# USE OF PILE TECHNOLOGY FOR WORK WITH BIOLOGICAL DATA

**Milan Tomeš**
Doctoral Degree Programme (2) , FIT BUT
E-mail: xtomes00@stud.fit.vutbr.cz

Supervised by: Jaroslav Zendulka
E-mail: zendulka@fit.vutbr.cz

## ABSTRACT

This paper deals with Pile technology and its use for work with biological data, which is the most growing fields today. This fields are very extensive and that is why research is difficult and there is a lot of problems. Interdisciplinarity and a large amount of data are typical problems of bioinformatics. Pile technology could have a great benefit in this field, because its basic advantages are working with data without redundancy, which is massive in biological data. Furthermore it is easy and fast searching, which is the basic activity there.

## 1. INTRODUCTION

Field of research and development is very rich in quantity of information, but suffer from an inability to intelligently integrate huge and complex data. Nevertheless we need to acquire knowledge from these data, for example understanding markets, or genomes.

One of the areas where this phenomenon is very emphasized is the field of bioinformatics. Today, thanks to large-scale lab procedures, we are able to predict behaviors of molecules by studying data collected all over the world. Accelerated research procedures are producing massive amounts of data, thousands of new projects were born, algorithms that have been around for decades found their role in the new field.

Gradually, our files and databases became too tight and inflexible for our needs. We are spending millions of dollars for supercomputers with terabytes of storage space to aid us in data mining, all in our race to represent nature. These investments helped us create models better or search databases faster, but will they help us represent the right way of nature?

## 2. GENERAL DESCRIPTION OF PILE

A PILE structure is a graph, which can be described as a combination of trees. A PILE object is an identification of two nodes from different trees, see Figure 1.

Pile always starts as a multiplicity of trees, growing and branching out as data is mapped. It grows in unusual ways so that branches start interacting with other branches, and even with roots, so that a set of initial trees evolves into a complex, yet very organized and coordinated network. When building more complex systems, we introduce heterogeneous da-

ta of different types and meanings simultaneously (for instance, a DNA sequence and its annotation) and want to link the two, see Figure 1. In the process, two complex networks are being created, and our structure ends up as a highly connected and co-ordinated network of heterogeneous trees (a network of networks) in which any object (such as a leaf or a branch) knows its predecessors, descendants and external links. [4]

In Pile, the Nucleic Acid is notated as the base system, defined as a system which does not contain any other subsystems, i.e., in which terminal values are basic variables, (int or char for example). Input values, regardless of their complexity, are called terminal values (TVs). Any data derived by combining TVs are referred to as "objects" in the system. Any object must have two parents — mother and father — where a single object can take the role of both father and mother, and one parent can have more children with any other parent in the system. Parents in Pile are referred to as normative (N) and associative (A); therefore one object can be N-parent, A-parent, and both at the same time. [4]

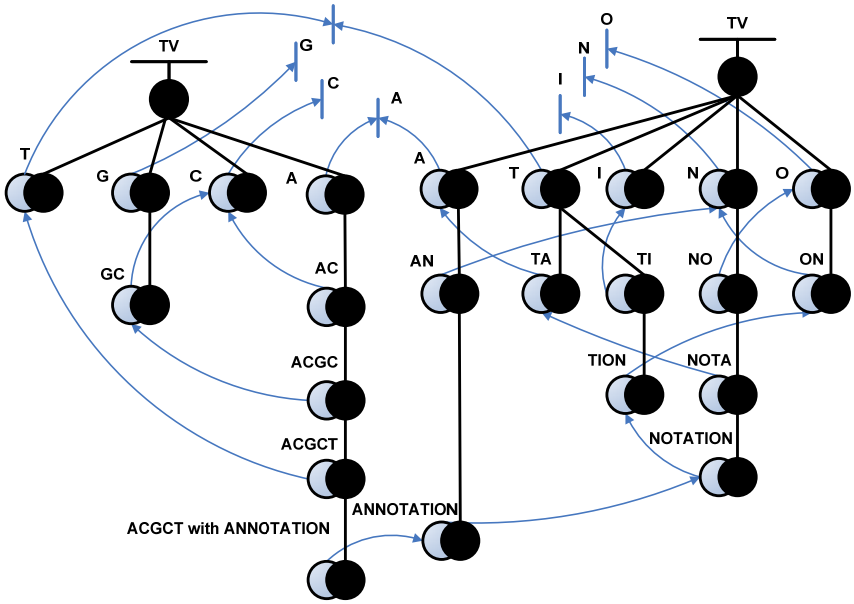


**Figure 1:** Pile structure

In addition to enabling objects to take multiple roles, in Pile every child node points to its parents, and to its children.

In addition to knowing its position within one tree (DNA sequence), an object can belong to multiple trees, and a new tree can start anywhere within the old tree's structure. This allows seamless integration of heterogeneous data, for example, DNA sequence, protein sequence and their annotation.[4]

PILE provides three main benefits:

- Saving storage capacity – because stores data without redundancy. There is a compression effect (although Pile is not a compression algorithm), which is very important at the time of exponential growth of amount of data in biological databases.
- Because is able to connect heterogeneous data (but system is clearly homogenous), facilitates finding the context inside the data. Data are assimilated within the context of their meaning, and that is why data mining have better conditions.
- The resulting structure can very well seek, quickly and without limitations

## 3. REPRESENTATION OF BIOLOGICAL DATA WITH PILE SYSTEM

In the Pile System, genomic DNA can be represented by structures that include a DNA sequence and all markers that can be inferred from the sequence. It is important to note that we are able to read most of the markers straight from nucleotide code. This can be accomplished by specifying rules that will manage DNA punctuation. We use punctuation to inform the Pile engine how to store the data upon loading, so the user, or other objects can access this data, not only as whole but also as punctuated units.

Genome and Proteome stored in Pile's way can then be accessed at any level needed. Any level can be linked with relevant data. As an example, the level of gene will be linked to its protein, EST, expression, polymorphism and disease data, which will be stored in different structures, but communicate, as they were one. Therefore, all structures are allowed to interact and support each other, even though they have their specialized role and form. Functioning in the same environment they act in perfect synergy and coordination.[2]

Pile provides a very fast substring search (exact as well as non-exact search), which is the first and simplest application of PILE, promises immediate profit for bioinformatics applications. These applications include [1]:

- searching in many databases at once (which can be connected in PILE)
- sequence alignment (comparing two DNA or protein sequences and "fitting" them)
- multiple sequence alignment – more sequences have to be "fitted" to others at once
- clustering of ESTs (pieces of genes, which have been sequenced and have to be ordered and placed in the genome)

## 4. BENCHMARKING THE PILE SYSTEM

Here are presented some results from experiments, that were performed in [3]:

- assimilating different types of input data and examining the correlation between the size of the input data, the size of the resultant Pile structure, and the number of relations created by the engine to assimilate the input data
- performing substring searches in random data files

To gain a broad perspective of the performance of the Pile System, different input data sources were used. There will be presented genetic data only. The sequences were obtained from NCBI Genbank and consisted of genomic data from Pufferfish, Rat, and Arabidopsis.

### 4.1. DATA ASSIMILATION

The parameter determined in these experiments was the dependency between the input data size and the number of relations generated by the engine to assimilate the given input data. As an auxiliary indicator to emphasize the capacity increase of the Pile structure with increasing input size, the ratio between these two values was determined, expressing the number of bytes assimilated on average in one relation generated by the Pile engine. In Figure 2, the depiction of the dependency between these two values is such that the input data size is shown on the ordinate (y-axis) as depending on the number of relations provided by the Pile system to assimilate this amount of input data. [3]

In the sections a) of the figure, the number of relations provided by the engine is shown on the x-axis, and the amount of input data assimilated by these relations is shown on the y-axis. The straight line from the origin to the upper right is the identical function f(x)=x, in-

cluded as a help for the eye, and the lines covering some of the data points represent the fitted functions. Credible fits could only be yielded for input data sizes larger than ca. 50kB. No polynomial or exponential functions could be found to faithfully approximate input data sizes smaller than this. In the b) section of the Figure 2, the input data size is given on the x-axis, while the y-axis displays the average amount of input data assimilated per relation by the Pile engine. This figure shows that the larger the amount of data already assimilated, the larger the average capacity of each new relation. [3]

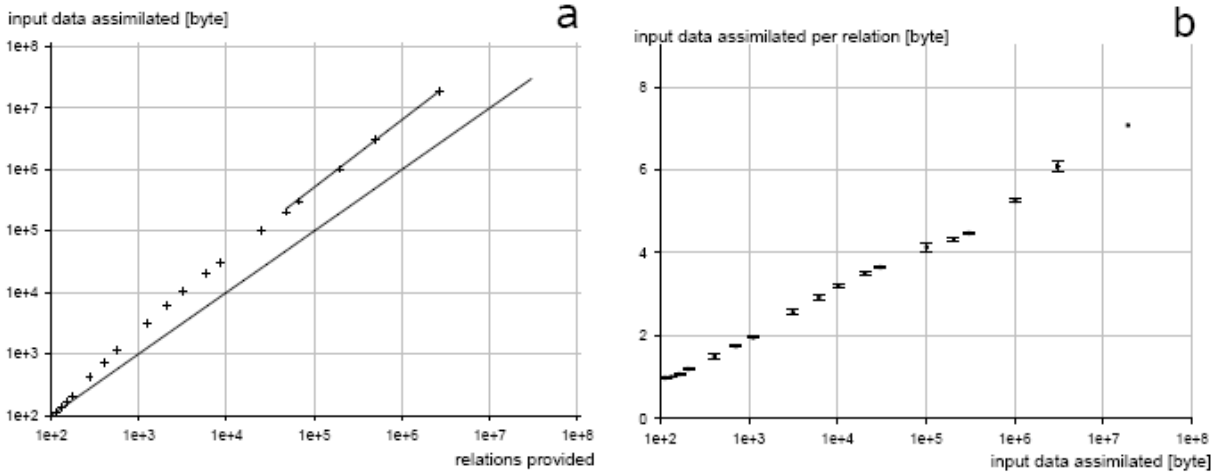


**Figure 2:** genetic data Arabidopsis

### 4.2. SUBSTRING SEARCH

Figure 3 and Figure 4 present the numerical results of the substring searches. Left and right part of Figure 3 presents the results for different substring lengths, and Figure 4 presents the same data again, with fixed file sizes for each figure, and one individual plot for each of the search string lengths. [3]

Figure 3 clearly shows that there is no obvious correlation of the retrieval times of each individual substring occurrence with the size of the file to be searched in. Neither does the number of occurrences have any quantifiable influence on the retrieval times. Figure 4 redisplays the same data again, and reemphasizes the finding, that neither the size of the file to be searched in, nor the size of the substring to be searched have any obvious influence on the retrieval times. [3]

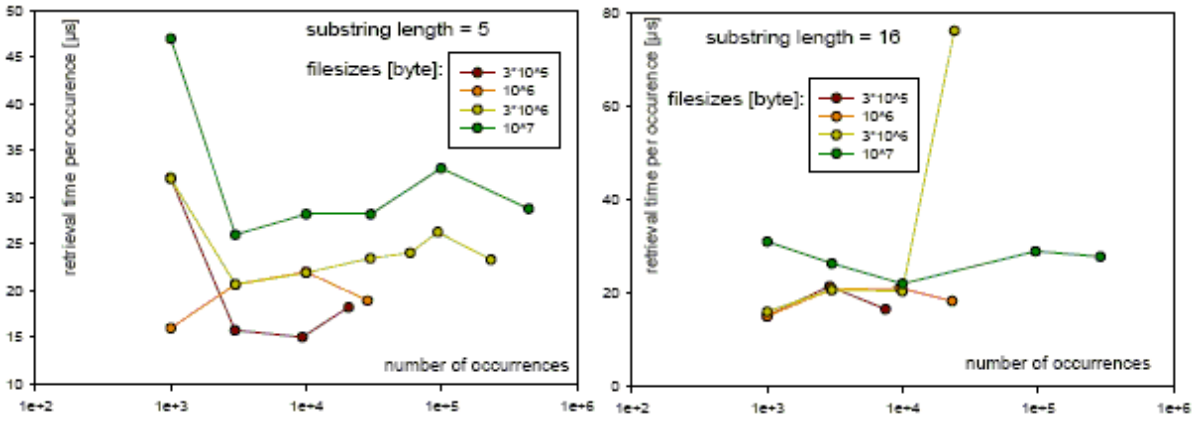


**Figure 3:** Substring search times with fixed substring lengths

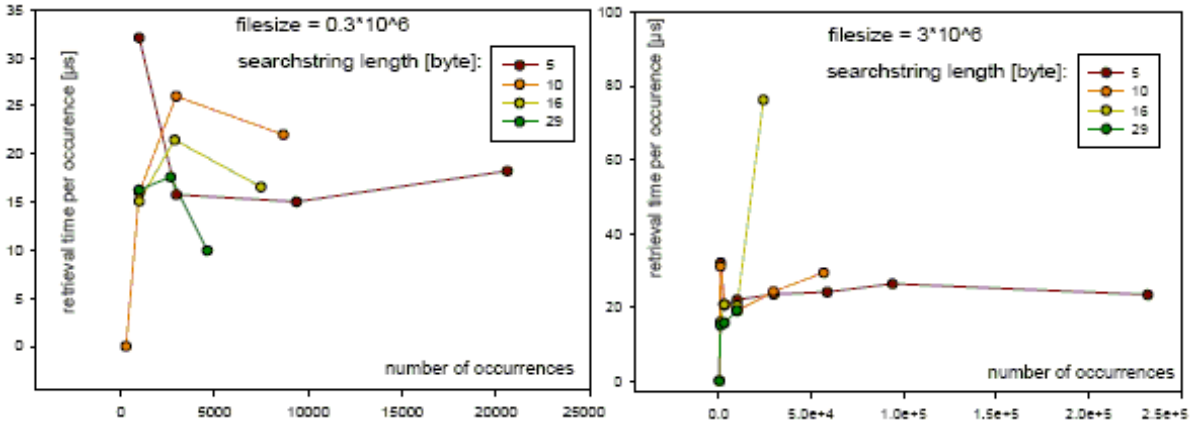


**Figure 4:** Substring search times with fixed file

### 4.3. SUMMARY

Two conclusions can be drawn from the current data [3]:

- There is a capacity increase with increasing size of the structure already assimilated, while maintaining full transparency and accessibility.
- The retrieval times for substring searches depend neither on the size of the substring to be searched, nor on the size of the string to be searched in.

## 5. CONCLUSION

Current technological advances leading to the exponentially increasing amount of biological data. New questions arise on the basis of these data. Area of using Pile is located just where the present computational methods aren‘t enough. Bioinformatics is this field, because the traditional methods of data representation don‘t correspond with the nature of biological conditions.

Available resources and tests show that Pile can be used for working with biological data, and its use can be very beneficial. This approach could be single, universal, homogeneous, flexible and interesting.

It means to overcome some fundamental principles, which also brings problems, because this approach is different. Pile is still only in the development phase and it is necessary to verify the theoretical aspects in practical applications and working on its further develop.

## REFERENCES

[1] Tomeš, M., The Pile system, Scientific papers of the University of Pardubice – Series D Faculty of Economic and Administration 12, 200--208, (2007)

[2] Acimovic, Y., Representing natural data – a Look at the Pile system, 2003, www.pilesys.com/Pile%20Bioninformatics%20ReportYA.pdf, (05.11.2008)

[3] Reuter, D., Processing Data by Assimilating Pure Relations, 2006, http://www.pilesys.com/new/Documents/Pile%20Benchmark.pdf, (05.11.2008)

[4] Acimovic, Y., The Pile system. October 2004, http://www.bioscienceworld.ca/ThePILEsystem, (12.01.2009)