ISSN 1870-4069

An Algorithm to Classify DNA Sequences of Hepatitis C Virus Based on Localized Conserved Regions and Heuristic Search

Sarahi Zúñiga-Herrera[1], Ivan Olmos-Pineda[1], Javier Garcés-Eisele[2], Mario Rossainz-López[1]

[1] Benemérita Universidad Autónoma de Puebla, Facultad de ciencias de la Computación, Puebla, Mexico
{sarahi.zuhe, ivanoprkl}@gmail.com, rossainz@cs.buap.mx
[2] Laboratorios Clínicos de Puebla, Puebla, Mexico
javier.garces@gmx.de

Abstract. In this work, the problem of classification in DNA sequences (STP Sequence-Typing Problem) is addressed. The main tool for molecular diagnosis is the polymerase chain reaction (PCR). However, the design of a PCR test is expensive (in time and economic resources) if the number of types of sequences to be classified is large and highly variable. The optimal design of PCR primers is essential to maximize the specificity and sensitivity of the test. This paper presents an algorithm to solve the problem of locating conserved sequences regions that help to discriminate between different types of hepatitis C virus using Shannon entropy concepts, information gain, a tree-based classifier and an evaluation function that considers specific parameters of PCR tests, which allow to identify patterns in conserved areas of DNA sequences.

Keywords: Shannon Entropy, Information Gain, DNA Sequences, Classification.

1 Introduction

DNA is made up of four molecules called nucleotides or nitrogenous bases. It contains all the genetic information of living organisms [1]. If we know a DNA sequence or part of it, we can determine to which organism it belongs. Thus, different types of sequence tests can be used in clinical analysis laboratories to identify: infectious agents present in a blood sample taken from a patient, for diagnostic tasks, and others applications.

One of the most frequent problems in diagnosis is to classify a given sequence and determine the class or subclass, which is known as sequence-typing problem (STP) [2]. There are different instances for the STP. Currently, the problem of classifying DNA sequences with a high variability rate is of great interest for the development of diagnostic tests. In this paper, the case for the hepatitis C virus is addressed. The hepatitis C virus (HCV), since its discovery in 1989, has been recognized as a serious

health worldwide problem [3]. The World Health Organization estimated that 71 million people had chronic HCV infections in 2015 [4]. At present, this virus represents the most frequent cause of chronic liver disease, cirrhosis and liver transplantation. HCV has a high genetic variability, identifying from its discovery to date seven main types associated with different behaviors of the virus in the host and responses to treatment [5]. To determine the type or subtype of HCV it is necessary to use a molecular diagnosis test which allows the doctor to be guided about the treatment for the patient.

One of the main tools in molecular diagnostics is the polymerase chain reaction (PCR) [1]. PCR revolutionized the field of molecular diagnostics, to the point that it currently represents the fastest growing segment in the clinical laboratories of the world [5].

The PCR process was originally developed to amplify short segments of a longer DNA molecule [6]. A typical amplification reaction includes target DNA, a thermostable DNA polymerase, two oligonucleotide primers, deoxynucleotide triphosphates (dNTPs), reaction buffer and magnesium. Once assembled, the reaction is placed in a thermal cycler, an instrument that subjects the reaction to a series of different temperatures for set amounts of time. This series of temperature and time adjustments is referred to as one cycle of amplification. Each PCR cycle theoretically doubles the amount of targeted sequence (amplicon) in the reaction. Ten cycles theoretically multiply the amplicon by a factor of about one thousand; 20 cycles, by a factor of more than a million in a matter of hours. Each cycle of PCR includes steps for template denaturation, primer annealing and primer extension [7]. PCR is characterized for being a technique with high sensitivity, reproducibility and efficiency, which generates reliable results in a short time and easy to analyze [8,9]. However, the design of a PCR diagnostic test can be very complex, if the number of sequences to be classified is large and, in addition, they are highly variable. The optimal design of primers is essential to maximize the specificity and efficiency of a PCR [10]. Poor design of the primers can result in small, nil or non-specific quantities of the amplification product. An appropriate design is one of the most important factors for the success of a PCR [8]. In general, the design of primers is summarized in 4 main points.

The first one is to obtain a database with the target genetic sequences, this database can be obtained in banks of international sequences such as GenBank or more selective sources such as ViRP where genetic sequences of viral pathogens are located, including HCV.

The second step is to process the database by aligning it using any of the currently available computational tools such as Jalview, Strap, ClustalX or Clustal Omega, among others [11] to locate the homologous and conserved regions. This is the process by which the sequences are compared by searching for common characters and establishing the correspondence residues between the related sequences. These regions are interesting because if there is a high degree of conservation, the probability of amplification increases.

The third step corresponds to the identification of the oligonucleotides or primers. Select those nucleotides that meet the chemical criteria that guarantee specificity and sensitivity.

The fourth corresponds to the verification and validation of the oligonucleotides proposed by PCR reactions.

The third step is described as the most expensive part of the entire design in terms of time and economic resources, especially for tests where you want to perform a classification of sequences as mentioned is the case for HCV, since two factors should be considered mainly: the first is to select regions that allow us to distinguish between one class and another; the second is that established chemical and thermodynamic conditions that guarantee the amplification of the PCR.

On the second factor, there are currently several computer programs that perform these tasks, such as oligo7 [11]. The level of development of these programs is so high that for identification tests it is enough to enter the sequence to the program and this will yield a series of the best proposals for the primers, leaving the researcher little or nothing to improve and accept the proposal made by the software. However, for the first factor and based on the study carried out in the state of the art, there are no computational tools that allow the selection of those primers that could classify the sequences of interest. Currently researchers are limited to observe the sequences and determine manually which is the right region to solve the classification problem.

2 State of the Art

The problem consists in deciding whether the unclassified DNA sequences belongs to a particular class. There are many highly efficient string-matching algorithms in the current state of the art that could be used to solve the problem of classification like Machine learning algorithms. However, in the clinical diagnosis a sample is extracted from a patient that contains a genetic material of which the sequence of nucleotides is unknown, so the algorithms available for the machine learning cannot be used since they require that the nucleotide sequence be known. Therefore, algorithms must be devised to help solve the classification problem by means of clinical diagnostic methods that currently exist, as has been mentioned one of the most used in PCR.

The researchers who design the primers for a PCR currently do the selection of the regions to solve the classification problem manually, this method is slow and complex, therefore it is necessary to design an algorithm that automates this process. Consequently, it is proposed to use the information gain criteria and look for those regions that after a mathematical analysis turn out to be better candidates to solve the classification problem.

To solve the STP problem on different pathogens, different proposals have been proposed. For example, in 2002 and 2004, two works were developed under the direction of Javier Garcés and his team, in which a solution to the STP problem for the human papillomavirus (HPV) is proposed using alignment tools of sequences, clustering processes, tools of Shannon's information theory such as entropy and information gain and decision theory [12, 13]. His proposal helped to design a molecular diagnostic test by RFLP-PCR (Restriction Fragment Lenght Polymophism coupled to Polymerase Chain Reaction) that is currently used in laboratories for the diagnosis of HPV [14, 15]. On the other hand, Benish WA [16] applied the Shannon Information Theory

to clinical analysis tests calculating the Gain of Information pre- and post-test. Information theory was also successfully applied to the codon classification problem to reveal the order in the genetic code [17]. It has also been used to clarify the interrelationships between structure, function and evolution of a family of genes or gene products [18]. Ebeling and Frommel [19] applied the concept of entropy as the ability to describe the structure of information carriers such as DNA, proteins, text and musical notes. The research of Solis et al. [20] proposes a method to extract the maximum amount of information available from peptide structures in fragments of sequences, finding that the manner in which the structure is represented affects the quantity and quality of structural information that can be extracted from sequences.

Despite the work already done and the tools designed, these tools focus on solving the problem of identifying conserved regions and designing primers regardless of which regions optimize the classification process and without consider hybridization criteria for PCR tests. In the case of the work done by Garcés, the entropy analysis and information gain focuses on the design of an RFLP-PCR. Therefore, it is necessary to do applied research that allows the development of software tools to solve the problem of classification of sequences by PCR diagnostic tests.

3 Formal Description of the Problem

A nucleotide is a fundamental organic chemical compound of nucleic acids (DNA and RNA), constituted by a nitrogenous base, a sugar and a molecule of phosphoric acid [1]. In a DNA sequence there are four nucleotides: adenine (A), cytosine (C), guanine (G) and thymine (T). A DNA sequence or chain represented by G_w is a sequence of letters representing the structure of a DNA molecule, with the capacity to transport information. The alphabet of a DNA sequence is composed of letters A, T, G and C which symbolize the four nucleotides:

$$G_w = \langle v_1, v_2, v_3 \dots v_n \rangle. \quad (1)$$

A string of elements of the alphabet $\sum_{ADN}$ where:

$$\sum_{ADN} = \{A, T, G, C\}, v_x \in \sum_{ADN}. \quad (2)$$

To be able to analyze a database with DNA sequences it is necessary to have the sequences aligned. Alignment is the process by which sequences are compared by searching for common characters and establishing the residues of correspondence between related sequences, to highlight areas of similarity, which could indicate functional, evolutionary or interesting relationships for analysis. These regions are interesting because if there is a high degree of conservation, the probability of amplification for a PCR increases A set of aligned strings is assigned the name set S:

$$S = \{G_w : w = 1,2, \dots m\}, |S| = m. \quad (3)$$

Each instance G_w belongs to a class C_y where y represents the *y-th* class value that is, the type of HCV. Each position or nucleotide of a set of aligned sequences S was

assigned the attribute name A_i where i indicates the position of the nucleotide. The domain $D(A_i)$ is equal to the set of values $v_x^i : x$ is the *x-th* value in the domain of the attribute A_i (See Figure 1).

Considering the concepts and the nomenclature previously described, we can formally describe the problem that we want to solve. In a set S of DNA sequences or instances G_w, we want to locate those attributes A_i that provide more information and are considered as the best attributes to solve the classification problem of the seven C_y classes of the hepatitis C virus that currently exist. These attributes must belong to a conserved region and consider the criteria that favor a molecular diagnostic test by PCR.

4 Proposed Solution

In the context of classification, the quality of an attribute A_i has to do with its capability to separate the instances G_w, between the different possible classes. If there is a direct relationship between the values of the attributes and the possible classes, it means that the attribute is very good to classify. The quality of an attribute has to do with what classes can be separated each time we instantiate that attribute A_i.

The classes are well separated when each subgroup is generated by the division of the product is homogeneous, that is: in each subgroup all the G_w belong to the same class once we instantiate the attribute with x. This is represented as S_x^i where *x-th* of the domain of A_i. Therefore, a homogeneity metric is necessary.

4.1 Shannon Entropy

Shannon entropy comes from information theory, which can be interpreted as the degree of error or the certification of a classification problem, which is a good way to measure homogeneity. It is defined as $H(S)$:

$$H(S) = -\sum_{y=1}^{N} P(C_y) \times Ln\left(P(C_y)\right). \quad (4)$$

Where $P(C_y) = \frac{|C_y|}{|S|}$ corresponds to the probability that an element belongs to the C_y class and N the total number of classes.

4.2 Information Gain

The information gain allows quantify the information provided by an attribute A_i with respect to the classification problem. The information gain is defined as the difference between the entropy of Shannon before $H(S)$ and then $H\ (S\ |A_i)$ of knowing the value of the attribute A_i:

$$I_G(A_i) = H(S) - H\ (S\ |A_i). \quad (5)$$

The attribute A_i subdivides the instances of S into z_i subgroups S_x^i $(x = 1, \dots, z_i)$ where $z_i = |D(A_i)|$ that is, the number of values that the attribute can present. To calculate the entropy of $H\ (S\ |A_i)$, it is calculated as the weighted average $\frac{|S_x^i|}{|S|}$ of the Shannon entropy in each subgroup S_x^i:

$$H(S\ |A_i) = \sum_{x=1}^{z_i} P\left(\frac{|S_x^i|}{|S|}\right) \times H\ (S|S_x^i), \tag{6}$$

Where:

$$H\ (S|S_x^i) = -\sum_{y=1}^{N} P\left(S_y^C(S_x^i)\right) \times Ln\left(P\left(S_y^C(S_x^i)\right)\right). \tag{7}$$

The function $P\left(S_y^C(S_x^i)\right) = \frac{|S_y^C(S_x^i)|}{|S_x^i|}$, is the probability that an element $S_y^C(S_x^i)$ belongs to the class C_y and if the element belongs to the subgroup S_x^i.

C_y	G_w	A_1	A_2	A_3	A_4	A_5
C_1	G_1	A	T	T	A	T
C_1	G_2	A	T	T	C	T
C_2	G_3	A	G	C	A	C
C_2	G_4	A	G	C	G	C
C_2	G_5	A	G	C	A	T
C_2	G_6	A	G	C	G	T
C_3	G_7	G	T	C	T	C
C_3	G_8	G	T	C	T	C
C_3	G_9	G	T	C	A	G
C_3	G_{10}	G	T	C	T	G
C_4	G_{11}	A	G	A	A	C
C_4	G_{12}	A	G	A	G	C

$S_x^i = S_G^1$ $S_y^C(S_x^i) = S_A^C(S_A^3)$

Fig. 1. Representation of S set of instances G_w, where each instance belongs to a class C_y. Each position or nucleotide of a set of aligned sequences S was assigned the attribute name A_i where i indicates the position of the nucleotide. The attribute A_i subdivides the instances of S into z_i subgroups S_x^i $(x = 1, \dots, z_i)$ where $z_i = |D(A_i)|$ $S_y^C(S_x^i)$ expresses that the subset S_x^i belongs to the C_y class.

To understand the above, in Table 1 an example associated with the values is shown. When applying the formula 5 for each of the attributes of the set S it is observed that the attribute A_3 is evaluated with the greatest information gain and divides in three subsets to the set S. The first one is S_C^3 where its instances belong to the classes C_2 and C_3. The second subset is S_A^3 with all its instances belonging to class C_4. The last subset is the S_T^3 where all instances belong to the class C_1. Because the subset S_C^3 does not have instances of a single class, the information gain analysis is performed again applying formula 5. If two instances do not have the same value for each attribute and belong to different classes, the attributes are suitable to carry out

the classification, as can be seen, the subgroup S_C^3 the attributes A_1 and A_2 allow to classify the elements correctly for the classes C_2 and C_3.

This example shows that it is possible to classify DNA sequences using the concepts of entropy and information gain. For this case the attribute A_3 with greater I_G is selected, this attribute allows to quickly discriminate the classes C_1 and C_4. When calculating again I_G of all the attributes, it is obtained that both A_1 and A_2 allow to discriminate the classes C_2 and C_3. The above can be displayed in a simple way in a decision tree where each vertex has a maximum of 4 possible values (Figure 2).

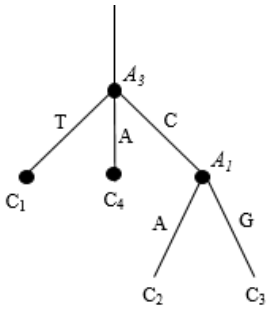

Fig. 2. Decision tree for the classification of the data presented in Fig. 1.

In addition to using the information gain criterion to select the attribute that best separates the classes, it is necessary contemplate that the results will be used to mount a PCR diagnostic test. It is necessary to consider some criteria that can optimize the design of the test.

Lefever et al. [21] demonstrated the effect that the type of mismatch has on the alignment and position in a primer on extension efficiency during the first cycles of the PCR. Finding minimal or no extension step [7] when they introduced a mismatch in the last 3 or 4 nucleotides of the primer at the 3' end. Their hypothesis was that the low extension was caused by the reduction in the binding of the enzyme DNA polymerase to the binding site [7]. He concluded that the closer the mismatch to the extreme 3' was, the greater the impact it had during the extension of the PCR, increasing the number of cycles where the fluorescence signal was detected in the PCR.

Using the analysis carried out by Lefever, an evaluation function was designed to consider its contributions and, in addition, to consider the criterion of information gain.

The criteria used for the evaluation function $E(A_i, d)$ was, the attribute of interest A_i must be evaluated in I_G to consider it a good attribute to discriminate between classes. The attributes that surround it must be also be evaluated to establish if it is a good option for the primer design.

The attributes surrounding the A_i attribute were called window $\varphi(A_i, d)$ (see Figure 3), the window $\varphi(A_i, d)$ can be expressed as:

$$\varphi(A_i, d) = \varphi(A_i, d)^- \cup \varphi(A_i, d)^+ ; d = distance, \quad (8)$$

Such that:

$$\varphi(A_i, d)^- = \{ A_J : J = i - d, \dots, i - 1 \} \text{ and } \varphi(A_i, d)^+ = \{ A_J : J = i, \dots, i + d \}. \quad (9)$$

C_y	$\mathcal{G}_w$			A_i		
C_1	$\mathcal{G}_1$	A	T	T	A	T
C_1	$\mathcal{G}_2$	A	T	T	C	T
C_2	$\mathcal{G}_3$	A	G	C	A	C
C_2	$\mathcal{G}_4$	A	G	C	G	C
C_2	$\mathcal{G}_5$	A	G	C	A	T
C_2	$\mathcal{G}_6$	A	G	C	G	T
C_3	$\mathcal{G}_7$	G	T	C	T	C
C_3	$\mathcal{G}_8$	G	T	C	T	C
C_3	$\mathcal{G}_9$	G	T	C	A	G
C_3	$\mathcal{G}_{10}$	G	T	C	T	G
C_4	$\mathcal{G}_{11}$	A	G	A	A	C
C_4	$\mathcal{G}_{12}$	A	G	A	G	C

$\varphi(A_i, d)^+$ ← A_i → $\varphi(A_i, d)^-$

Fig. 3. Representation of the evaluation window $\varphi(A_i, d)$ where d indicates the number of attributes that relate to the analysis of the window to the right and to the left of the attribute of interest A_i.

Therefore, the selection criterion was established by means of the evaluation function $E(A_i, d)$ is simply the product between $I_G(A_i)$ and the evaluation of the window $E\varphi(A_i, d)$:

$$E\varphi(A_i, d) = -I_G(A_i) \times E\varphi(A_i, d). \quad (10)$$

Where $E\varphi(A_i, d) = \sum_{-d}^{d} (W_i^{-1} \times H(A_i))$ such that $E\varphi(A_i, d)$ is the evaluation of window $\varphi(A_i, d)$ with d attributes before and after position A_i.

W_i is the weight of the attribute W_i taken from the Lefever experiments defined as $\sum \frac{dC_q}{R_i}$ the sum of the differences between C_{qMM} number of cycles for the amplification with the unadjusted alignment and the number of cycles for the amplification with the alignment perfect C_{qP} [21]. It is simplified in Table 1, based on the results obtained by Lefever.

Table 1. Weights W_i and its associated value by position based on Lefever's results.

$\boldsymbol{W_i}$	Value associated with $\boldsymbol{W_i}$
0	0.687
1	0.057
2	0.031
3	0.016
4	0.012
5-20	0.014

The entropy of the attribute A_i is defined as:

$$H(A_i) = \sum_{x=1}^{z_i} P(v_x^i) \times ln\left(P(v_x^i)\right). \quad (11)$$

Such that $P(v_x^i) = \frac{|v_x^i|}{s}$.

The previous analysis indicates that the concepts of entropy and information gain allow to classify DNA sequences and due to the study of Lefever can be considered criteria that can favor the design of primers for a PCR. Based on the above analysis we design the Algorithm 1 that receives a database with instances of DNA sequences and selects the best attributes by analyzing them using (10). Finally, it returns a decision tree where each node is an attribute that solves the classification problem.

Algorithm 1. Receive from a database with instances of DNA sequences. Returns a decision tree where each node is an attribute that solves the classification problem.

```
id3Ev (instances, target_attribute, attributes)
begin
  Create a new root node to the tree;
  If all instances have the target_attribute belonging to the
same Cv
    Return the tree with single root node with label Cv;
  If attributes is empty, then
    Return the tree with single root note with the most common
label of a target_attibute in instances;
  Else
    Ai := The attribute in attributes which best classifies in-
    stances;
    root decision attribute := Ai;
    For each possible value vi of Ai
      Add new ramification below root, corresponding to the test
      Ai = vi;
      Let instancesvi be the subset of instances with the value
      vi  for Ai;
      If instancesvi is empty then
        Bellow this ramification, add a new leaf node with the
        most common value of target_attribute in instances;
      Else
        DTC (instancesvi, target_attribute, attributes);
end
```

5 Discussion

The in silico design [21] of the primers can allow a tremendous saving of time and money in the development of diagnostic tests, however, they can never replace the experimental verification tests since it is not possible to predict the specificity of a first in silico.

The proposed methodology seems to be an adequate solution to help researchers design primers to solve STP problems. As mentioned in the introduction, researchers are currently limited to observing the sequences and manually determining which regions are suitable for solving the classification problem. This paper presents the design of a methodology as a solution proposal to solve the classification problem through the concepts of information gain and an evaluation function that quantifies the value of the conserved area where the attribute are of interest. This evaluation function uses the W_i weight, which is a heuristic value, calculated with the results offered by Lefever in its work.

Algorithm 1 is currently in the implementation stage to perform tests later. It is expected that the algorithm will be evaluated not only by a mathematical analysis but also be analyzed by experts in the area of molecular biology and diagnostics that can determine the quantitative and qualitative quality of the method.

6 Conclusions

The example presented in section 4.2 shows that it is possible to find a way to classify all instances of the database with DNA sequences using the concepts of entropy and information gain.

The proposed algorithm in addition to using these concepts incorporates an evaluation function that establishes criteria that could favor the design of primers for a PCR. This can be useful to solve the task of classifying genetic sequences with high variability rates. Algorithm 1 finally proposes a decision tree where the nodes are the suggested attributes for researchers to design primers and a PCR diagnostic test to detect the seven types of Hepatitis C virus that currently exist.

References

1. Panduro, A.: Biología molecular en la clínica. McGraw-Hill Interamericana (2000)
2. Eisele, J. G., Roldán, C. Y. C., Galindo, M. O., Gil, M. P.: Usefulness of solution algorithms of the traveling salesman problem in the typing of biological sequences in a clinical laboratory setting. In: 14th IEEE International Conference on Electronics, Communications and Computers CONIELECOMP 2004, pp. 264–269 (2004)
3. Mohd Hanafiah, K., Groeger, J., Flaxman, A.D., Wiersma, S.T.: Global epidemiology of hepatitis C virus infection: new estimates of agespecific antibody to HCV seroprevalence. Hepatology 57(4), 1333–1342 (2013)
4. World Health Organization, http://apps.who.int/iris/bitstream/10665/255016/1/9789241565455-eng.pdf?ua=1
5. Messina, J. P., Humphreys, I., Flaxman, A., Brown, A., Cooke, G. S., Pybus, O. G., Barnes, E.: Global distribution and prevalence of hepatitis C virus genotypes. Hepatology 61(1), 77–87 (2015)
6. Sharkey, D.: Antibodies as thermolabile switches: High temperature triggering for the polymerase chain reaction. Biotechnology 12, 506–509 (1994)

7. Eckert, K.A., Kunkel, T.A.: DNA polymerase fidelity and the polymerase chain reaction. PCR Methods Appl. 1, 17–24 (1991)
8. Mas, E., Poza, J., Ciriza, J., Zaragoza, P., Osta, R., Rodellar, C.: Fundamento de la Reacción en Cadena de la Polimerasa (PCR). AquaTIC 15, 70–78 (2016).
9. Tamay de Dios, L., Ibarra, C., Velasquillo, C.: Fundamentos de la reacción en cadena de la polimerasa (PCR) y de la PCR en tiempo real. Investigación en discapacidad 2(2), 70–78 (2013)
10. Abd-Elsalam, K. A.: Bioinformatic tools and guideline for PCR primer design. African Journal of biotechnology 2(5), 91–95 (2003)
11. Bioinformatics at COMAV, bioinf.comav.upv.es/courses/intro_bioinf/multiple.html
12. Lozano Yécora J.: La genotipificación del virus de papiloma humano: una familia de problemas de optimización combinatoria. Undergraduate thesis, Universidad de las Américas, Puebla(2004)
13. Fernández, R.; Desarrollo de Algoritmos para la Clasificación de Secuencias. Master's Thesis, Universidad de las Américas, Puebla (2002)
14. Rodríguez, R. S., Roldán, C. Y. C., Eisele, J. G., Gil, P. G., Galindo, M. J. O.: Algorithms for the typing of related DNA sequences. In: 15th IEEE. International Conference on Electronics, Communications and Computers CONIELECOMP'05, pp. 268–271 (2005)
15. Benish, W.A.: Relative entropy as a measure of diagnostic information. Med Decis Making 19(2), 202–206 (1999)
16. Jiménez, M.: On the syntactic and redundancy distribution of the genetic code. BioSystems 32, 11–23 (1994)
17. Gotoh, O.: Multiple sequence alignment: algorithms and applications. Adv Biophys 36, 159–206 (1999)
18. Ebeling, W., Frommel, C.: Entropy and predictability of information carriers. Biosystems 46(1-2), 47–55 (1998)
19. Solis, A.D, Rackovsky, S.: Optimized representations and maximal information in proteins. Proteins 38(2), 149–164 (2000)
20. Lefever, S., Pattyn, F., Hellemans, J., Vandesompele, J.: Single-nucleotide polymorphisms and other mismatches reduce performance of quantitative PCR assays. Clinical chemistry 59(10), 1470–1480 (2013)
21. Okada, M., Tsukamoto, M., Ohwada, H., Aoki, S.: Consensus scoring to improve the predictive power of in-silico screening for drug design. In: Proceedings of the 2nd International Conference on Engineering and Meta-Engineering, pp. 27–30 (2011)