Species Identification through DNA String Analysis - Literature Review

Mark Vorster

Supervisor: Prof. Philip Machanick

29 May 2012

1 Introduction

Since DNA sequences where stored in the first bioinformatics databases computers have been aiding and speeding up the process of analysis for bioinformaticians. The Rhodes University Department of Biochemistry, Microbiology & Biotechnology has found need to identify the distinct species of bacteria given a large set of DNA sequences, however they have identified a means to decrease the time taken in the processing of bacterial DNA for phylogenetic analysis. While the algorithms work well for small numbers of samples increasing the number of samples into the tens of thousands has started to take an inordinate amount of time. After examining the broad area of bioinformatics focusing on framing the problem, this paper looks at how progression in the fundamentals of genetics along with the rise of computers began bioinformatics. It then focuses more specifically on sequence analysis including mentioning the problems of sequence alignment and phylogenetic analysis. Finally the problem is expanded upon.

2 Bioinformatics

The Biomedical Information Science and Technology Initiatives Definition Committee which was chaired by Dr Huerta defined Bioinformatics in the year 2000 as

> Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioural or health data, including those to acquire, store, organize, archive, analyse, or visualise such data. [7]

A more concise definition is found in the Oxford English Dictionary:

> The branch of science concerned with information and information flow in biological systems, esp. the use of computational methods in genetics and genomics. [1]

While Biology, the study of life, has fascinated scientists and philosophers for hundreds of years, originating as the passing on of knowledge of plants and animals for survival purposes it progressed into medicine, agriculture, botany and zoology to name a few. At this stage the classification of living organisms was done in a qualitative manner through observation. As an example from an observation that animals look the same it can be deduced that they are of the same species, or the grouping of all animals with mammary glands being called mammals. The development of the microscope allowed biologists to move beyond the reach of the naked eye, the study of cells became paramount to all facets of biology. [8]

Further technological advances in the 19th century in genetics and cellular biology and 20th century advances in molecular biology as explained by Xu et. al. [15] has made it possible to observe the fundamental units of biology and genetics. Being able to quantify hereditary information allows the use of computers to aid in the analysis of the data.

2.1 Genetics

Where biology is the science of life and genetics is the study of hereditary through genes, the fundamental units of hereditary. All life goes through a reproductive cycle and the DNA chain molecules hold the inherited informa-

tion that passes from generation to generation. The macromolecules form a ladder-like helix from two strands, the runs of which are formed by pairs of a nitrogenous bases either adenine and thymine or guanine with cytosine.[1] Baldi and Brunak [3] name a fundamental feature of these chain molecules as their ability to be represented digitally. The representation most commonly used represents the bases with four letters: A, T, G and C [13, p. 18]. Reviewing this information from the bottom up, the bases together form genes of which there are multiple in each chromosome and carry the information required for cell growth, division and function.

2.2 History of Bioinformatics

Now that we have discussed some concepts we can look at where bioinformatics started. While the term bioinformatics only came about in the 1990s work which would now be classified under it began many years before. Watson, Crick, Wilkins and Franklin first discovered the structure of DNA in 1953 [5] [6, p. 163] and they quickly understood how it allowed a copying mechanism for the hereditary information. The structure of DNA enables an abstraction of the molecule down to just four base units which allows for an easy digital representation and this is where bioinformatics begins. The storing of this information in databases was the first problem posed to bioinformaticians due largely to the sheer volume of data, [10, p. 7] the number of base pairs in a DNA molecule ranges from thousands to hundreds of thousands, but this, as pointed out by Tramontano, [13, p. 6] would hardly be

[1] There is an additional base, 'uracil', that forms part of RNA sequences but for simplicity it not in the scope of this project.

useful without annotations of the knowledge of each sequence alongside it.

3 Sequence Analysis

The analytical study of DNA, RNA or any amino acid sequence are included under the umbrella of sequence analysis. Won, Park, Yoon and Kim identify the reason for DNA sequence analysis as being that often knowledge can be inferred about newly sequenced samples through knowledge of functions of other sequences [14]. Tramontano too describes how inference can be made from similar sequences to homology, as homologous sequences are expected to have higher similarity than unrelated sequences [13, p. 77]. Sequence analysis is more general than this. Gibas and Jambeck list the 5 main types of sequence analysis as :

- Knowledge-based single sequence analysis for sequence characteristics.
- Pairwise sequence comparison and sequence-based searching.
- Multiple sequence alignment.
- Sequence motif discovery in multiple alignments.
- Phylogenetic inference.

[6, pp. 159,160]

3.1 Sequence Alignment

One of the major issues in sequence analysis is the alignment of sequences, and while for the purpose of this project the sequences will be assumed to be

globally aligned, it is still important to understand. Brudno et. al. recognise sequence alignment among the most successful applications of Computer Science in Bioinformatics [4]. Before any meaningful analysis into the genetic relationship between genes can take place sequences must be aligned [12, p. 771]. The process of alignment is often simply matching the positions of sequences to their least different (or most similar) locations. This is particularly important with partial samples where one sample is being aligned within another with unknown starting locations. This can be very computationally expensive, as not only does each potential starting location need to be examined, but due to the nature of genes there is inherent fuzziness, bases can be inserted or deleted in mutations, so at each point there may be one or more bases extra or missing from either sequence. [13, p. 55] [11]

3.2 Phylogenetic Analysis and Species Identification

Phylogenetics is a branch of taxonomy, the study of homogeneity of organisms including determining their evolutionary relationships for example their species, that deals with numerical data such as DNA sequences. One of the main applications of phylogenetics is the construction of phylogenetic trees through studying similarities between organisms [9]. These phylogenetic trees are a visualisation of sequences' genetic relationships. The steps in building a phylogenetic tree from a set of DNA sequences is first to determine the similarity of each of the sequences, then starting with the most similar pair draw the branch of the tree and calculate the average distance between that branch and each of the remaining sequences [13, pp. 66 - 73]. This research focuses on the first step, the building of the similarity table,

which, although trivial for small numbers of sequences becomes very slow when dealing with thousands due to is exponential nature.

3.3 FASTA format

The FASTA format is the data format chosen by the Rhodes University Department of Biochemistry, Microbiology & Biotechnology to represent DNA sequence data, it is one of the most commonly used and simple formats [6, p.180]. A file in the FASTA format can represent many DNA sequences each has a single header line followed by multiple lines of sequence data. The header line requires a leading greater than character '>' and a single word which is the name of the sequence, the rest of the line is a comment or description of the sequence. The data can contain multiple lines, new line characters are ignored, and whitespace, periods or underscores can have application specific meaning. Below is an example of a single sequence in the FASTA format.

```
>SequenceName description of the sequence
CCGGAATACCTAGGACATAGCAGAGGCGTCTTGCCTATACAG
TGTTTTTCTCCGAGACGCCTGATTACCTGCTAGTCGGGATGA
TAACCAAGAATTTGTGTCTGCTGCGCGCCATTTGCCAACCGA
GCCTTCATCCCCCGCCGGTCTGTGATGTCCCAATGGACCGGA
```

4 String Matching

String matching in computer science is applied to many fields, largely in text analysis but also in speech or character recognition and many others as patterns can be similarity found in the binary data. Approximate string matching algorithms are used when an exact copy is expected, such as with determining the similarity of sequences. Baase [2, pp. 504 - 508] discusses an algorithm for building a difference table between two strings taking into account insertions and deletions as they may occur in DNA sequences. The approach uses a dynamic programing technique to speed up the process building a two dimensional matrix where each point D[i][j] has the minimum number of differences between two strings segments P and T each ending at p_i and t_j respectively. The matrix is built up column by column where D[i][j] is calculated as the minimum of three possible numbers either a 'matchCost' (if $p_i = t_j$) or a 'reviseCost'(if $p_i \neq t_j$), a 'insertCost' and a 'deleteCost' where each is defined as the following:

$$\textbf{matchCost} = D[i-1][j-1] \quad \text{, if } p_i = t_j \text{ or}$$

$$\textbf{reviseCost} = D[i-1][j-1] + 1 \quad \text{, if } p_i \neq t_j$$

$$\textbf{insertCost} = D[i-1][j] + 1$$

$$\textbf{deleteCost} = D[i][j-1] + 1$$

5 Problem Discussion

The Rhodes University Department of Biochemistry, Microbiology & Biotechnology have taken large aquatic samples of bacterial DNA on which they wish to do some phylogenetic analysis. They have found their existing tools lacking in the area of identifying the different bacterial species from the samples. They have found the gene samples have areas of highly conserved and other highly variable sections, by focusing on these sections it should be possible to greatly reduce the amount of time to process the samples. The samples have therefore been processed, simplifying the problem as the sequences are assumed to be aligned to the same location. They require the sequences to be grouped with no more than one percentage difference. The proposed method is to calculate the differences with the approximate string matching algorithm. At least two efficiencies can be made on this algorithm due to the assumptions mentioned above. Firstly as we are searching for similarities within one percentage difference branches can be pruned from the search tree once D[i][j] exceeds the threshold, as it is the minimum difference at that point in the matrix no further processing is needed. Secondly as the sequences are assumed to begin at the same location, the first row can include a dependency on the previous column's first row such that D[0][j] = D[0][j-1] if $p_i = t_j$ or D[0][j-1]+1 if $p_i \neq t_j$ which will reduce the possible 'starting columns' down to only one percentage of the shortest gene rather than its entire length, saving processing but more significantly space. Further efficiencies are hoped to be explored by noting that given two similar sequences a third sequence significantly different from the first need not be compared at all with the second.

References

[1] Oxford English Dictionary: Bioinformatics. [online]. Accessed on 2 April 2012. Available from: `http://www.oed.com/view/Entry/255935`.

[2] Baase, S., and Van Gelder, A. *Computer Algorithms: Introduction to Design and Analysis*. Addison-Wesley, 2000.

[3] Baldi, P., and Brunak, S. *Bioinformatics: The Machine Learning Approach*. Adaptive Computation and Machine Learning. Mit Press, 2001.

[4] Brudno, M., Do, C. B., Cooper, G. M., Kim, M. F., Davydov, E., Program, N. C. S., Green, E. D., Sidow, A., and Batzoglou, S. Lagan and multi-lagan: Efficient tools for large-scale multiple alignment of genomic dna. *Genome Research 13*, 4 (2003), 721–731.

[5] Crick, F., and Watson, J. Molecular structure of nucleic acids: A structure for dna. *Nature 171* (April 1953), 737 – 738.

[6] Gibas, C., and Jambeck, P. *Developing Bioinformatics Computer Skills*. O'Reilly Series. O'Reilly, 2001.

[7] Huerta, M., Downing, G., Haseltine, F., Seto, B., and Lie, Y. Nih working definition of bioinformatics and computational biology. Online, July 2000.

[8] Jones, N., and Pevzner, P. *An Introduction To Bioinformatics Algorithms*. Computational Molecular Biology. Mit Press, 2004.

[9] Kanehisa, M. *Post-Genome Informatics*. Post-genome Informatics. Oxford University Press, 2000.

[10] Krawetz, S., and Womble, D. *Introduction to Bioinformatics: A Theoretical And Practical Approach*. Humana Press, 2003.

[11] Li, H., and Homer, N. A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in Bioinformatics 11*, 5 (2010), 473–483.

[12] Nardone, J., Lee, D. U., Ansel, K. M., and Rao, A. Bioinformatics for the 'bench biologist': how to find regulatory regions in genomic dna. *Nature Immunology 5*, 8 (Aug. 2004), 768–774.

[13] Tramontano, A. *Introduction to Bioinformatics*. Chapman and Hall mathematics series. Chapman & Hall/CRC, 2007.

[14] Won, J.-I., Park, S., Yoon, J.-H., and Kim, S.-W. An efficient approach for sequence matching in large dna databases. *Journal of Information Science 32*, 1 (2006), 88–104.

[15] Xu, D., Keller, J., Popescu, M., and Bondugula, R. *Applications of Fuzzy Logic in Bioinformatics*. Series on Advances in Bioinformatics and Computational Biology. Imperial College Press, 2008.