Quality of Algorithms for Sequence Comparison

Mikhail Roytberg[1,2]

[1] Institute of Mathematical Problems in Biology RAS, Institutskaya, 4, Pushchino, Moscow Region, 142290, Russia
[2] National Research University Higher School of Economics, Myasnitskaya, 20, Moscow, 101000, Russia
mroytberg@lpm.org.ru

Abstract. Pair-wise sequence alignment is the basic method of comparative analysis of proteins and nucleic acids. Studying the results of the alignment one has to consider two questions: (1) did the program find all the interesting similarities ("sensitivity") and (2) are all the found similarities interesting ("selectivity"). Definitely, one has to specify, what alignments are considered as the interesting ones. Analogous questions can be addressed to each of the obtained alignments: (3) which part of the aligned positions are aligned correctly ("confidence") and (4) does alignment contain all pairs of the corresponding positions of compared sequences ("accuracy"). Naturally, the answer on the questions depends on the definition of the correct alignment. The presentation addresses the above two pairs of questions that are extremely important in interpreting of the results of sequence comparison.

Keywords: alignment, seed, sequence comparison, sensitivity, selectivity, accuracy, confidence.

1 Seeds, Sensitivity and Selectivity

Many programs of sequence similarity search (e.g. BLAST, FASTA) are based on the filtration paradigm; they firstly mark the regions of putative similarity and then restrict the search with the regions only. To perform the first step the seeding scheme is usually implemented: one searches only for the similarities containing the strong similarity of special form, e.g. the similarities containing k consecutive matches. This seeding scheme leads to the drastic speed up compared to the more rigorous dynamic programming based methods at the price of possible loss of some interesting similarities.

In the framework of similarity search in biological sequences, a *seed* specifies a class of short sequence motifs which, if shared by two sequences, are assumed to witness a potential similarity. We say that a seed *matches* a similarity (or a similarity is recognized by a seed) if it contains a sub-similarity corresponding to a seed. To define what is sensitivity and selectivity of a seed we have to make some preliminary definitions. First, we have to describe the set of considered possible sequence alignments and the subset of interesting similarities ("target similarities"). For example, we may consider all ungapped similarities

S.O. Kuznetsov et al. (Eds.): PReMI 2011, LNCS 6744, pp. 17–20, 2011.

(alignments) of a given length and the set of target similarities consisting of all ungapped alignments having identity level higher than a given cut-off. Second, we have to consider the two probability distributions on alignments: the *background* distribution, corresponding to the random alignments and the *foreground* distribution that corresponds to the target alignments. E.g. one can consider both distributions as Bernoulli distributions in two-letter alphabet (match-mismatch) and define the probability of match as 0.25 for the background distribution and as (say) 0.7 for the foreground distribution.

Given the set of target alignments and the distributions, the *sensitivity* of a seed is the probability that a random similarity is recognized by a seed according to a foreground distribution and the *selectivity* of a seed is the probability that a random similarity is recognized by a seed according to a background distribution. For the Bernoulli distribution the selectivity is often defined as a probability that a seeding similarity can be found for two random independent sequences of a length equal to the seeds length.

The seed implemented in BLASTN program [1] describes a class of k consecutive matches (default $k = 11$). The selectivity of the default seed is $0.25^k = 0.25^{11} \sim 10^{-6}$. The sensitivity of the seed for ungapped nucleotide similarities of length 64 with 70% identity is ~ 0.3. Several years ago Ma, Tromp and Li [2] have proposed to use k nonconsecutive letters as a seed. This change surprisingly led to a significant improvement of sensitivity without loss of selectivity that depends only on the desired number of matches k and on the background match probability. E.g. the seed `110100110010101111` (`1` stands for the match positions and `0` stands for "spaces") has the sensitivity 0.46 with the same number of matches $k = 11$. The seminal work of Ma, Tromp and Li (2002) have caused the investigation of various seed models both for nucleic and amino acid sequences, e.g. vector seeds, subset seeds, multyseeds, etc [3]-[12].

We will consider advantages and disadvantages of the models and will present the unifying framework to compute the seed sensitivity.

2 Alignments, Accuracy and Confidence

For many applications it is important to evaluate the quality of algorithmically obtained alignments, i.e. how close the algorithmic alignment is to the evolutionarily true one. Here the evolutionarily true alignment is an alignment superimposing the positions originating from the same position of the common predecessor [13].

Moreover, it is important not only to know the quantitative measure of the average similarity of alignments but also to understand the typical differences between the algorithmic and the evolutionary true alignments. However, the evolutionarily true alignment of given sequences is usually unknown, and thus an approximation is needed.

There are two possible ways to obtain such an approximation: (1) to use artificial sequences pairs obtained according to a proper evolutionary model [14,15] and (2) to use alignments based on the superposition of the protein 3D-structures (that is possible only for the comparison of amino acid sequences) [13,16].

Accuracy and confidence of global and local alignments were studied in several papers [13,14], [16]-[19]. The data show that the main difference between the algorithmic and true alignments is the number of gaps while the average length of a gap is approximately the same. Surprisingly, the 3D-structure based protein alignments contain significant number of ungapped fragments of negative score that can not be restored in algorithmic alignments.

The significant gain both in accuracy and in confidence of protein alignments can be achieved using the information on the secondary structure (experimentally obtained or predicted) [20,21].

Acknowledgments

The work was supported by grant RFBR 09-04-01053-a.

References

1. Altschul, S.F., Gish, W., Miller, W., et al.: Basic local alignment search tool. J. Mol. Biol. 215, 403–410 (1990)
2. Ma, B., Tromp, J., Li, M.: PatternHunter: Fasterand more sensitive homology search. Bioinformatics 18(3), 440–445 (2002)
3. Brejová, B., Brown, D.G., Vinař, T.: Optimal Spaced Seeds for Hidden Markov Models, with Application to Homologous Coding Regions. In: Baeza-Yates, R., Chávez, E., Crochemore, M. (eds.) CPM 2003. LNCS, vol. 2676, pp. 42–54. Springer, Heidelberg (2003)
4. Brejová, B., Brown, D.G., Vinař, T.: Vector seeds: An extension to spaced seeds allows substantial improvements in sensitivity and specificity. In: Benson, G., Page, R.D.M. (eds.) WABI 2003. LNCS (LNBI), vol. 2812, pp. 39–54. Springer, Heidelberg (2003)
5. Brejova, B., Brown, D., Vinar, T.: Optimal spaced seeds for homologous coding regions. Journal of Bioinformatics and Computational Biology 1(4), 595–610 (2004)
6. Brown, D.: Optimizing multiple seeds for protein homology search. IEEE Transactions on Computational Biology and Bioinformatics 2(1), 29–38 (2005)
7. Buhler, J., Keich, U., Sun, Y.: Designing seeds for similarity search in genomic DNA. In: Proceedings of the 7th Annual International Conference on Computational Molecular Biology (RECOMB 2003), Berlin, Germany, April 2003, pp. 67–75. ACM Press, New York (2003)
8. Kucherov, G., Noé, L., Roytberg, M.: Multiseed lossless filtration. IEEE Transactions on Computational Biology and Bioinformatics 2(1), 51–61 (2005)
9. Li, M., Ma, B., Kisman, D., Tromp, J.: Pattern Hunter II: Highly sensitive and fast homology search. Journal of Bioinformatics and Computational Biology (2004), Earlier version in GIW 2003 (International Conference on Genome Informatics)
10. Kucherov, G., Noé, L., Roytberg, M.: A unifying framework for seed sensitivity and its application to subset seeds. Journal of Bioinformatics and Computational Biology 4(2), 553–569 (2006)
11. Xu, J., Brown, D.G., Li, M., Ma, B.: Optimizing Multiple Spaced Seeds for Homology Search. In: Sahinalp, S.C., Muthukrishnan, S.M., Dogrusoz, U. (eds.) CPM 2004. LNCS, vol. 3109, pp. 47–58. Springer, Heidelberg (2004)

12. Yang, I., Wang, S., Chen, Y., Huang, P., Ye, L., Huang, X., Chao, K.: Efficient methods for generating optimal single and multiple spaced seeds. In: Proceedings of the IEEE 4th Symposium on Bioinformatics and Bioengineering(BIBE 2004), Taichung, Taiwan, May 19-21, 2004, pp. 411–416. IEEE Computer Society Press, Los Alamitos (2004)
13. Sunyaev, Bogopolsky, G.A., Oleynikova, N.V., Vlasov, P.K., Finkelstein, A.V., Roytberg, M.A.: From Analysis of Protein Structural Alignments Toward a Novel Approach to Align Protein Sequences. PROTEINS: Structure, Function, and Bioinformatics 54(3), 569–582 (2004)
14. Stoye, J., Evers, D., Meyer, F.: Rose: generating sequence families. Bioinformatics 14, 157–163 (1998)
15. Polyanovsky, V., Roytberg, M., Tumanyan, V.: Reconstruction of Genuine Pair-Wise Sequence Alignment. J. Comput. Biol. (April 24, 2008) (Epub ahead of print)
16. Vogt, G., Etzold, T., Argos, P.: An assessment of amino acid exchange matrices in aligning protein sequences: the twilight zone revisited. J. Mol. Biol. 249, 816–831 (1995)
17. Domingues, F.S., Lackner, P., Andreeva, A., et al.: Structure-based evaluation of sequence comparison and fold recognition alignment accuracy. J. Mol. Biol. 297, 1003–1013 (2000)
18. Mevissen, H.T., Vingron, M.: Quantifying the local reliability of a sequence alignment. Prot. Eng. 9, 127–132 (1996)
19. Vingron, M., Argos, P.: Determination of reliable regions in protein sequence alignments. Prot. Eng. 3, 565–569 (1990)
20. Litvinov, I.I., Lobanov, Yu, M., Mironov, A.A., et al.: Information on the Secondary Structure Improves the Quality of Protein Sequence Alignment. Mol. Biol. 40, 474–480 (2006)
21. Wallqvist, A., Fukunishi, Y., Murphy, L.R., et al.: Iterative sequence/secondary structure search for protein homologs: Comparison with amino acid sequence alignments and application to fold recognition in genome databases. Bioinformatics 16, 988–1002 (2000)