AU2EU: Privacy-Preserving Matching of DNA Sequences

Tanya Ignatenko, Milan Petković

AU2EU: Privacy-Preserving Matching of DNA Sequences [⋆]

Tanya Ignatenko[1], Milan Petković[2,3]

[1] Electrical Engineering Department,
Eindhoven University of Technology, The Netherlands
t.ignatenko@tue.nl
[2] Mathematics and Computer Science Department,
Eindhoven University of Technology, The Netherlands
[3] Philips Research Eindhoven, The Netherlands
milan.petkovic@philips.com

Abstract. Advances in DNA sequencing create new opportunities for the use of DNA data in healthcare for diagnostic and treatment purposes, but also in many other health and well-being services. This brings new challenges with regard to the protection and use of this sensitive data. Thus, special technical means of protection should safeguard critical DNA data and create trust for patients and consumers of lifestyle services. In particular an interesting research challenge is to design secure operations on DNA sequences in the encrypted domain that allow a person to engage into a DNA-based service and obtain required (medical) answers without revealing his/her DNA. We focus in this paper on this topic and present a solution to a particular problem of privacy-preserving matching of DNA sequences which can be used in clinical trials or other DNA services.

1 Introduction

To be competitive and efficient, multiple independent organizations often have to form virtual collaborations to work together on critical applications or to share sensitive data. In order to facilitate such collaborations, a widely deployed network infrastructure can be used to allow access to either cloud-based environments or directly grant access through the involved parties to each other's systems and resources. Given the fact that such applications have to deal with sensitive data, organizations feel reluctant to move their resources to the cloud or safely rely on the (authentication) claims coming from the members of the collaboration. As a result trust becomes an important component of such collaborations. Cross-organizational and jurisdictional nature of these collaborations makes it hard to relate different attributes and policies of different collaborating parties and thus build mutual trust.

[⋆] This work has been partially funded by the EC via grant agreement no. 611659 for the AU2EU project.

The problem of establishing trust lies in the lack of authentication and authorization infrastructures supporting high level of assurance, privacy as well as cross-domain and jurisdictional collaborations. Thus, there is a clear need for an adequate infrastructure for authentication and authorization in establishing trust. To address this need and general requirements of distributed eAuthentication and eAuthorization infrastructures for trusted secure information sharing, the EU FP7 project called AUthentication and AUthorization for Entrusted Unions (AU2EU) was initiated. The project is a joint collaboration between seven EU partners and five research institutes and universities in Australia. The main objectives of this project are to: (a) deploy a composable architecture that builds on the best existing practices and novel emerging techniques to design the eAuthentication and eAuthorization infrastructure for cross-organizational and jurisdictional collaborations; (b) extend the joint eAuthentication and eAuthorization framework with four novel functionalities: (i) assurance of claims to increase trust by introducing the ability to assess reliability of claims; (ii) trust indicators to assess trustworthiness of the involved devices, platforms and services; (iii) cryptographic policy enforcement to ensure data confidentiality in cloud-based and offline scenarios inherent to distributed systems; and (iv) mechanism to perform operations under encryption to enable processing data in a privacy-preserving way; (c) implement the resulting joint eAuthentication and eAuthorization framework and deploy it in two real-life pilots in Australia and Europe; and (d) evaluate its security, maturity, scalability, and usability.

The project is driven by four use cases, namely bio-security, eHealth, and two healthcare use cases related to picture archiving and communication systems (PACS) and DNA data management. The joint eAuthentication and eAuthorization infrastructure is designed by combining the XACML-based authorization framework [3], ABC4Trust [2] authentication architecture and TDL [1] authentication framework. In our framework we deploy novel mechanisms for semantic mapping to translate policies and attributes to the required authentication claims that can be verified in our authentication framework. To guarantee strong authentication and privacy at the same time, idemix technology [4] for attribute-based authentication is used as a building block of our authentication architecture. To support privacy as well as ease of use for collaborating parties, the idemix is integrated as a cloud-based service in the AU2EU architecture. The platform will be deployed and evaluated in the bio-security and eHealth pilots.

In the rest of this paper, we focus on the DNA use case that fall under one of the AU2EU research directions, which is operations under encryption. When sensitive data, e.g. patient DNA or bio-security incident data, need to be accessed and processed by various parties in the distributed collaborative systems, restricting access to only partial data is a difficult task, since processing and extracting partial information from the data often requires access to the whole dataset. Policy enforcement mechanisms alone cannot guarantee this fine-grained level of access control. It would be ideal if we could perform operations on the encrypted data that are equivalent to the operations one need to perform on the corresponding non-encrypted data, without the need to decrypt them.

Processing in the encrypted domain that builds on homomorphic encryption techniques [5], secure multi-party computation [6], and code-based security [7] suggest possible solutions. It would allow a party to engage into applications dealing with sensitive data (e.g. medical trial or disease risk profiling that use DNA data) and obtain required answers without revealing particular data or without 'seeing' user's data in plain text.

In this paper we concentrate on the problem of privacy-preserving similarity search in DNA databases for clinical trials and medical research. Clinical trials and genome-wide association studies are typical tools to evaluate effectiveness of certain treatments and drugs, and to determine dependencies between DNA patterns and diseases. In clinical trials, the eligibility criteria for inclusion in a trial might include patients with DNA sequences that have similar phenotype (e.g. race) and functionality (e.g. a gene is on or off). In genome-wide association studies, to conduct tests, one need to select DNA sequences that can be formed into cases (e.g. sequences that contain a mutation) and controls (sequences that do not contain a mutation). Therefore, to find eligible patients for a clinical trial or data for research purposes, various parties like pharmacies that conduct a trial, and clinical researchers have to be able to look up patient's primary medical records and research repositories containing DNA information and check DNA sequences against inclusion criteria. However, accessing DNA information in such databases poses privacy and security concerns. Remarkably, DNA sequences are self-identifying sensitive data. They are a unique identifier of human beings; moreover these sequences contain information used for disease risk profiling, ancestry determination and, potentially, other more personal physiological aspects. It is important to realize that since DNA data are identifiers by themselves, DNA sequences, unlike other medical information, cannot be anonymized. Thus realization of clinical trials and research experiments that use genetic information as a subject selection criterion requires a proper DNA management infrastructure in place. In the next section we present our solution to a particular problem of privacy-preserving indexing of DNA sequences to support similarity search.

2 Privacy-Preserving DNA Indices

Consider the problem when a third party (e.g. a pharmacy) has to query a (distributed) DNA database in order to find patients (e.g. volunteers for a clinical trial) whose DNA sequences are similar to a query (example) DNA sequence. To guarantee privacy of DNA information stored in the database, we only store privacy-preserving DNA indices that can be used for similarity measurements. We propose to use DNA sequences processed into context trees as index-information that can be used to facilitate privacy-preserving matching and similarity search in DNA databases. The context trees are built by estimating the underlying model of (a set of similar, in a certain sense) DNA sequence(s) using the universal compression technique called context-tree weighting (CTW) [8]. As a retrieval criterion the mutual information, that characterizes the inherent dependence of two variables, see e.g. [19], between a query DNA sequence and database sequences is used. We compute this mutual information as a difference between the codeword length of a query DNA sequence computed using CTW and the codeword length of this DNA sequence computed given the stored DNA-indices. Privacy of DNA information is achieved by only storing the context trees that represent the DNA source generating the sequences, as the context trees along are insufficient to reconstruct the underlying DNA sequences.

2.1 DNA Preliminaries

Genome is entire organism hereditary information containing the complete set of instructions for constructing an organism. The human genome consists of tightly coiled threads of deoxyribonucleic acid (DNA) which basic building blocks are four nucleobases or bases that are adenine (A), thymine (T), cytosine (C), and guanine (G). The particular order of the nucleobases is called the DNA sequence which is measured in base pairs (bp). The human genome contains roughly 3 billion bp. DNA sequences contain instructions for manufacturing all proteins, in this way to form proteins triplets of DNA bases (codons) are interpreted as amino acids, and amino acids in their turn are added to a growing chain that forms protein. Thus DNA sequences are not random and have logical sequential organization.

Individual genomes vary in about 1 in 1000 bp. Remarkably these small variations account for significant phenotype differences including disease susceptibility, medication response etc. Typically variations are accumulated over time from mutations, structural polymorphisms, chromosome recombination. On a structural level, these variations include single-nucleotide polymorphisms (SNP), that is substitutions variation in a DNA at a single nucleotide position; structural polymorphism, that is a large scale structural changes characterized by indels - insertion or deletion of short nucleotide sequences.

2.2 Context-Tree Weighting (CTW) Method

The context-tree weighting (CTW) method [8] is a universal source coding method that finds a good coding distribution for an observed (DNA) source sequence in a sequential way. This coding distribution corresponds to all tree-models whose depth does not exceed D. The distribution can be used to compress an observed sequence using arithmetic coding techniques. The CTW method also approaches entropy for ergodic stationary sources. The CTW method can also be used as a two-pass method [18]: in the first step it is used to determine the statistical model matching an observed (DNA) sequence, and in the second step this model is encoded and the observed (DNA) sequence is encoded (compressed) given the model.

The CTW method uses a concept of context trees. A context tree, see Fig.1, consists of nodes that correspond to contexts s up to a certain depth D. Each node s in the context tree is associated with the subsequence of source symbols that occurred in the observed sequence after context s. Moreover, to each node s

there corresponds a parameter θ_s that give s the probability of the next symbol being one (in the binary case) in the sequences when context s was observed.

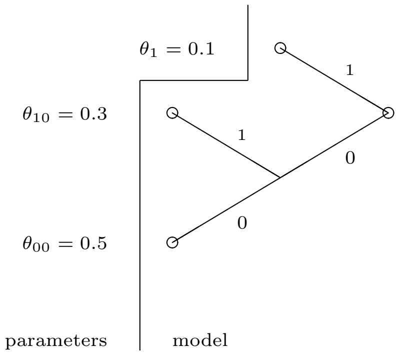


Fig. 1. An example of a binary context-tree of depth 2 with parameters, from [8].

The DNA sequence structure is such that it codes for amino acids and subsequently for proteins in a sequential way. This makes the CTW particularly suitable for this type of data. For example, it was shown in [9] that CTW performs well for DNA compression. Let x^T denote an observed DNA sequence. Then CTW can be used to estimate $P(x^T)$, where x^T corresponds to a DNA vector with values from alphabet $A = \{1, 2, 3, 4\}$[4]. Denote with x_t a symbol at position t in the observed sequence x^T. A statistical model for the DNA sequence is estimated by building the context tree and estimating the distribution $P(x^T)$ using the CTW algorithm as $P(x_t|\{x_{t-b}, b \in B\})$, where B is a set of well-chosen integers. The "context" $\{x_{t-b}, b \in B\}$ consists of a set of values from alphabet A obtained from $|B|$ different locations of x^T. Typically, B is defined as a set of values preceding x_t. All possible contexts (that actually occurred in the observed DNA sequence) together with probability distribution $P(x_t|\{x_{t-b}, b \in B\})$ constitute the context-tree (model) and the parameters, respectively. Thus, for a DNA sequence the context-tree model constitutes an ensemble of short subsequences that occurred there.

2.3 Similarity Measure

Consider an observed DNA sequence x^T. Suppose $\{S, \Theta_S\}$ are a model (contexts) and parameter set (conditional probabilities) describing some tree source of depth not larger than D. Then if we use $\{S, \Theta_S\}$ to compress this sequence, the length of the compressed sequence will be given by

$$L(x^T|x^1_{-D}, \{S, \Theta_S\}) = -\sum_{t=1}^{T} \log_2 P(x_t|x^{t-1}_{-D}, \{S, \Theta_S\}) = -\sum_{t=1}^{T} \log_2 \theta^{x_t}_{\sigma_{\{x^{t-1}_{-D}\}}}, \tag{1}$$

[4] Since DNA alphabet is $\{A, T, G, C\}$, we can replace it with $\{1, 2, 3, 4\}$ without loss of generality.

where $\sigma_{\{x_{-D}^{t-1}\}}$ is a mapping of x_{-D}^{t-1} to a context from S and

$$P(x_t|x_{-D}^{t-1}, \{S, \varTheta_S\}) = \theta_{\sigma_{\{x_{-D}^{t-1}\}}}^{x_t} \in \varTheta$$

is the probability of symbol x_t to occur after subsequence $\sigma_{\{x_{-D}^{t-1}\}}$ was observed in x^T. When $\{S, \varTheta_S\}$ describes the actual source that produced x^T then $L(x^T|x_{-D}^1, \{S, \varTheta_S\})$ corresponds to the ideal codeword length. However, if $\{S, \varTheta_S\}$ describes some other source then $L(x^T|x_{-D}^1, \{S, \varTheta_S\})$ will be larger than the ideal codeword length as the used model does not help to describe the observed sequence, this also relates to MDL principle see e.g. [15]. Note that when CTW is used to estimate model and parameters of an observed (DNA) sequence, then the resulting codeword length will have the smallest distance (redundancy) from the ideal codeword length.

Now suppose y^N and x^T are two observed DNA sequence not necessarily of the same length. Let $\{S_x, \varTheta_{S_x}\}$ be the model and parameter set for x^T, estimated using the CTW method and $L_{ctw}(y^N)$ be the codeword length for y^N estimated using the CTW method. Then if we consider the difference

$$\begin{aligned}
&\frac{1}{N}L_{ctw}(y^N) - \frac{1}{N}L(y^N|\{S_x, \varTheta_{S_x}\}) \\
&\quad = -\frac{1}{N}\sum_{t=1}^{N}\log_2 P_{ctw}(y_t|y_{-D}^{t-1}) + \frac{1}{N}\sum_{t=1}^{N}\log_2 P(y_t|y_{-D}^{t-1}, \{S_x, \varTheta_{S_x}\}) \\
&\quad = -\frac{1}{N}\sum_{t=1}^{N}\log_2 \frac{P_{ctw}(y_t|y_{-D}^{t-1})}{P(y_t|y_{-D}^{t-1}, \{S_x, \varTheta_{S_x}\})} = -\frac{1}{N}\sum_{t=1}^{N}\log_2 \frac{P_{ctw}(y_t|y_{-D}^{t-1})}{\theta_{S_x, \sigma_{\{y_{-D}^{t-1}\}}^{y_t}}}, \qquad (2)
\end{aligned}$$

we see that this difference tells how much we can gain if we use the distribution of x^T instead of y^N in order to describe (compress) y^N. If the gain is high then $\{S_x, \varTheta_{S_x}\}$ corresponds to the source that fits well y^N, and it is more likely that both y^N and x^T are generated by the same source, thus they belong to the same sequence class. If the gain is low, then codeword length for y^N estimated using $\{S_x, \varTheta_{S_x}\}$ has very high redundancy and thus $\{S_x, \varTheta_{S_x}\}$ does not help to describe y^N, which means that it corresponds to some other source generating other types of (DNA) sequences. Hence we can say that y^N and x^T are generated by different sources and they are not similar. In general, the higher the gain the better the model and parameter set describe sequence y^N. Thus it is the more likely, that the source with $\{S_x, \varTheta_{S_x}\}$ generated y^N.

The codeword length per source symbol estimated using the CTW method gives (a good) estimate of the entropy of the (DNA) source sequence. Hence the similarity measure given above can be seen as an estimate of the mutual information between a DNA sequence Y^N and a DNA source that produced some DNA sequence X^T. Note that mutual information is non-negative, while our estimate can take up negative values. This underestimate partially comes from the fact that query sequence can have deletions, insertions and substitutions that are not part of the source model used for the codeword length estimates.

Now consider the mutual information between a database DNA sequence X^N and the query DNA sequence Y^N

$$\begin{aligned} I(Y^N; X^T) &= I(Y^N; X^T, \{S_x, \mathit{\Theta}_{S_x}\}) \\ &= I(Y^N; \{S_x, \mathit{\Theta}_{S_x}\}) + I(Y^N; X^T | \{S_x, \mathit{\Theta}_{S_x}\}), \end{aligned} \tag{3}$$

where in the first step we use the fact that $\{S_x, \mathit{\Theta}_{S_x}\}$ is a function of X^T and the data-processing inequality, see e.g. [19]. Note that when $\{S_x, \mathit{\Theta}_{S_x}\}$ matches the source that generated Y^N, the second term in the last step becomes negligible, on the other hand if Y^N and X^N are produced by different sources this term is also small. We see that mutual information between X^N and Y^N sequences is equivalent to the mutual information $I(Y^N; \{S_x, \mathit{\Theta}_{S_x}\})$. Thus in order to find the closest sequence we may concentrate on finding the estimated model and parameters that maximize $I(Y^N; \{S_x, \mathit{\Theta}_{S_x}\})$.

2.4 Proposed System

Set-up:
Create a database of privacy-preserving DNA-indices for a (sets of) DNA sequence(s) $x_i^{T_i}, i = 1, 2, \ldots, n$. In order to do it, estimate the models and parameters for each (sets of) DNA sequences $x_i^{T_i}, i = 1, 2, \ldots, n$ applying the CTW method. Store $\{S_{x_i}, \mathit{\Theta}_{S_{x_i}}\}$ in the database together with some other relevant information.
Retrieval:
Given the query (example) DNA sequence y^N, perform the following steps:

1. Apply CTW and estimate the codeword length per source symbol

$$\frac{1}{N} L_{ctw}(y^N), \text{ for } y^N; \tag{4}$$

2. For each DNA record $i, i = 1, 2, \ldots, n$ in the database, compute the estimate of the codeword length for y^N given $\{S_{x_i}, \mathit{\Theta}_{S_{x_i}}\}$, by mapping subsequences in y^N to the contexts from S_{x_i} and using the corresponding parameters as

$$\frac{1}{N} L(y^N | \{S_{x_i}, \mathit{\Theta}_{S_{x_i}}\}) = - \sum_{t=1}^{N} \log_2 \theta_{S_{x_i}, \sigma^{y_t}_{y^{t-1}_{-D}}}, \tag{5}$$

note that if there is no context in S_{x_i} for some subsequence from y^N, then the corresponding parameter equals 1/2. Observe that this parameter 1/2 will also contribute to dissimilarity of the close DNA sequences when deletion or insertion occurred.

3. Find the record i that maximizes

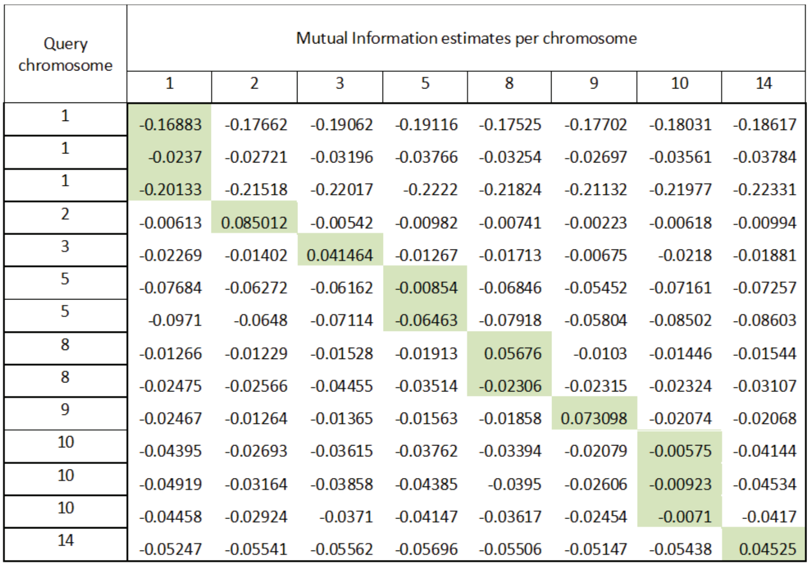
$$\frac{1}{N} L_{ctw}(y^N) - \frac{1}{N} L(y^N | \{S_{x_i}, \mathit{\Theta}_{S_{x_i}}\}) \tag{6}$$

and return the relevant information to the querying party.

2.5 Security Discussion

Observe that in the DNA database in order to perform privacy-preserving similarity search one only need to store the model and the parameter set $\{S_{x_i}, \Theta_{S_{x_i}}\}$ corresponding to a (set of) DNA sequence(s). Note that the model consists of short subsequences that occurred in the DNA sequences, but contains no information on temporal arrangement of the subsequences. Moreover, due to DNA variable length, also probabilistic information contained in the parameters is insufficient to characterize the number of the subsequences. Note also that an average typical length of DNA sequence is around 3.2×10^9bp. Thus our model and parameter set can be seen as a hash of DNA sequences that allows for prohibitively large number of sequences being produced based on it.

2.6 Toy Example



Query chromosome	Mutual Information estimates per chromosome							
	1	2	3	5	8	9	10	14
1	-0.16883	-0.17662	-0.19062	-0.19116	-0.17525	-0.17702	-0.18031	-0.18617
1	-0.0237	-0.02721	-0.03196	-0.03766	-0.03254	-0.02697	-0.03561	-0.03784
1	-0.20133	-0.21518	-0.22017	-0.2222	-0.21824	-0.21132	-0.21977	-0.22331
2	-0.00613	0.085012	-0.00542	-0.00982	-0.00741	-0.00223	-0.00618	-0.00994
3	-0.02269	-0.01402	0.041464	-0.01267	-0.01713	-0.00675	-0.0218	-0.01881
5	-0.07684	-0.06272	-0.06162	-0.00854	-0.06846	-0.05452	-0.07161	-0.07257
5	-0.0971	-0.0648	-0.07114	-0.06463	-0.07918	-0.05804	-0.08502	-0.08603
8	-0.01266	-0.01229	-0.01528	-0.01913	0.05676	-0.0103	-0.01446	-0.01544
8	-0.02475	-0.02566	-0.04455	-0.03514	-0.02306	-0.02315	-0.02324	-0.03107
9	-0.02467	-0.01264	-0.01365	-0.01563	-0.01858	0.073098	-0.02074	-0.02068
10	-0.04395	-0.02693	-0.03615	-0.03762	-0.03394	-0.02079	-0.00575	-0.04144
10	-0.04919	-0.03164	-0.03858	-0.04385	-0.0395	-0.02606	-0.00923	-0.04534
10	-0.04458	-0.02924	-0.0371	-0.04147	-0.03617	-0.02454	-0.0071	-0.0417
14	-0.05247	-0.05541	-0.05562	-0.05696	-0.05506	-0.05147	-0.05438	0.04525

Fig. 2. Estimates for mutual information from the toy example. A shaded cell corresponds to the maximum mutual information

Here we consider a toy example where we use 14 DNA sequences from GenBank. Suppose we need to arrange the database per chromosome. Then we create the corresponding privacy-preserving indices using CTW with depth 9 (this corresponds to three codons) by estimating the models and parameter sets for each chromosome, i.e. for chromosome 1, 2, 3, 5, 8, 9, 10, 14 in our example. These models and parameter sets are stored in the database.

Next a researcher presents a piece of a DNA sequence and he would like to find from which chromosome it comes from. Using the system described above he can calculate the estimates of the mutual information between the available piece of the DNA sequence and the models and parameters corresponding to different chromosomes, and then find the chromosome that maximizes the mutual information. Fig.2 shows the results of such estimates for a number of query sequences. From this table we observe that the proposed method can correctly detect which chromosome the query piece of DNA comes from.

3 Related work

Work in the direction on privacy-preserving operations on DNA data focuses on privacy-preserving calculation of edit (Levenshtein) and set distances. E.g. in [10] oblivious automata for privacy-preserving approximate DNA matching and searching is proposed. This approach is using Levenshtein distance as a similarity metric for DNA sequences. In [11] edit distance between two DNA sequences is derived using homomorphic encryption. In [12] and [17] homomorphic encryption and secure two-party computations ares used to match short tandem repeats that are used for human identification and for parental tests. Finally, in [13] Privacy-Enhanced Invertible Bloom Filter (PEIBF) is proposed for set distance computations based on compressed DNA sequences, where DNA sequence compression is defined as sets of differences from the DNA reference string.

The approaches in [10] and [13] are effective for human authentication and identification, a well as verification if a certain pattern is a part of a given DNA. Methods based on homomorphic encryption like [11] are prohibitively expensive to be used in large databases and can be effectively used for authentication. The approach presented in [12] is applicable to human authentication and identification, forensic investigations and parental tests. However, the approaches mentioned above are not sufficient if one has to determine whether DNA sequences have a similar functionality, since e.g. it was shown that chimpanzee and human genomes are 96% similar [16], while the corresponding edit distance between two genomes is very large. Therefore, to compare DNA sequences more complex similarity metrics than edit or set distance, like divergence [16] and mutual information [14] are needed, as these metrics also takes into account temporal structure of DNA sequences. Note that work in [16] and [14] does not focus on privacy, while our approach aims at privacy-preserving similarity search based on mutual information. To the best of our knowledge this is the first work in this direction.

4 Conclusions

In this paper we have presented a particular solution for privacy-preserving search and matching in DNA databases. Our approach is based on the universal source coding technique, CTW [8]. Further investigations of the proposed solution, as well as design of a wide range of operations on DNA sequences still remain as the future work in the AU2EU project.

References

1. Ronny Bjones, “Architecture serving complex Identity Infrastructures”, Trust in Digital Life, p.21, 2012
2. J. Camenisch, I. Krontiris, A. Lehmann, G. Neven, C. Paquin, K. Rannenberg, H. Zwingelberg, “Architecture for Attribute-based Credential Technologies”, ABC4Trust, Available at https://abc4trust.eu/index.php/pub/results/107-d21architecturev1
3. Erik Rissanen, “eXtensible Access Control Markup Language (XACML) Version 3.0”, OASIS standard, 2010, http://docs.oasis-open.org/xacml/3.0/xacml-3.0-core-spec-cs-01-en.pdf
4. IBM Research Zurich, “Specification of the Identity Mixer Cryptographic Library Version 2.3.1”, 2010
5. C. Fontaine, F. Galand, “A Survey of Homomorphic Encryption for Nonspecialists,” EURASIP Journal on Information Security, 2007.
6. I. Damgård, V. Pastro, N. Smart, S. Zakarias, “Multiparty Computation from Somewhat Homomorphic Encryption,” Advances in Cryptology - CRYPTO 2012, LNCS 7417, pp. 643–662, 2012.
7. M. Finiasz, N. Sendrier, “Security Bounds for the Design of Code-Based Cryptosystems,” Adv. in Cryptology - ASIACRYPT 2009, LNCS 5912, pp. 88–105, 2009.
8. F.M.J. Willems, Y.M. Shtarkov, T.J. Tjalkens, “The Context-Tree Weighting Method: Basic Properties,” IEEE Trans. on Information Theory, vol.41, no.3, pp.653–664, 1995.
9. T. Matsumoto, K. Sadakane, and H. Imai, “Biological Sequence Compression Algorithms,” Genome Informatics Workshop, vol.11, pp. 43-52, 2000.
10. J. R. Troncoso-Pastoriza, S. Katzenbeisser, and M. Celik, “Privacy Preserving Error Resilient DNA Searching Through Oblivious Automata,” In Proceedings of the 14th ACM Conference on Computer and Communications Security (CCS’07), ACM, pp.519–528, New York, NY, USA, 2007.
11. Somesh Jha , Louis Kruger , Vitaly Shmatikov, “Towards Practical Privacy for Genomic Computation,” The 2008 IEEE Symposium on Security and Privacy, pp.216-230, May 18-21, 2008.
12. F. Bruekers, S. Katzenbeisser, K. Kursawe, and P. Tuyls, “Privacy-Preserving Matching of DNA Profiles,” IACR Cryptology ePrint Archive, 2008.
13. D. Eppstein, M. T. Goodrich, P. Baldi, “Privacy-Enhanced Methods for Comparing Compressed DNA Sequences,” CoRR abs/1107.3593, 2011.
14. C. Adami, “Information Theory in Molecular Biology,” Physics of Life Reviews, vol.1, no.1, pp.3–22, April 2004.
15. R. Cilibrasi and P. Vitanyi, “Clustering by Compression,” IEEE Trans. on Information Theory, vol.51, no.4, pp.1523–1545, 2005.
16. Chimpanzee Sequencing and Analysis Consortium, “Initial Sequence Of The Chimpanzee Genome And Comparison With The Human Genome,” Nature 437 (7055), pp.69–87.
17. E. De Cristofaro, S. Faber, P. Gasti, G. Tsudik, “Genodroid: Are Privacy-Preserving Genomic Tests Ready For Prime Time?”, WPES’2012, pp. 97–108, 2012.
18. F.M.J. Willems, A. Nowbahkt-Irani, and P.A.J. Volf, “Maximum a-Posteriori Tree Models,” ITG 2002, Berlin, Germany, February, 2002.
19. T. M. Cover and J. A. Thomas,“Elements of Information Theory,” 2nd Ed., New York: John Wiley and Sons Inc., 2006.