Neural Networks and Bioinformatics

Term paper 498Bio; Peter Fleck; 12/11/2001

Sequence alignment (SA) of DNA, RNA and protein primary structure forms an integral, if not the most important part of bioinformatics. We have discussed an overview of SA methods in this course [1], which share the procedure of finding the "best" match of two given strings of letters and evaluating its quality or extent of agreement consistently with alternative alignments of the same two strings or those of other strings. The evaluation consists of assigning scores to matches and mismatches of individual letters plus finally adding these up, where both the choice of the scores and the minimum search in the space of possible alignments are constructed from statistical methods.

Let me draw your attention to two particular points:

- These methods tacitly assume a particular model of evolution when claiming to map evolutionary distances between species with the obtained scores, namely that of random independent point mutations in either string.

- Any scoring method designed according to the principles sketched above will at best rank alignments exactly according to the evolutionary model assumed, but cannot give any feedback to the quality of this assumption.

As mapping of evolutionary relationships not yet being an exact science so far, an elaborate or even "perfect" scoring system might even get across a false sense of certainty regarding the evolutionary distance of two particular strings.

Considering now the complications arising to construct such a "perfect" method together with some remaining uncertainty about the accuracy of the underlying model of evolution and gotten inspired by the fact, that most currently used heuristic methods get gauged by known examples of qualitatively good, intermediate or bad alignment, the author of this paper asked himself, whether artificial neural networks (ANN) might provide an alternative solution to this problem. These as well have to initially "learn" how to rank alignments, but

- do not a priori assume a particular mechanism of evolution, and
- might be able to tolerate non-standard deviations of an overall good alignment when evaluating.

Evaluating alignments in this manner would obviously be not exact and depend in functionality and accuracy on the design of the network and the examples it got trained with. On the other hand, this very dependence on the "teaching" allows in principle to train ANNs to rank alignments according to any evolutionary model of interest, and probably even more importantly, can help reveal yet unknown evolutionary mechanisms: assume the case a close evolutionary distance between two particular DNA or protein sequence were known independent of the alignment; training a network with this and like examples plus analyzing it afterwards might yield additional understanding of the original development.

Artificial neural networks have been used intensively in bioinformatics[2], except in the context of sequence alignment. There is no indication whether this suggestion has been discarded due to inherent problems, for mere reasons of impracticability, due to it being too ambitious to use such a technique in this context or simply hasn't yet been thought of. In any case this might offer fertile ground for research.

Next it might be helpful to take a look at previous applications of ANNs in bio- or genome informatics, leading to a review of a recent introduction into the topic[2] in the following.

ANNs have been used "to predict protein secondary and tertiary structure, to distinguish protein encoding regions from non-coding sequences, to predict bacterial promoter sequences, and to classify molecular sequences" (Introduction). The book quickly introduces the basic concepts of ANNs, i.e. possible functionalities of a single neuron (chap. 2), basic networking and layers (chap. 3), frequently used network architectures (chap. 4) and learning procedures (chap. 5). More advanced topics are briefly summarized in context, i.e. in a later chapter (chap. 8) or whenever the need arises. The literature on using ANNs in nucleic acid sequence analysis (chap. 9), protein structure prediction (chap. 10) and protein sequence analysis (chap. 11) is reviewed subsequently, providing the reader rather with an guideline through the existing literature than a hands-on tutorial how to use the methods or discuss them in a mutual context. Finally future directions of the field including a possible combination with statistical methods (chap. 12) and progress in feature extraction from a network (chap. 13) are briefly discussed. This

book therefore comprises a quick introduction into the overall field, leaving to extract the details for an application of interest from the existing literature to the reader.

Artificial neural networks in bioinformatics therefore seem to have been used rather in a traditional manner, namely for pattern recognition of various kinds instead of evaluative comparison of sets or sequences. Indications for the impossibility to realize this paper's suggestion have not been found so far, which may encourage its reader to cultivate this field.

References

[1] N. Goldenfeld. 498bio. lecture, fall 2001.

[2] Cathy H. Wu and Jerry W. McLarty. *Neural Networks and Genome Informatics*, volume 1 of *Methods in Computational Biology and Biochemistry*. Elsevier Science Ltd, Oxford, 2000.