Ecole Doctorale STIC

Sciences & Technologies de l'Information et de la Communication

ED STIC - Proposition de Sujets de Thèse

pour la campagne d'Allocation de thèses 2017

Axe Sophi@Stic :	aucun
Titre du sujet :	DNA CODING FOR LONG-TERM DIGITAL DATA STORAGE
Mention de thèse :	ATSI
HDR Directeur de thèse inscrit à l'ED STIC :	Marc Antonini

Co-encadrant de thèse éventuel :

Nom :	BARBRY
Prénom :	Pascal
Email :	barbry@ipmc.cnrs.fr
Téléphone :	04 93 95 77 93

Email de contact pour ce sujet :	am@i3s.unice.fr
Laboratoire d'accueil :	I3S

Description du sujet :

CONTEXT

Storage of digital data is becoming challenging for the humanity due to the relatively short life span of storage devices. Indeed, the durability of digital equipment, either hard drives, flash drives, floppies or CD-ROMs, compare very badly with the one of clay, vinyl or paper. Servers, for instance, have to be replaced about every five years. Leave a server farm alone too long and its stored data will degrade and become inaccessible more rapidly than any of its analog predecessors. Several projects, for instance at the University of Southampton or at Hitachi, are currently considering new forms of very long term digital storage, using molding silica glass, which estimated storage length time in the range of the 100 millions year. However, these projects are currently stymied by an important problem related to space: both developed at most a storage capacity that

does not exceed 40MBytes per inch, i.e. a very low value compared to the one Terabyte per square inch capacity reached by any standard hard disk.
An interesting alternative approach may stem from the use of DNA, the support of heredity in living organisms. This comes from recent biotechnological developments that allow easy and affordable DNA writing (synthesis) and DNA reading (sequencing). This implies that DNA can appear as an attractive support for long-term data storage, representing a relevant alternative for any future archiving. A key first property of DNA is its longevity, when stored in appropriate conditions, i.e. in an oxygen and water-free environment. This is well illustrated by the capacity to analyze and sequence ancient DNA, even when stored in suboptimal conditions. For instance, a team reconstructed in 2013 the full mitochondrial genome of a bear aged more than 300,000 years, proving that authentic ancient DNA can be preserved for hundreds of thousand years outside of permafrost (Dabney et al.).

A second interesting property of DNA storage is its high-density encoding capacities, well illustrated by the fact that a full human genome, corresponding to a succession of 3 billion characters written with an alphabet composed of four distinct nucleotides, is stored in a cell nucleus, i.e. a volume of 3.10-8mm^3. As pointed out by Church et al, DNA has thus very high-density encoding capacities and could potentially represent an ad hoc storage device for up to petabytes of data in a limited volume: 5.10^15 bits/mm^3 for DNA, to be compared to 3.10^9 bits/mm^3 for Hard Disk, 1.10^7 bits/mm^3 for DVD and 4.10^5 bits/mm^3 for compact disks.

OBJECTIVES OF THE PhD

The present Ph.D. project considers the study of a very long-term storage system where large mass of digital information may be stored as DNA. We envision that this approach could provide a credible solution for long-term storage of critical information, and therefore offer an appropriate solution for the creation of high-value strategic repositories and owns an enormous economical potential. DNA Sequencing can be thought of as the “reading” of a linear text composed by a succession of 4 possible nucleotides: adenine (A), thymine (T), cytosine (C), guanine (G). This is accomplished through a sequential reading of each base in the DNA molecule. Skyrocketing development of Next Generation Sequencing (NGS) during the last 10 years now enables to sequence billions of DNA molecules in just a few hours/days. Recent progresses are well illustrated by the announcement in 2014 by the company Illumina of a full human genome sequencing for $1000, followed by a new announcement in 2017 of a further decrease below $100 before 2020.

Current state of the art of coding is based on binary coding because it is required by the technology and the way nowadays computers work. In computer science, coding a source data requires the association of a binary code word composed by a set of “0” and “1”, to each different value the source can take. Basically, code words try to reduce the redundancy and represent the source with fewer bits that carry more information. A compact representation is obtained when the amount of bits that represents the source

corresponds to its entropy, which gives the minimum number of bits necessary to describe completely the source without error.
Thanks to its intrinsic properties, DNA coding is a strong opportunity to make a breakthrough for long-term storage of multimedia data, pending some adaptations for using it in
digital coding. Indeed, DNA coding is a quaternary coding processing that must be adapted to digital data (sequence of bits) if one wants to use DNA as a storage support.
Then, as DNA structures are based on an assembly of 4 “bases” {A, C, G, T} and if one wants to exploit the DNA mechanism as a compact source code, the most important
issue to solve is to extend the concept of binary code word to quaternary code word.

In this Ph.D., the student aim will be to develop an encoding process that allows generating a quaternary stream adapted to DNA sequencing. The proposed solution should
be:
- Efficient in term of compression, i.e., generating quaternary stream with compact length.
- Decodable and robust to the synthesis and sequencing errors. Because the processes of DNA synthesis and sequencing are highly prone to errors, inclusion of robust
encoding and decoding processes is mandatory. The Ph.D. will focus on several possible solutions to robustify the DNA code.
- Adapted to DNA sequencing constraints (avoid homopolymers, limit the number of nucleotides by read, create plausible pairs of nucleotide etc.).

Compression efficiency and noise robustness are two issues of DNA coding/sequencing that we propose to study in that Ph.D. To this end, we aim at developing an efficient
compression strategy adapted to the nature of the signal to be encoded, and to define a cost-effective protocol to translate any digital document into a DNA library.

URL : http://www.i3s.unice.fr/~am/sujets/SujetTheseDNA-EDStic.pdf

English version: