Online journal ISSN - 1676-5680
GMR Genetics and Molecular Research
EVOLUTION AND TECHNOLOGY

Routine libraries for pattern recognition in quasispecies

E.A. Marucci[1], G.F.D. Zafalon[1], A.C.G. Jardim[2], L.H.T. Yamasaki[2], C. Bittar[2], P. Rahal[2] and J.M. Machado[1]

[1]Laboratório de Bioinformática, Universidade Estadual Paulista,
São José do Rio Preto, SP, Brasil
[2]Laboratório de Estudos Genômicos, Universidade Estadual Paulista,
São José do Rio Preto, SP, Brasil

Corresponding author: G.F.D. Zafalon
E-mail: zafalon@gmail.com

Genet. Mol. Res. 7 (3): 970-981 (2008)
Received june 2, 2008
Accepted August 11, 2008
Published September 30, 2008

ABSTRACT. The results obtained through biological research usually need to be analyzed using computational tools, since manual analysis becomes unfeasible due to the complexity and size of these results. For instance, the study of quasispecies frequently demands the analysis of several, very lengthy sequences of nucleotides and amino acids. Therefore, bioinformatics tools for the study of quasispecies are constantly being developed due to different problems found by biologists. In the present study, we address the development of a software tool for the evaluation of population diversity in quasispecies. Special attention is paid to the localization of genome regions prone to changes, as well as of possible hot spots.

Key words: Bioinformatics; Quasispecies; Pattern recognition; Genetic variability

INTRODUCTION

The emerging field of bioinformatics is constantly under development, and has recently stimulated a deep interest among computer scientists (Liew et al., 2005). This interest is related mainly to the fact that biologists need the power of computer processing to analyze the results obtained from experimental research, making it possible to recognize patterns in different sequences of nucleotides or amino acids. The patterns can show many important pieces of information for the biologists in their studies.

Very often, the amount of data that biologists obtain in their research becomes unmanageable with manual analysis. Therefore, both computers and specialized algorithms play an essential role in performing this study.

Choosing and constructing the correct algorithm for each case is the most important step in obtaining the correct answer in an acceptable time (Anbarasu et al., 1999). The main issue is to define which algorithm will be chosen and how to implement it to achieve the correct answer and the best optimization.

Generally, quasispecies analysis needs a special algorithm designed for each case to be considered. In other words, this is an important field of bioinformatics and must be explored through studies trying to develop the most generic algorithm to treat the quasispecies problem. In the next section, we will realize that quasispecies analysis leads to many distinct problems.

Quasispecies problem

An important feature of viruses with RNA genome is their genetic variability. This variability may have important consequences on the pathogenesis of infections and related diseases (Pawlotsky, 2003).

Hepatitis C virus (HCV) is an RNA virus that represents the major causes of liver disease in the world. The resulting genetic variability defines a classification in clades, genotypes, subtypes, isolates, and quasispecies (Le Guillou-Guillemette et al., 2007). Furthermore, in infected individuals, quasispecies circulate as a population of many different but closely related viral variants (Martell et al., 1992; Domingo et al., 2006). The average nucleotide differences, along with the entire HCV genome, are approximately 20-30%, 10-20% and less than 10% for HCV genotypes, subtypes and quasispecies, respectively (Zhou et al., 2007). Genetic variability has been studied in many regions of the HCV. However, most studies analyzing HCV quasispecies have been directed toward short sequences of the viral genome, producing little information about entire regions of the virus with encoded proteins relevant to mechanisms of viral replication and propagation.

In quasispecies, the analysis of lengthy nucleotide sequences of the viral genome is not feasible without the use of bioinformatics tools, because variants may have several nucleotide mutations among them.

Therefore, we developed a bioinformatics tool we called LOCQSPEC ("Localizador de Quasiespécies", or Quasispecies Locator). It is a tool built on C++ programming language, which makes it possible to compare a set of sequences, determining whether they are similar or different sequences. In the different sequences, it is possible to find the spot places where changes of nucleotides or amino acids are present.

Based on the results, the population diversity of quasispecies may be evaluated, and it is possible to investigate a genome permissive to change and possible hot spots.

To the extent of our knowledge there is no software available that performs the same work as ours. The way our software was built had a particular intention: to improve the work of research biologists in the Laboratory of Genomics Studies, who need software that is easy to operate for analysis of results. Before we created the software, these scientists performed an exhaustive search for software similar to LOCQSPEC, but they could not find any.

This set of routines is intended to be an open source. We intend to keep the source code available for download and also to create a web version for researchers' use. This web version will be published in a future paper.

Sequence alignments

There are several softwares to perform sequence alignments, each of them working in a particular way. Some of them can be used in general to align sequences and uncover patterns among them, and many others are for biological inferences. For instance, Gao and Qiao (2000) described a general parallel implementation algorithm. It performs the operations of pairwise alignment, star alignment, phylogeny reconstruction, and generalized tree alignment.

In the context of pattern recognition, there are also several efficient tools for revealing these patterns. A method has been described that detects binding sites in coding regions (Blanchette, 2003).

Otherwise, there are softwares for specific proposals. For instance, in the study of Kececioglu and Starrett (2004), we may find an algorithm refinement for sequence alignments where it is intended to minimize the linear gap-costs. With respect to quasispecies research, there are researchs showing some progress in this area. Buendia and Narasimhan (2004) used multiple distance matrices and correlation rules to output the phylogenetic tree.

In this paper, we consider a more simplified approach for the problem of quasispecies search. For the applications we devised, this simplified approach results in ease of software maintenance and more user-friendly interfaces, when compared to other methods.

MATERIAL AND METHODS

Methods applied to routines

Here, we give an overview of how the routines were developed and how they were applied in the context of quasispecies problems. The flowchart presented in Figure 1 shows the order in which each section is performed in the algorithm.

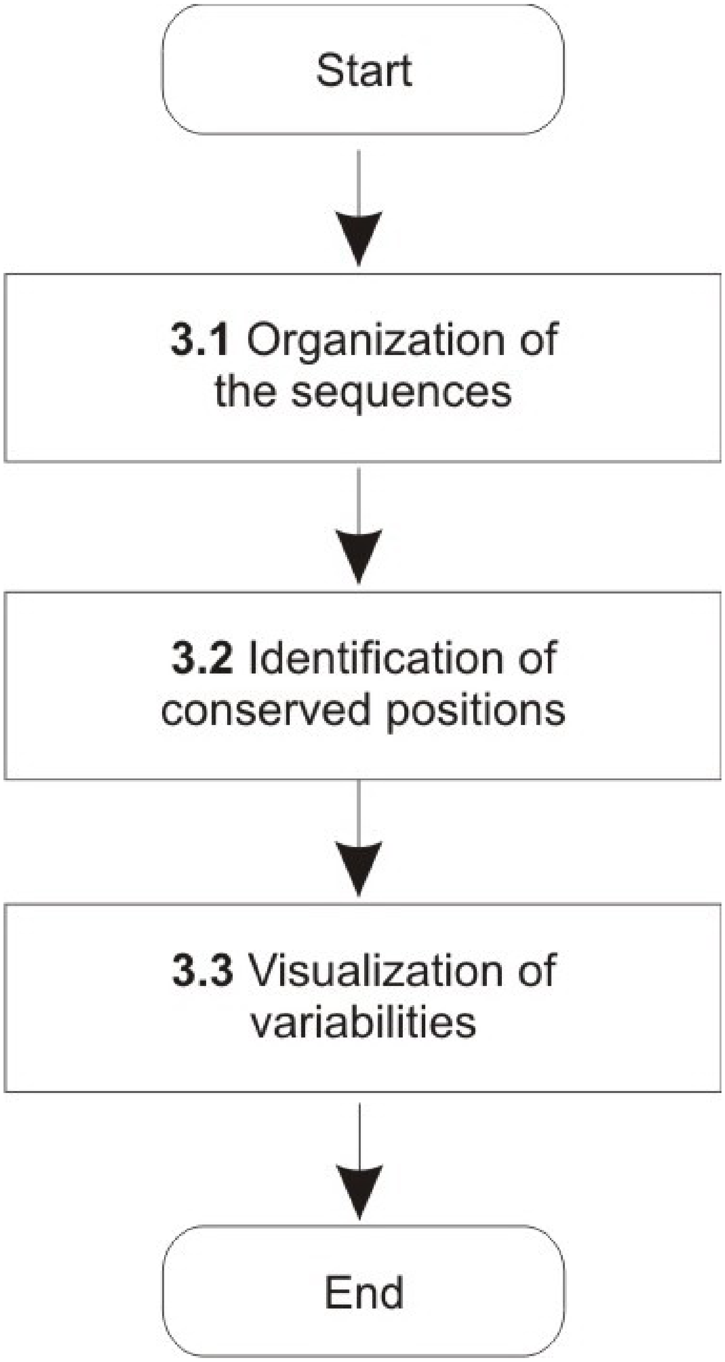

Start
3.1 Organization of the sequences
3.2 Identification of conserved positions
3.3 Visualization of variabilities
End

Figure 1. Overview of the algorithm.

Organization of the sequences

The first step is to create groups of sequences. The flowchart in Figure 2 illustrates this part of the algorithm.

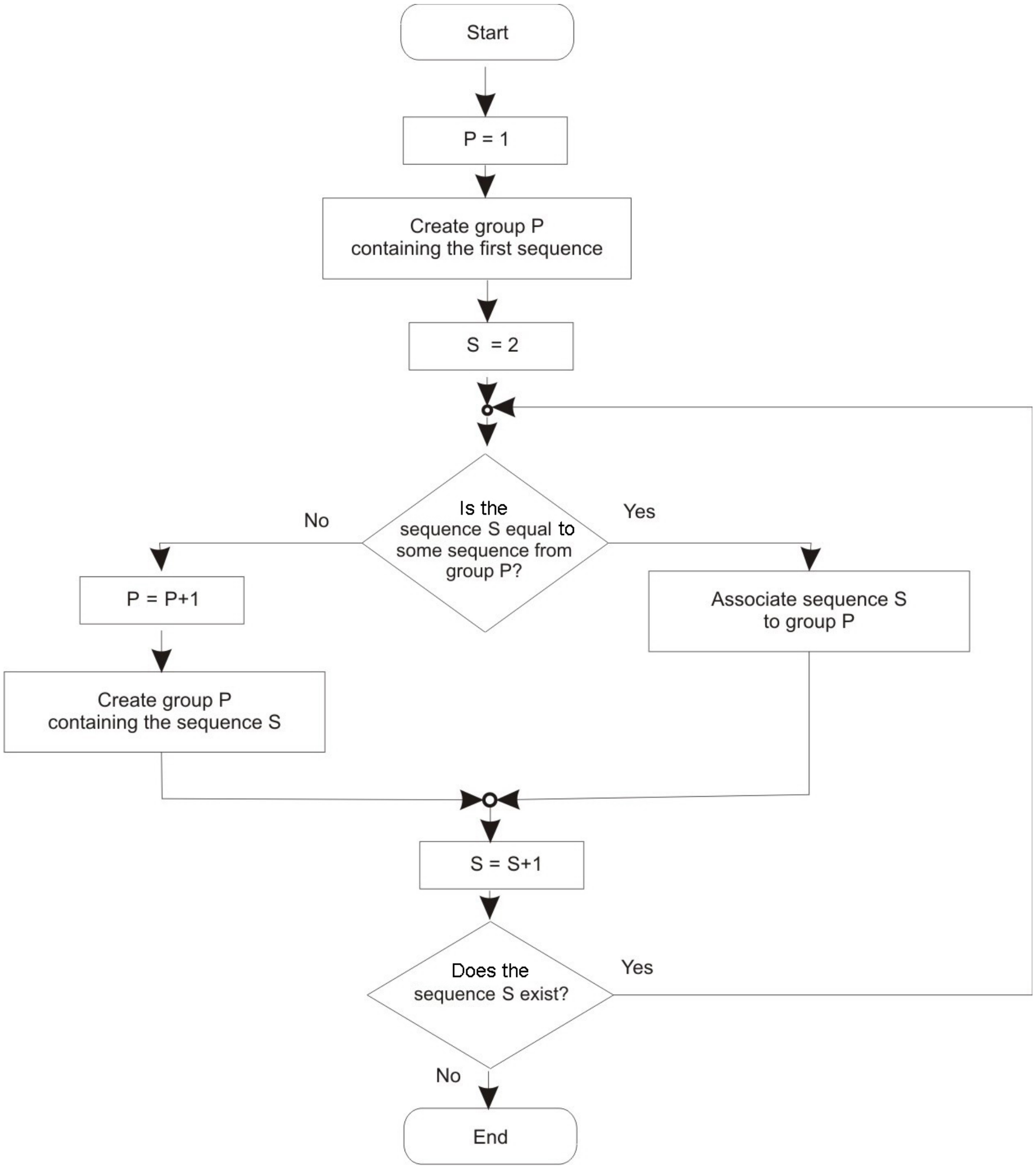

Start
P = 1
Create group P containing the first sequence
S = 2
Is the sequence S equal to some sequence from group P?
No
Yes
P = P+1
Create group P containing the sequence S
Associate sequence S to group P
S = S+1
Does the sequence S exist?
Yes
No
End

Figure 2. Organization of the sequences in the flowchart.

Each group (P) corresponds to a set of identical sequences (S) that can be found more than once with distinct names in the sequence entry sets. The construction of each group starts with an identity verification based on parity among sequences that have been grouped. For the first grouping, verification is not necessary. The first sequence of the set is directly associated with the first group. For the other sequences, their identity is verified relative to the entire set of existing groups. There is an optimization in this process to stop the verification whenever a mismatch is found. Otherwise, when there is an identity, the sequence is associated with the group, and the next sequence to be verified is begun. If all existing groups are verified and no identity is found, a new group is formed, and the new sequence is associated with it. The entry sequences must be given as FASTA format in order to be correctly processed by LOCQSPEC.

This grouping is the first step to verify the differences among all or among some sequences of the set. As in the study of the species' variability, some sequences may not show variation. Thus, one grouping should be used to avoid redundance in a later analysis, and to make the visualization of the results clearer.

Identification of conserved positions

When a grouping is completed, the tool starts by searching the conserved positions among the sequences of interest. The sequences of interest may be from only one patient, from many of them, from particular sequences, or from a whole entry set. The identification of conserved positions is performed in groups that contain these sequences. These groups are defined from the parameter *setgroupstart* that is passed to the tool. This parameter allows the definition of all groups, which contain sequences whose names start with a determined value. An example of the usage for *setgroupstart* can be seen in the Results section. The advantage of this option is the possibility to use only one entry file to analyze different sets of sequences, thus improving the tool's efficiency.

The process of identifying the conserved positions is carried out for the defined groups.

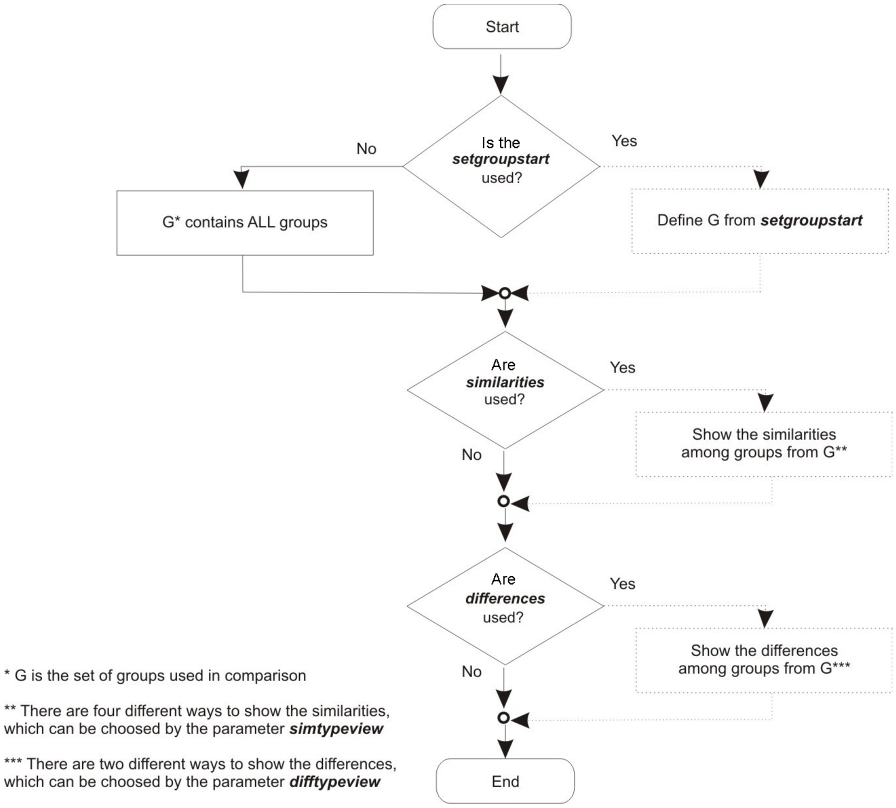
This identification locates, among the groups, which columns are conserved. Each of these columns is conserved with a specific residue. Both the column position and residue are stored in an array of similarities.

Each sequence column is compared, and those not showing variation are stored in this array. The different positions are not present in the array.

From this array, the information of interest is shown on the screen according to the parameter, which was provided to the program. The *similarities* flag shows the similarities among groups. The original form shows the symbols in the conserved positions and lines in variation locations. Nevertheless, other forms of visualization are also available. One of them shows all conserved positions; another exhibits motifs and is useful in pattern recognition.

In the same way, the positions where differences occur are exhibited using the flag *differences*. However, all differences are only shown in the variability visualization step, together with the related statistics for them. The statistics indicate the relative amount of occurrences for specific profiles. As the variability shows a large set of information, the possibility to define the places of interest makes the investigation of quasispecies feasible.

The flowchart in Figure 3 shows the order in which the parameter *setgroupstart* and the flags *similarities* and *differences* are analyzed and how they work. The continuous line region shows the default option for the algorithm.


Start
Is the *setgroupstart* used?
No
Yes
G* contains ALL groups
Define G from *setgroupstart*
Are *similarities* used?
Yes
No
Show the similarities among groups from G**
Are *differences* used?
Yes
No
Show the differences among groups from G***
End

* G is the set of groups used in comparison

** There are four different ways to show the similarities, which can be choosed by the parameter ***simtypeview***

*** There are two different ways to show the differences, which can be choosed by the parameter ***difftypeview***

Figure 3. Identification of conserved positions in the flowchart.

Visualization of variabilities

Useful information for identifying which kind of variability has occurred is shown only in one region or in a specific place of sequences, although it is possible to exhibit this information in the whole sequence. This kind of generalization makes a specific investigation unfeasible, due to the large set of information that needs to be shown on the screen. Thus, it is desirable to define one region to make the analysis. This place of analysis is defined with knowledge of information obtained in the last step and it is passed to the program through the parameter *pos.*

For one specific position, the program shows all the residues that are present in that position. For a range of positions from a given initial position to a final position, one set of grouped residues is exhibited. All the profiles obtained are shown. Together with them some important information is presented. Using the flag *showsimgroups*, it may be demonstrated, for instance, in which groups each of these profiles appear.

The flowchart in Figure 4 shows the order in which the parameter *pos* and the flag *showsimgroups* are analyzed and how they work. The continuous line region shows the default option for the algorithm. In the dotted line region, we have another conditional structure, whose default option is also indicated.

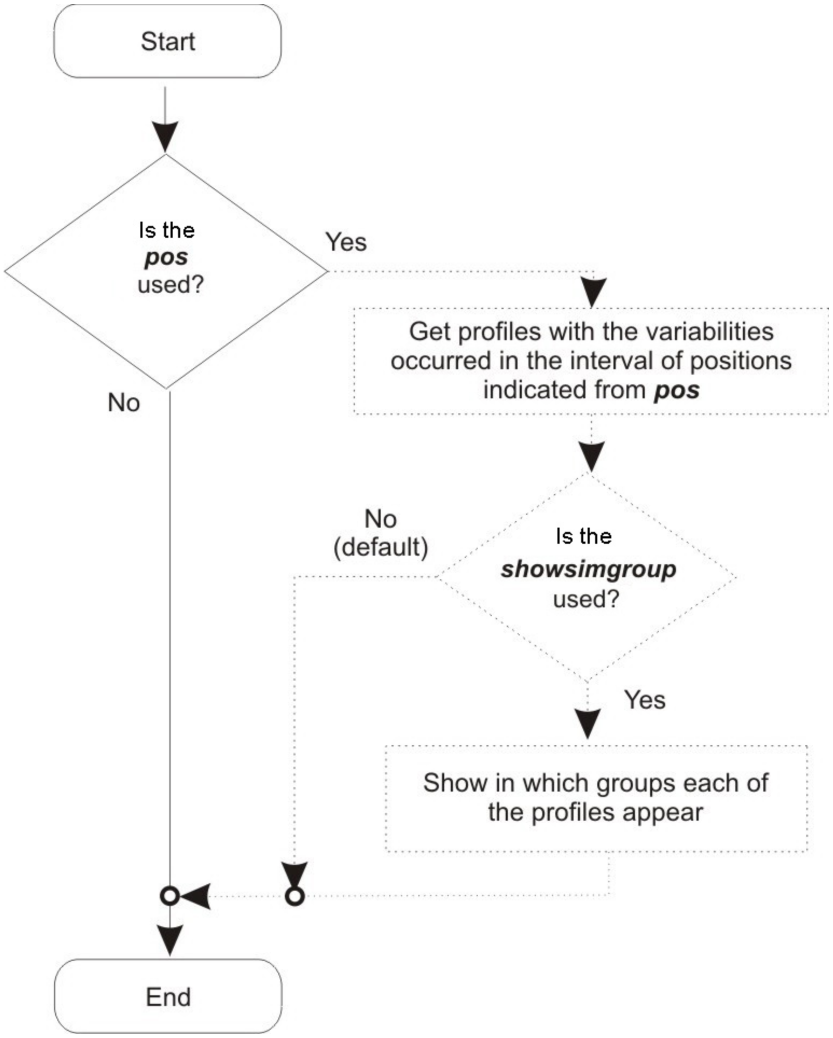

Start
Is the *pos* used?
Yes
No
Get profiles with the variabilities occurred in the interval of positions indicated from *pos*
No (default)
Is the *showsimgroup* used?
Yes
Show in which groups each of the profiles appear
End

Figure 4. Visualization of variabilities in the flowchart.

RESULTS

To demonstrate the functionality of the tool, we executed some tests taking many variations of input parameters. The FASTA file used has 165 aligned sequences.

The first test executes the tool in its basic configuration, when only the input file was informed without any additional parameter. The grouping of these sequences is shown in the screen and it identifies the identical sequences. In this test, 118 distinct groups were identified. Almost all of them have only one sequence. Nevertheless, some sequence groups showed up to 12 identical sequences. One part of these groups is presented as follows:

```
# locqspec –in input.fas

 Group [   1]: NS5A1a


 Group [   2]: 0314, 0310
 Group [   3]: 0316
 ...
 Group [ 101]: 4417
 Group [ 102]: 3904, 3906, 3907, 3908, 3910, 3912, 3916, 3913, 3903,
               3915, 3918, 3902
 Group [ 103]: 3917
 ...
 Group [ 117]: 4213
 Group [ 118]: 4204
```

In the second test, the flag *similarities* is passed to the tool, and is required from the program to identify the similarities among all 118 groups. The way to exhibit these similarities is through symbols, showing only the conserved residues. Those that had variation in at least one group are shown with a trace.

```
# locqspec –in input.fas –similarities –simtypeview symbol

Similarities among groups (Symbols):
S-SWL-D--DWIC-VL-DF-TWL-----P-LPG-P--SCQ-GY-GVW--DGIM-T-C- CGA-ITGHVKNG-
MRI--P-T--N-W-GTFP-NAYTTGPC-P-P--NY---LW-V--- EYVE---VGDFHYV-GM-TDN-KCPC--
P-P-FF---DG-R-HR-A----P-LR--V-F-VGL--Y----QLPCE-EP-V-V-TSML----HITAE-A---
LA-GSPPS--SS-AS QLSAPSL--TCT--HDSPDA-LIEANLLWRQ-MG--I-RVESE-KVV-LDSF--L-AE
E-E-E-S-PAEILR--R-F--A-P-W-R-DYNP-L-E-WK--DY-P--VHGC-LP--- -P--PPPR--R--
VLTES-----L-EL--K----S--S-------------------- -SD-ES-S-MPP-EGEPGD-DL-----
STVS-----EDVVCC
```

In case it is desired to show the specific position of differences, the *differences* flag may be used. The results are shown below.

```
# locqspec –in input.fas –differences

Differences among groups (Positions):
2:6:8:9:14:17:20:24:25:26:27:28:30:34:36:37:41:44:48:49:54:56:58:62:71:75:7
6:78:80:81:83:85:90:99:101:103:104:107:108:109:112:114:115:116:121:122:123:
131:134:138:143:144:146:148:151:152:153:156:158:161:163:164:165:166:168:171
:172:174:176:180:181:183:184:185:186:192:195:197:199:204:205:206:207:213:21
5:216:217:220:226:227:230:240:241:245:246:253:264:267:268:270:276:280:285:2
86:288:292:294:296:298:305:306:308:310:311:313:315:317:319:324:326:328:331:
332:335:337:338:343:346:347:348:349:351:352:357:358:360:361:367:368:369:370
:371:373:376:377:379:380:381:382:384:385:387:388:389:390:391:392:393:394:39
5:396:397:398:399:400:401:402:403:404:405:406:407:410:413:415:419:426:429:4
30:431:432:433:438:439:440:441:442:
```

Although the results show only the locations where variabilities have occurred, biological inferences cannot be made. We also need to know the kind of variability. It is also necessary to know which residues are found in this place, in order to make some inference. In the next test, specific information of regions is shown. With the parameter *pos*, the stretches of interest are indicated (if it is only one position or a range of positions) and whole differences are shown. In this test, the *showsimgroups* flag was also used to show the groups where a specific pattern occurs. The range between positions 25 and 30 has the following result:

```
# locqspec -in input.fas -pos 25-30 -showsimgroups

Differences between the 118 groups in the positions of range (25-30):

AKLMPQ - 68 occurrences (57.6271%)
 Groups: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19,
20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 37, 38, 39,
40, 41, 42, 43, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59,
60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70

AKLVPQ - 2 occurrences (1.69492%)
 Groups: 36, 44

SRVLPR - 1 occurrence (0.847458%)
 Groups: 71

SKLLPR - 47 occurrences (39.8305%)
 Groups: 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87,
88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104,
105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118
```

The last test shows the function, which selects specific parts of the input file. In this case we selected only the sequences that start with 22, those that start with 39 and sequence NS5A1a. In the following, the result is shown together with the similarities, the positions where differences occurred, and specific information for positions between 25 and 30.

```
 # locqspec -in input.fas -setgroupstart 22,39,NS5A1a
 -similarities -simtypeview symbol –differences -pos 25-30
 -showsimgroups

  Group [   1]: NS5A1a
  Group [  84]: 2201, 2207, 2203, 2204, 2213, 2216, 2208, 2205, 2210, 2209
  Group [  85]: 2220
  Group [  86]: 2211
  Group [  87]: 2202
  Group [  88]: 2206, 2214
  Group [ 102]: 3904, 3906, 3907, 3908, 3910, 3912, 3916, 3913, 3903, 3915,
                3918, 3902
  Group [ 103]: 3917
  Group [ 104]: 3909, 3911

  Similarities among groups (Symbols):
 SGSWL-D-WDWIC-VL-DFKTWL--KL-P-LPG-PF-SCQRGY-GVWRGDGIM-T-C- CGA-ITGHVKNG-
 MRIVGP-TC-N-W-GTFPINAYTTGPCTP-P-PNY--ALWRV--E EYVE--RVGDFHYV-GMTTDN-KCPCQ-
 P-PEFF-E-DGVRLHR-AP-CKPLLREEV- F-VGL--Y-VG-QLPCEPEPDV-V-TSMLTDPSHITAE-A-
 RRLARGSPPS-ASSSAS QLSAPSLKATCT--HDSPDA-LIEANLLWRQEMGGNITRVESENKVVILDSFDPL-
 AE EDE-EVSVPAEILRK-R-F-RA-P-WARPDYNPPL-E-WK-PDY-PPVVHGC-LPP-- -PP-PPPR-KRT-
 VLTEST-S-ALAEL--K-F--S--S-------T------------ -
 SDVESYSSMPPLEGEPGDPDLSDGSWSTVS-----EDVVCC
 Differences among groups (Positions):
6:8:14:17:24:25:28:30:34:37:44:54:56:58:62:71:78:81:83:85:101:103:107:108:1
14:115:121:122:131:138:144:146:151:153:161:164:174:176:180:181:183:186:197:
199:213:215:226:245:246:253:288:294:306:308:310:313:315:326:328:331:335:343
:347:348:349:352:357:361:368:370:376:377:379:381:382:384:385:387:388:389:39
0:391:392:393:395:396:397:398:399:400:401:402:403:404:405:406:407:438:439:4
40:441:442:

 Differences among 9 groups in positions of range (25-30)

 AKLMPQ - 1 occurrence (11.1111%)
 Groups: 1

 SKLLPR - 8 occurrences (88.8889%)

 Groups: 84, 85, 86, 87, 88, 102, 103, 104
```

CONCLUSIONS AND FUTURE PERSPECTIVES

The presenty study was intended to show a bioinformatics tool that works with two approaches for the analysis of quasispecies in an easy and efficient way. These two approaches were adopted according to the needs of biologists who collaborated on this study with empirical knowledge.

The use of parameters such as an input data improved the way to work with the tool, which operates in text mode. Generally, in professional tools every input is entered in the program in this way, making this the most important reason to adopt this approach.

Based on the results of last section, one sees that the tool shows the sequences and the results of sequence analysis in an easy way, to be conveniently analyzed by biologists.

The future perspectives are the development of a graphic user interface in desktop mode, in which the user will be able to execute the tool in a simpler manner. Later, a web version will be created, so that the user can access the tool on the internet and use it on any computer.

REFERENCES

Anbarasu LA, Narayanasamy P and Sundararajan V (1999). Multiple sequence alignment using parallel genetic algorithm. *Lecture Notes Artif. Intell.* 1585: 130-137.

Blanchette M (2003). A comparative analysis method for detecting binding sites in coding regions. In: Proc. Seventh Annu. Int. Conf. Res. Comp. Mol. Biol., Berlim, 57-66.

Buendia P and Narasimhan G (2004). MinPD: distance-based phylogenetic analysis and recombination detection of serially-sampled HIV quasispecies. *Proc. IEEE Comput. Syst. Bioinform. Conf.*, *Los Alamitos*, 110-119.

Domingo E, Martin V, Perales C, Grande-Perez A, et al. (2006). Viruses as quasispecies: biological implications. *Curr. Top. Microbiol. Immunol.* 299: 51-82.

Gao W and Qiao S (2000). Multithreaded implementation of a biomolecular sequence alignmentalgorithm-software/information technology. In: Proc. Canadian Conf. Electrical Comput. Engin., Ontario, 494-498.

Kececioglu J and Starrett D (2004). Align alignments exactly. In: Proc. 8th ACM Conf. Res. Comput. Mol. Biol. - RECOMB'04, San Diego, 85-96.

Le Guillou-Guillemette H, Vallet S, Gaudy-Graffin C, Payan C, et al. (2007). Genetic diversity of the hepatitis C virus: impact and issues in the antiviral therapy. *World J. Gastroenterol.* 13: 2416-2426.

Liew AWC, Yan H and Yang M (2005). Pattern recognition techniques for the emerging field of bioinformatics: A review. *Pattern Recognit.* 38: 2055-2073.

Martell M, Esteban JI, Quer J, Genesca J, et al. (1992). Hepatitis C virus (HCV) circulates as a population of different but closely related genomes: quasispecies nature of HCV genome distribution. *J. Virol.* 66: 3225-3229.

Pawlotsky JM (2003). Hepatitis C virus genetic variability: pathogenic and clinical implications. *Clin. Liver Dis.* 7: 45-66.

Zhou D, Fan X, Tan D, Xu Y, et al. (2007). Separation of near full-length hepatitis C virus quasispecies variants from a complex population. *J. Virol. Methods* 141: 220-224.