

Automatic Parameter Selection for Sequence Similarity Search

Jianping Zhou
Dept. of Computer Science
Univ. of Mass., Lowell
jzhou@cs.uml.edu

Gary Livingston
Dept. of Computer Science
Univ. of Mass., Lowell
gary@cs.uml.edu

Georges Grinstein
Dept. of Computer Science
Univ. of Mass, Lowell
grinstein@cs.uml.edu

Abstract

We show that simulated annealing search can be used to automatically select parameters and find highly similar data regions using a modified version of the DNA-DNA Sequence Similarity Search program. We call this modified program AutoSimS. We use the average score of high-scoring chains to measure the goodness of the resulting sequence similarity search, and use adaptive simulated annealing to perform automatic search within a space of parameter values to maximize this goodness measure. We tested our program using pairs of DNA sequences, and the results show that although close-to-optimal parameter settings are very difficult to find manually, there are many different parameter settings that yield close-to-optimal search results. We suggest that our approach is able to successfully and automatically select parameters for programs used to finding close-to-optimal solutions, such as highly similar sequence regions.

1. Introduction

The DNA-DNA search or similarity algorithm (DDS/SIM) is a hash-based pairwise sequence comparison algorithm, having integrated features from Smith-Waterman, BLAST, FastA, and Haste (Hash-Accelerated Search) [1]. DDS/SIM's inherent ability to handle gaps and multiple high-scoring pairs make it attractive. Its use of several efficient computational techniques, including dynamic programming and hashing, make it particularly effective for sequence screening with linear space complexity. It has been rated as one of fastest and least space consuming tools for universal sequence alignment [2].

However, the DDS/SIM algorithm appears to be seldom used. We conjecture that the main reason is that its 11 parameters or cutoffs need to be manually set, which often require the use of heuristics obtained by experience or several trials. The optimal settings for these parameters are highly dependent on the sequences and their type.

Our experiments show that DDS/SIM performance is sensitive to parameter setting, with parameters affecting the quality of the search results, run time, and memory usage. These parameters include sliding window size w , distance cutoffs $d1$ and $d2$, segment score cutoff d , extension drop cutoff $d3$, overlap score cutoff ci , and chain score cutoff f , as well as scores m for symbol match and u for symbol mismatch, penalties p for open gap and r for repetition gap. Some products, such as Paracel PFP, PCP, and Pedant-Pro Sequence Analysis Suite, use DDS/SIM either through default settings based on expert experience (which may be inappropriate to many applications), or by being set in a custom manner by users.

2. Parameter Selection Problem

The output of the DDS/SIM similarity search is a one-dimensional array whose elements are called high-scoring chains. The size of the array depends on parameter settings. Every high-scoring chain represents a pair of similar regions in the sequences being searched, and the chain's score measures how similar the regions are. Since the goal of sequence similarity search is to find a search result with as many similar regions each with as high a score as possible, we use the average of high-scoring chain scores to measure the "goodness" of the search result.

When automatically selecting parameters for a program built on non-linear models and describing complex behavior, it is very important to retain and respect the nonlinearities inherent in these models, as they are probably present in the complex systems they model. But this requirement conflicts with feasibility of computation. Simulated annealing handles the fitting of nonlinear models and attempts to search for appropriate settings by simulating the metallurgic annealing process. Statistically, simulated annealing attempts to find the close-to-optimal fit of a nonlinear constrained non-convex cost-function over a D-dimensional space.

An alternative to selecting parameters, the wrapper approach [5], requires almost as much tuning as the parameter selection process itself.

For the DDS/SIM parameter selection problem, we use adaptive simulated annealing (ASA) [3] and an annealing schedule with temperature T decreasing exponentially with annealing-time t , $T = 20 * \exp(-0.005*t)$ if $t < 100$ and 0 otherwise [4]. ASA's use of re-annealing aids adaptation to changing sensitivities in the multi-dimensional parameter-space. In addition, ASA has over 100 options that may be used to provide robust tuning over many classes of nonlinear stochastic systems.

Our modified version of DDS/SIM, called AutoSimS (Automatic Sequence Similarity Search), uses ASA to automatically select parameters and find highly similar regions within a given search range. In the program, ASA functions as a wrapper around DDS/SIM (see Fig 1). Table 1 below shows three run results over a 100 and 200 symbol long pair of DNA sequences within the parameter ranges: $w = 3$ to 14, $d1 = 200$ to 219, $d2 = 20$ to 23, $d3 = 5$ to 7, $ci = -10$, $f = 3$ to 17, $m = 2$, $u = -5$, $p = -10$, and $r = -2$.

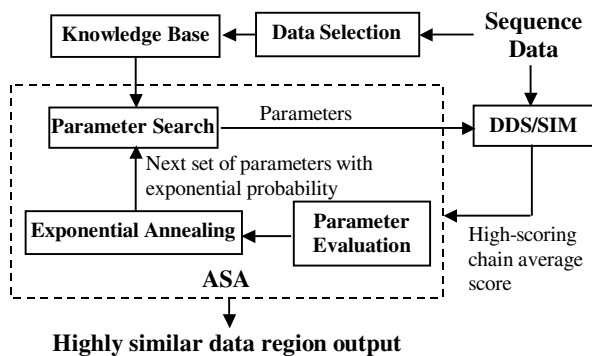


Figure 1: AutoSimS Program Flow

Table 1: Automatic Parameter Selections and Results

Result	1 st run	2 nd run	3 rd run
w (sliding window size)	9	13	10
d1 (far distance cutoff)	212	208	205
d2 (near distance cutoff)	22	23	23
d3 (extension drop cutoff)	5	5	5
f (chain score cutoff)	3	13	3
Average score of high-scoring chain	124.6	123.7	146.7

It is well known that statistical searches usually work well for problems with multiple fits close to the optimal. Referring to our experimental data in Table 1, the high scoring chain average score search results are close, but the parameter selections are quite different. We suggest that ASA is very suitable for resolving the DDS/SIM parameter selection problem. We also suggest it is very hard to predict any

relationship between the search results and parameter selection. It is this property that makes manual selection of parameters difficult.

3. Future Work

To make the AutoSimS program fully functional, a knowledge base and a data selection utility are necessary. The knowledge base and selection utility provide and use rules for selecting parameter search ranges or specified parameters in term of the identified sequence data type.

The performance for large-scale sequence data might be a concern, and techniques such as the stochastic beam search algorithm and parallel simulated annealing may need to be incorporated.

4. Conclusion

Most sequence similarity search programs, including the widely used BLAST and FASTA, need user-determined parameter or cutoff selections. Our AutoSimS program, an integration of DDS/SIM and ASA, is able to automatically select the parameters and find highly similar regions within a given search range.

We showed a successful application of ASA for automatic parameter selection in sequence similarity search. The techniques we used in our program can also be applied to other bioinformatics analysis programs including alignment and clustering.

References

- [1] Huang, X. and Miller, W. (1991) A Time-Efficient, Linear-Space Local Similarity Algorithm. *Advances in Applied Mathematics* 12, 337-357.
- [2] Tech Topics, Michigan Technological University, Nov. 3, 1995, Vol. XXVIII, No.9
- [3] Ingber, L., Adaptive Simulated Annealing, Special Issue of the Polish Journal Control and Cybernetics on Simulated Annealing Applied to Combinatorial Optimization, 1995
- [4] Russel, S. and Norvig, P., *Artificial Intelligence, A Modern Approach*, second edition, Prentice Hall, 2003. ISBN 0-13-790395-2
- [5] Kohavi, R. and John, G.H., (1995), Automatic Parameter Selection by Minimizing Estimated Error, in Prieditis & Russell, *Proceedings of the 12th International Conference on Machine Learning*, Morgan Kaufman, SF.