BALSA: Bayesian Algorithm for Local Sequence Alignment

Bobbie-Jo M. Webb [1,2], Charles E. Lawrence [1,3], Jun S. Liu [4]

[1] The Wadsworth Center for Laboratories and Research, New York State Department of Health, Albany, NY 12201, USA
[2] Department of Decision Sciences and Engineering Systems, Rensselaer Polytechnic Institute, Troy, NY 12180, USA
[3] Department of Computer Science, Rensselaer Polytechnic Institute, Troy, NY 12180, USA
[4] Department of Statistics, Harvard University, Cambridge, MA 02138, USA

1. Abstract

Local sequence alignment is a fundamental tool in biomedical research and despite its long history, it still has a number of well-known shortcomings. Dynamic programming methods yield a single alignment, albeit optimal, which can be strongly affected by the scoring matrix and gap penalties chosen. Additionally, the scores obtained are dependent upon the lengths of the sequences aligned, requiring a post-analysis conversion. We undertook a study to examine the utility of Bayesian statistics to overcome these issues. We developed a Bayesian algorithm for local sequence alignment, BALSA, that takes into account the uncertainty associated with all unknown variables by incorporating in its forward sums a series of scoring matrices, gap parameters, and all possible alignments. The algorithm returns a representative sample of alignments and the posterior probabilities of gap penalties and scoring matrices. Furthermore, it inherently adjusts for variations in sequence lengths. BALSA was compared to SSEARCH with E-values, to date the best performing dynamic programming algorithm in the detection of structural neighbors.

2. Introduction

Dynamic programming and heuristic methods were significant advances in local sequence alignment algorithms. Local alignment is typically the method of choice for aligning a pair of biopolymers; obtaining the best common subsequence is usually more advantageous to detect distantly related proteins than an alignment end to end, globally (Needleman and Wunsch, 1970; Smith and Waterman, 1981). The most generally used are SSEARCH, FASTA (Pearson *et al.*, 1988), and BLAST (Altschul *et al.*, 1990). However, these algorithms require the specification of a scoring matrix and set of gap penalties, and return only a single alignment and an associated score that must be adjusted for the lengths of the sequences. Bayesian statistics provides a means to relax these requirements and adjustments. Furthermore, since all required sums can be completed using modified dynamic programming recursions, exact inferences on all variables are available.

3. Methods

Most sequence alignment methods can be viewed as optimizing an objective function, typically a log-likelihood for two sequences $R^{(1)} = \{R_1^{(1)} \text{K } R_I^{(1)}\}$ and $R^{(2)} = \{R_1^{(2)} \text{K } R_J^{(2)}\}$. This requires setting specific parameter values, Θ^o, an amino acid scoring matrix, and Λ^o, gap penalties, in order to find the optimal alignment, A^*, over all possible alignments. Mathematically this is:

$$\log(P(R^{(1)}, R^{(2)} \mid \Theta^o, \Lambda^o)) = \max_{\text{all } A} \{\log(P(P(R^{(1)}, R^{(2)} \mid A, \Theta^o)) + \log(P(A \mid \Lambda^o))\}$$

(Liu and Lawrence, 1999; Pearson, 1995).

The premise behind Bayesian inference is that everything is a random variable, observed data, missing data, and unknown parameters alike. Thus the Bayesian procedure doesn't require that Θ and Λ be fixed, defining the logarithm of the likelihood for the pair of sequences as:

$$\log P(R^{(1)}, R^{(2)} \mid A, \Theta) = \sum_{i=1}^{I} \log \Theta(r_i^1, o) + \sum_{j=1}^{J} \log \Theta(o, r_j^2) + a_{ij} \log \Psi_{r_i^1, r_j^2}$$

In this implementation of the algorithm, scoring matrices and gap parameters are seen as pairs, (Θ, Λ), the full joint probability, *Joint = likelihood*priors*, can be defined as:

$$P(R^{(1)}, R^{(2)}, A, \Theta, \Lambda) = P(R^{(1)}, R^{(2)} \mid A, \Theta) P(A \mid \Lambda) P(\Theta, \Lambda)$$

Prior probabilities are used to incorporate previous knowledge about the parameters. A full Bayesian method uses non-degenerate priors, but in this case, lacking *a priori* information, uninformed priors are utilized. Thus the following assumptions are made to implement the model. All scoring matrix gap penalty pairs, (Θ, Λ), are equally likely and $P(A \mid \Lambda)$ is the probability of any allowable path A prior to seeing the sequence data.

The unknown variable, A, can be removed from the joint distribution by summing over all alignments as follows:

$$P(R^{(1)}, R^{(2)} \mid \Theta, \Lambda) = \sum_{A} P(R^{(1)}, R^{(2)} \mid A, \Theta) P(A \mid \lambda_o, \lambda_e) =$$

$$\frac{\sum_{A} P(R^{(1)}, R^{(2)} \mid A, \Theta) \lambda_o^{k_g(A)} \lambda_e^{l_g(A) - k_g(A)}}{\sum_{A'} \lambda_o^{k_g(A')} \lambda_e^{l_g(A') - k_g(A')}} \qquad (1)$$

Examining the posterior distributions for the unknowns gives inferences on the scoring matrices and gap parameters. Then by Bayes rule, the desired posterior can be obtained:

$$P(\Theta, \Lambda \mid R^{(1)}, R^{(2)}) = \frac{P(R^{(1)}, R^{(2)} \mid \Theta, \Lambda) P(\Theta, \Lambda)}{\sum_{\Theta, \Lambda} P(R^{(1)}, R^{(2)} \mid \Theta, \Lambda) P(\Theta, \Lambda)}.$$

The posterior distribution for an alignment, $P(A^* \mid R^{(1)}, R^{(2)})$, can similarly be expressed. Given that the number of possible alignments for even small biopolymer sequences is immense, it is not feasible to calculate this distribution directly, but a good approximation can be obtained by sampling.

3.1 *Algorithm*

The Bayesian model must capture the idea of local alignment, aligning the related subsequences, ignoring the unrelated sections of the sequence on the ends. The summation over all alignments takes into account alignments that begin at any point and end at any point in the two sequences while adhering to one constraint; an alignment may not end before it has begun. The summation over all alignments can be achieved via a recursive algorithm. The algorithm calculating the numerator of Equation 1 can be written as follows:

$$Pm(i, j) = \{Pm(i-1, j-1) + Pi(i-1, j-1) + Pd(i-1, j-1) + Pn(i-1, j-1)\} \Psi(r_i^{(1)}, r_j^{(2)})$$

$$Pi(i, j) = \lambda_e Pi(i-1, j) + \lambda_o \{Pm(i-1, j) + Pn(i-1, j)\}$$

$$Pd(i, j) = \lambda_e Pd(i, j-1) + \lambda_o \{Pm(i, j-1) + Pn(i, j-1)\}$$

$$Pn(i, j) = \Psi(r_i^{(1)}, r_j^{(2)})$$

$$Pe(i, j) = Pm(i, j) + Pi(i, j) + Pd(i, j) + Pn(i, j)$$

$$P(i, j) = \sum_{k=1}^{i} \sum_{l=1}^{j} Pe(k, l)$$

The initial conditions are: $Pm(i,0)$, $Pi(i,0)$, $Pd(i,0)$, $Pn(i,0)$, and $Pe(i,0) = 0 \ \forall i$ and $Pm(0, j)$, $Pi(0, j)$, $Pd(0, j)$, $Pn(0, j)$, and $Pe(0, j) = 0 \ \forall j$. The denominator of equation 1, the summation over all possible alignments, can be computed in a similar manner as the recursive algorithm above by replacing $\Psi(r_i^{(1)}, r_j^{(2)})$ by 1.

The backward recursive algorithm for sampling alignments from their exact posterior distribution is comparable to the algorithm used by the 'Bayes Aligner' to obtain the posterior alignment distribution (Zhu *et al.*, 1998). Sampling an alignment can be broken down into three steps: (i) The parameters Θ and Λ are sampled from their exact posterior distribution, $P(\Theta, \Lambda \mid R^{(1)}, R^{(2)})$, obtained from the forward algorithm; (ii) Conditioning on Θ and Λ, an endpoint from which to start the backtrace is sampled; and (iii) Each choice, match, insertion, deletion, or beginning a new alignment, is sampled dependent upon the previous choice. The sample ends when a new alignment is chosen. Each sample gives a choice of a specific alignment, A^*, yielding an estimate for $P(A \mid R^{(1)}, R^{(2)})$.

4. Results

A major development in optimal local sequence alignment was the development of probabilistic scoring schemes to take into account the dependence of score on the lengths of the two sequences. BALSA includes terms that adjust for variations in sequence length. Specifically, the denominator of the likelihood function $\sum_{A'} \lambda_o^{k_g(A')} \lambda_e^{l_g(A')-k_g(A')}$ is only dependent upon the lengths of the two sequences, not the amino acids they consist of. An analysis of BALSA score versus sequence length returned no correlation between the two. Thus homology can be determined directly from the BALSA score.

The performance of sequence alignment algorithms has been evaluated and compared on several occasions. The two most extensive comparisons were by Pearson (1995) and Brenner (1998). Both found SSEARCH with optimally chosen gap penalties and E-values to outperform the others, correctly identifying the most remote homologs in the database. In order to keep consistency with earlier evaluations, the same databases used by Brenner, PDB40D-B and PDB90D-B, were employed in the evaluation of our alignment procedure, BALSA. PDB40D-B and PDB90D-B include only domains less than 40% and 90% identical to any of the others and contain 1323 and 2079 domains respectively. In an all-vs-all comparison there are 1,749,006 ordered pairs of which approximately, 9,044, 0.5%, are distantly related for PDB40D-B. PDB90D-B consists of 4,320,162 ordered pairs of which 53,988, or approximately 1.2%, are distantly related (Brenner *et al.*, 1998).

An all-vs-all comparison of the database was conducted using BALSA, the results were sorted in descending order, and a cutoff was drawn at which the number of related sequences above the threshold was acceptable with a given error rate of, false positives. The various thresholds and related errors were evaluated by utilizing coverage vs. error plots as in Brenner (1998). The coverage corresponds to the fraction of homologs detected at a specified error rate, errors per query (EPQ), which is the number of nonhomologs above the cutoff divided by the total number of query sequences.

In previous studies SSEARCH with E-values detected 18.4% of homologous pairs at a 1% EPQ on PDB40D-B and 38% for PDB90D-B. BALSA with the set of four matrix/gap penalty pairs previously defined obtained 19.8% and 41.3% of all relationships at the same error rate for the two respective databases. In evaluating the nonoverlaping sequences between the two databases, sequences with between 41 and 90% sequence identity, an even larger increase in detecting homologous pairs was noted; 60% for SSEARCH to 67.2% for BALSA at a 1% EPQ. In fact BALSA outperformed SSEARCH at all EPQ levels for all three databases, Figure 1. Using the same parameters as SSEARCH, BALSA found 19.2% of the homologous pairs for PDB40D-B.

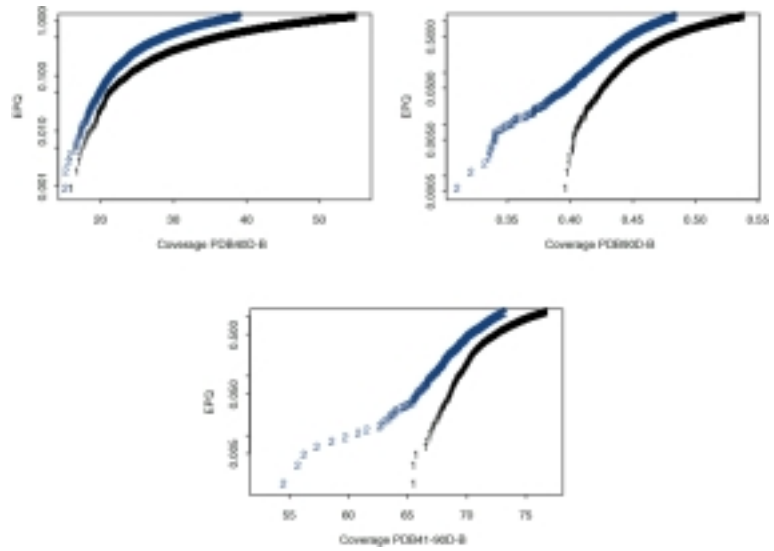

Figure 1: Coverage vs. errors per query (EPQ) plots of BALSA with four given matrix gap parameter pairs and SSEARCH with optimal gap parameters and E-values. BALSA obtained a larger coverage, detection of more homologous pairs than SSEARCH at all EPQ levels for PDB40D-B, PDB90D-B, and PDB41-90D-B.

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990) Basic Local Alignment Tool. *J. Mol. Biol.*, **215,** 403-410.

Brenner, S., Chothia, C. and Hubbard, T. J. P. (1998) Assessing Sequence Comparison Methods with Reliable Structurally Identified Distant Evolutionary Relationships. *Proc. Natl. Acad. Sci. USA*, **95,** 6073-6078.

Liu, J. S. and Lawrence, C. E. (1999) Bayesian Inference on Biopolymer Models. *Bioinformatics*, **15**(1)**,** 38-52.

Needleman, S. B. and Wunsch, C. D. (1970) A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. *J. Mol. Biol.*, **48,** 443-453.

Pearson, W. R. and Lipman, D. J. (1988) Improved Tools for Biological Sequence Comparison. *Proc. Natl. Acad. Sci. USA*, **85,** 2444-2448.

Pearson, W. R. (1995) Comparison of Methods for Searching Protein Sequence Databases. *Protein Science*, **4,** 1145-1160.

Smith, T.F. and Waterman, M.S. (1981) Identification of Common Molecular Subsequences. *J. Mol. Biol.*, **147**, 195-197.

RESUME

L'alignement de sequence locale est un outil fondamental dans la recherche bio-medicale et, malgre son existence ancienne, il persiste encore de nombreux points faibles bien reconnus. Les methodes de programmation dynamique produisent un alignement simple, quoique optimal, qui peut etre serieusement affecte par les parametres choisis. De plus, les resultats obtenus dependent de la longueur des sequences alignees, ce qui exige une analyse complementaire de conversion. Nous avons entrepris une methode pour examiner l'utilite des statistiques bayesienne pour surmonter ces difficultes. Nous avons developpe un algorithme bayesienne pour l'alignement de sequence locale, appele BALSA, qui fait etat des incertitudes associees avec toute variable inconnue en incorporant dans ses calculs avances de parametres et tout alignement possible. L'algorithme renvoie un specimen qui represente les alignements et les probabilites suivantes de parametres. En outre, cet algorithme s'adapte fondamentalement aux variations dans les longueurs de sequence. BALSA a ete compare a SSEARCH avec Valeurs " E" qui a ete jusqu'a present le meilleur algorithme de programmation dynamique en operation pour detecter les homologues.