *frequency distribution* of the symbols x_{i+k}, y_{j+k} in a maximally scoring segment?
- 4 What is the *tail probability distribution* of the quantities M_l and R_l beyond their expected values?

It turns out that Question 4 is answered using entirely different techniques from Questions 1 through 3.



Assumptions Behind Problem Formulation

Important Note: These assumptions are crucial to BLAST theory as it is currently used!

- 1 The two sequences $\{x_1^l\}, \{y_1^l\}$ are *realizations of independent, and identically distributed* random variables, with distributions ϕ on \mathbb{A} and ψ on \mathbb{B} .
- 2 If $\mu = \phi \times \psi$ is the joint distribution on $\mathbb{A} \times \mathbb{B}$,

$$E[F, \mu] < 0, \text{ and } \exists i, j \text{ s.t. } F(a_i, b_j) > 0.$$

Here $E[F, \mu]$ denotes the ‘expected value’ of the scoring function:

$$E[F, \mu] = \sum_{a_i \in \mathbb{A}} \sum_{b_j \in \mathbb{B}} F(a_i, b_j) \phi(a_i) \psi(b_j).$$

So the scoring function is negative ‘on average’ but assumes positive values.



Outline

- 1 The Role of Sequence Alignment in Biology
 - Statistical vs. Symbolic Similarity
- 2 Probability Theory Preliminaries
 - Relative Entropy (Kullback-Leibler Divergence)
 - Expected Values, Conjugate Distribution
- 3 The BLAST Algorithm
 - Problem Formulation
 - **Main Results of BLAST Theory**
- 4 Proposed Research
 - Problem Formulation and Expected Results
 - Limitations of Proposed Approach

Main Results of BLAST Theory – 1

Theorem:

- 1 $R_l \sim (\ln l)/\lambda^*$ as $l \rightarrow \infty$.
- 2 The length of a maximal scoring segment L_l is asymptotically equal to $l/H(\theta\|\mu)$, where $H(\theta\|\mu)$ is the 'relative entropy' between the two probability vectors.
- 3 On any maximal scoring segment, the frequency distribution of (x, y) is asymptotically equal to θ .

So this theorem tells us: (i) the expected maximal segmental score, (ii) the length of a maximal scoring segment, and (iii) the frequency distribution of the symbols on a maximal scoring segment.



Main Results of BLAST Theory – 2

Similar results hold for M_l .

Theorem:

- 1 $M_l \sim (2 \ln l) / \lambda^*$ as $l \rightarrow \infty$.
- 2 Let θ denote the conjugate distribution of μ . Then the length of a maximal scoring segment L_l is asymptotically equal to $2l / H(\theta \| \mu)$, where H denotes the relative entropy.
- 3 On any maximal scoring segment, the empirical distribution of (x, y) is asymptotically equal to θ .



Main Results of BLAST Theory – 3

Theorem: There exists a constant K^* , which can be estimated closely, such that for all $x > 0$, the following inequality holds:

$$\begin{aligned} \lim_{l \rightarrow \infty} \Pr\left\{R_l - \frac{\log l}{\lambda^*} > x\right\} &= 1 - \exp(-K^* \exp(-\lambda^* x)) \\ &\approx K^* \exp(-\lambda^* x) \text{ if } x \gg \frac{\log K^*}{\lambda^*}. \end{aligned}$$

This is the so-called ‘Gumbel distribution’. This theorem allows us to determine the number of expected high scoring segments (so-called E score), and whether a high score occurred purely by chance.

Similar result holds for M_l .



Reverse Engineering the Scoring Function

Key Question: Where does the scoring function come from?

Various scoring functions (or scoring matrices) like Blosum come from 'reverse-engineering' sequences that have been aligned by hand to reflect biological reality.

- On a matched (high-scoring) sequence pair, construct the frequency distribution ϕ of the \mathbb{A} symbols, and ψ of the \mathbb{B} symbols.
- Similarly, compute the *joint* frequency distribution θ of the pair (a_i, b_j) .
- Define the scoring function

$$F_{ij} := \frac{1}{c} \log \frac{\theta_{ij}}{\mu_{ij}} = \frac{1}{c} \log \frac{\theta_{ij}}{\phi_i \cdot \psi_j}.$$

Here c is any constant that we choose.



Outline

- 1 The Role of Sequence Alignment in Biology
 - Statistical vs. Symbolic Similarity
- 2 Probability Theory Preliminaries
 - Relative Entropy (Kullback-Leibler Divergence)
 - Expected Values, Conjugate Distribution
- 3 The BLAST Algorithm
 - Problem Formulation
 - Main Results of BLAST Theory
- 4 Proposed Research
 - Problem Formulation and Expected Results
 - Limitations of Proposed Approach

Need for an Amended Problem Formulation

The *key limitation* of existing BLAST theory is the assumption that the sequences come from *independent samples*.

Studies of both nucleotide as well as amino acid sequences show that the assumption of independence is *not justified*. It is more realistic to assume that the sequences represent *a Markov process*.

How can we amend the problem formulation accordingly?

What results can be expected?

Amended Problem Formulation

- 1 Assume that the symbols $\{x_1^l\}, \{y_1^l\}$ to be aligned come from a Markov process.
- 2 The scoring function is not 'local' any more but depends on *both the current and the previous symbol*, thus: The score of a segment of length L is

$$\sum_{k=1}^L F[(x_{i+k}, x_{i+k-1}), (y_{j+k}, y_{j+k-1})].$$

The problem is to study maximal segmental scores by maximizing the above for all choicds of starting points, and all lengths L .

The Questions to be Answered

- 1 What is the *expected value* of the maximal segmental scores M_l or R_l ?
- 2 Let L_l denote the length of a maximal scoring segment. What is the *expected value* of L_l ?
- 3 What is the *frequency distribution* of the symbols x_{i+k}, y_{j+k} in a maximally scoring segment?
- 4 What is the *tail probability distribution* of the quantities M_l and R_l beyond their expected values?

It turns out that Question 4 *has already been answered* for Markov processes in a paper written in 1992! Again it follows a Gumbel distribution.

So we need to study only the first three questions.



The Expected Answers

Preliminary research shows that even in the case of Markov processes, the answers will be very similar to those for the i.i.d. case.

Instead of distributions on the symbol sets \mathbb{A}, \mathbb{B} , we will have 'pair distributions' on $\mathbb{A}^2, \mathbb{B}^2$. Call these distributions ϕ, ψ , and let $\mu = \phi \times \psi$.

There will be a 'conjugate distribution' θ , and expressions for (i) the expected maximal segmental score, (ii) the length of a maximal scoring segment, and (iii) the frequency distribution of the symbols on a maximal scoring segment will all be similar to earlier formulas.

The scoring function can once again be reverse-engineered from matched sequences, as before.



Outline

- 1 The Role of Sequence Alignment in Biology
 - Statistical vs. Symbolic Similarity
- 2 Probability Theory Preliminaries
 - Relative Entropy (Kullback-Leibler Divergence)
 - Expected Values, Conjugate Distribution
- 3 The BLAST Algorithm
 - Problem Formulation
 - Main Results of BLAST Theory
- 4 Proposed Research
 - Problem Formulation and Expected Results
 - Limitations of Proposed Approach

Limitations of Proposed Research

The main limitation will be *the absence of sufficiently many matched sequences*.

One Blosum matrix (for aligning protein sequences) contains $20^2 = 400$ entries. To apply that approach to Markov processes, we would need $20^4 = 160,000$ entries!

There is some hope of being able to do this for genome sequences as the matrices would have only $4^4 = 256$ entries.



Conclusions

- The theory of BLAST can (apparently) be extended to handle the case where the sequences to be aligned come from Markov processes.
- It is also (apparently) possible to reverse-engineer the scoring functions from aligned sequences.
- In the case of protein sequences (most widely used application of BLAST), there won't be enough aligned sequences to reverse engineer the scoring functions.
- But in the case of genomes this will be possible.
- Will 'generalized BLAST' lead to new and surprising alignments? Who knows!

Thank You!

