BIOINFORMATICS

Vol. 17 no. 1 2001
Pages 16–22

# PAQ: Partition Analysis of Quasispecies

*Prasith Baccam[1,6], Robert J. Thompson[2], Olivier Fedrigo[3], Susan Carpenter[4] and James L. Cornette[5]*

*[1]Department of Mathematics and Department of Veterinary Microbiology and Preventive Medicine, [2]Interdepartmental Genetics Program, [3]Department of Zoology and Genetics, [4]Department of Veterinary Microbiology and Preventive Medicine and [5]Department of Mathematics, Iowa State University, Ames, IA 50011, USA*

Received on December 22, 1999; revised on March 14, 2000 and August 15, 2000; accepted on August 16, 2000

## ABSTRACT

**Motivation:** The complexities of genetic data may not be accurately described by any single analytical tool. Phylogenetic analysis is often used to study the genetic relationship among different sequences. Evolutionary models and assumptions are invoked to reconstruct trees that describe the phylogenetic relationship among sequences. Genetic databases are rapidly accumulating large amounts of sequences. Newly acquired sequences, which have not yet been characterized, may require preliminary genetic exploration in order to build models describing the evolutionary relationship among sequences. There are clustering techniques that rely less on models of evolution, and thus may provide nice exploratory tools for identifying genetic similarities. Some of the more commonly used clustering methods perform better when data can be grouped into mutually exclusive groups. Genetic data from viral quasispecies, which consist of closely related variants that differ by small changes, however, may best be partitioned by overlapping groups.

**Results:** We have developed an intuitive exploratory program, Partition Analysis of Quasispecies (PAQ), which utilizes a non-hierarchical technique to partition sequences that are genetically similar. PAQ was used to analyze a data set of human immunodeficiency virus type 1 (HIV-1) envelope sequences isolated from different regions of the brain and another data set consisting of the equine infectious anemia virus (EIAV) regulatory gene *rev*. Analysis of the HIV-1 data set by PAQ was consistent with phylogenetic analysis of the same data, and the EIAV *rev* variants were partitioned into two overlapping groups. PAQ provides an additional tool which can be used to glean information from genetic data and can be used in conjunction with other tools to study genetic similarities and genetic evolution of viral quasispecies.

[6]To whom correspondence should be addressed at Los Alamos National Laboratory, MS K-710, Theoretical Biology and Biophysics, Los Alamos, NM 87545, USA.

**Availability:** http://www.vetmed.iastate.edu/units/carplab/PAQ/main.html

**Contact:** pbaccam@lanl.gov

## INTRODUCTION

The rapid accumulation of genetic data resulting from high-throughput sequencing and the various genome sequencing projects necessitates a variety of computer programs to analyze the genetic data. Phylogenetic programs like PHYLIP (Felsenstein, 1993) and PAUP (Swofford, 1999) have been increasingly utilized to analyze the evolutionary relationship among genetic sequences. Genetic conversions and recombinations, however, may not be modeled well under the branching assumptions of phylogenetic reconstruction. Progress is being made on inferring genetic relationships when recombination has occurred. Some examples of these methods include spectral analysis (Charleston, 1998), split decomposition (Dopazo *et al.*, 1993), and median networks (Guenoche, 1986). Another method of finding similarity among sequences is to perform ordination analyses, which attempt to describe the multidimensional information in low dimensions. Examples of these methods include Principal Components Analysis (PCA) (Mardia *et al.*, 1979), principal coordinates analysis (PCOORD) (Gower, 1966; Higgins, 1992), and multivariate statistical sequence analysis (van Heel, 1991). The algorithms used to reduce the dimensionality of the data set may, however, be non-intuitive to some users, and the results of the analysis may likewise be non-intuitive.

The program described here, Partition Analysis of Quasispecies (PAQ), is an intuitive tool that allows the user to explore genetic data and identify natural groupings of sequences that are most similar. The program's intuitive nature comes from the fact that it does not assume complex models of evolution like some of the methods mentioned previously. The number of clustering techniques

currently available is large, but the number of clustering programs for analyzing sequence data seems to be limited. Hierarchical methods of clustering are the most popular, partly because the output is a dendogram which illustrates the clusters. Of the methods that group by hierarchical techniques, the unweighted pair-group method using arithmetic averages (UPGMA) method (Lance and Williams, 1966) is one of the most widely used. Other hierarchical techniques include single linkage (SLINK) (Sneath, 1957), complete linkage (CLINK) (Johnson, 1967), and Ward's method (Ward, 1963). These hierarchical methods all belong to the category of agglomerative methods, since each element begins within a cluster and the number of clusters is reduced by merging one cluster to another cluster. The agglomerative methods merge data together to form clusters which grow larger in a process termed 'chaining'. This chaining process can identify mutually exclusive clusters but may have difficulty with overlapping clusters.

While PAQ shares common features with other clustering methods, its main differences are its non-hierarchical clustering algorithm and the fact that PAQ allows for overlapping clusters. Unlike the dendogram output of hierarchical methods, PAQ outputs partitions, or groups, of genetically similar sequences. The algorithm utilized by PAQ to identify the partitions is similar to the algorithm of K-means (MacQueen, 1965), which is the only notable non-hierarchical clustering method. The method of K-means requires that the user specify the desired number of partitions prior to cluster analysis. PAQ differs from K-means, however, by presenting the potential groups and allows the user to determine the most appropriate partitions. The main difference between PAQ and K-means is the fact that PAQ allows for overlapping groups while K-means forces sequences into mutually exclusive groups. While PAQ can be used to analyze many types of sequence data, it was designed specifically to analyze viral sequence data. Viruses exist as a heterogeneous population of closely related variants, commonly referred to as a quasispecies (Domingo *et al.*, 1996; Eigen, 1993; Holland *et al.*, 1992), and thus may not be grouped appropriately by mutually exclusive clusters. PAQ is designed to group sequences by using spherical clusters which are defined by a radius parameter. This program is not meant to replace phylogenetic analysis or the other methods mentioned above, but it offers an alternative exploratory tool for investigating groupings of genetically similar quasispecies sequences.

## SYSTEMS AND METHODS

PAQ was written in ANSI C++ and can be compiled and operated on Macintosh, PC, and UNIX platforms. The input file should consist of aligned nucleotide or amino acid variants. The most basic measure of genetic distance, nucleotide or amino acid change (referred to as genetic change hereafter), was used to define the distance between all sequences. After the input file is opened, the user has the option to consider or ignore gap positions in the distance calculations. A genetic distance matrix is created, where entries represent the number of genetic differences between all pairwise sequences summed across the entire sequence, also known as a Hamming distance. After all the pairwise sequence distances are calculated, the distances are sorted. A gap in the list of sorted distances may indicate the presence of distinct, non-overlapping groups. The program identifies the largest such region, which may serve as a good radius value to begin searching for groups. The remainder of the program is menu-driven and offers the following options:

1. Display the genetic distance matrix.
2. Display the groups for a range of radius values.
3. Display the groups for a single radius value.
4. Search a group for sub-groups.
5. Display a rearranged genetic distance matrix.
6. Display all potential groups for a single radius value.
7. Display the average distance from all centers in a range of radius values.
8. Display all genetic distance from a single sequence.
9. Display the genetic distance between two sequences.

## ALGORITHM

The basic assumption of the program is that sequences separated by the fewest genetic differences are more similar and should thus group together. A radius is selected by the user, and each sequence is used as a center to define spherical groups. From all the potential groups centered around each sequence, options 2–5 utilize an algorithm which automatically identifies all distinct groups (groups containing more than one variant and not a subset of other groups). The center genotype of each group is the sequence that is most representative of all variants within the group. The first distinct group is the group that contains the most variants within it. If more than one group contains the same number of variants, then the group that is the most 'compact' is chosen. Compactness is characterized by having more variants surrounding the center rather than near the boundaries of the group. Computationally, the most compact group minimizes the average distance between the center and all other variants within the group (neighbors), which is given by the

following equation:

$$\text{av.dist.} = (1/n)\sum_{i=1}^{n}(D_{ic})^2 \qquad (1)$$

where $n$ is the number of neighbors within the group with center variant $c$, and $D_{ic}$ is the genetic distance between variants $i$ and $c$. Subsequent distinct groups are found in the same manner as the first distinct group.

The second and third menu options utilize the automatic algorithm described above and outputs the distinct groups, the center genotype and its neighbors, the variants that belong to more than one group, and the variants that were not contained within any distinct groups. The fourth menu option searches for sub-groups within the distinct groups. After an initial radius is chosen and the distinct groups are displayed, the user can select a group and input a smaller radius to search for sub-groups. The rearranged genetic distance matrix, menu option 5, can be a useful tool to search for groups and sub-groups. Using a user-defined radius value, the genetic distance matrix is rearranged so that sequences partitioned together are grouped together in the new distance matrix. The result is a symmetric matrix with small distances around the matrix diagonal and larger distances elsewhere (Figure 1).

In the final analysis of sequence grouping, there may be multiple groups with different radius values. Menu options 6–9 offer the user additional information that aids in defining the most representative groups. For example, option 6 displays all the potential groups rather than the algorithm-derived final groups. Information concerning the average distance between the center genotype and all its neighbors can be obtained with menu option 7. The change in average distances may help to define appropriate radius values for different groups. Menu options 8 and 9 provide information concerning distances between specific sequences.

### *Pseudocode*

```
input filename (includes numseq, numsites)
choose to consider gaps in calculating sequence distances
calculate Hamming distance between all pairwise sequence comparisons
      (store in distmatrix)
sort the genetic distances
      output largest region of genetic distance not present in sorted list

menu:
(1) display the genetic distance matrix
      output distmatrix
(2) display the groups for a range of radius values
      input rmin and rmax
      for r = rmin, rmax (radius of group)
            for c = 1, numseq (center of group)
                  identify variants within r of c
      for all centers and all radius values, store the number of variants
            within r of c and the average
            distance of all variants within r of c (equation 1)
      for r = rmin, rmax (radius of group)
            for c = 1, numseq (center of group)
                  identify distinct groups as those with the most variants
                     within r of c (if multiple groups have the same number
                     of variants, then choose the group with the smallest
                     average distance of all variants within r of c); distinct
                     groups are not subsets of other groups
      for r = rmin, rmax (radius of group)
                  output the distinct groups, including the center and
                     all variants within r of the center, any variants
                     that belong to more than one group, and
                     all variants not contained within any groups
(3) display the groups for a single radius value
      same procedure as menu option (2), with rmin = rmax
(4) search a group for sub-groups
      input r (radius)
            use the procedure for menu option (3) to find distinct groups input
               clus#, the group to search for sub-groups create a new genetic
               distance matrix for the variants within clus# use the procedure
               for menu option (3) to find distinct sub-groups of clus# output
               the variants that were contained within clus# and the distinct
               sub-groups
(5) display a rearranged genetic distance matrix
      input r (radius)
            use the procedure for menu option (3) to find distinct groups
            rearrange the genetic distance matrix so that variants within
               the distinct groups are grouped together in the new genetic
               distance matrix
            output the new genetic distance matrix
(6) display all potential groups for a single radius value
      input r (radius)
      for c = 1, numseq
            identify and output all variants within r of c
(7) display the average distance for all centers in a range of radius values
      input rmin and rmax
      as in menu option (6), identify all potential groups for each radius value
      for r = rmin, rmax
            for c = 1, numseq
               output the average distance of c
(8) display all genetic distance from a single sequence
      input seq
      from the genetic distance matrix, output all genetic distance from seq
(9) display the genetic distance between two sequences
      input seq1
      input seq2
      from the genetic distance matrix, output the distance between seq1
         and seq2
```

## IMPLEMENTATION

PAQ has been tested with several data sets including human immunodeficiency virus type 1 (HIV-1) envelope sequences described by Shapshak *et al.* (1999). The HIV-1 sequences consisted of the V1–V5 domains of the surface envelope gene (approximately 1129 bases) isolated from different regions of the brain. Sixty-three sequences from this study and another sequence (HIVHXB3), used as the outgroup (accession numbers AF125810–AF125874, M14100), were analyzed by PAQ.

The nucleotide sequences were used to create the genetic distance matrix, ignoring gap positions (the analysis when gap positions were considered gave similar findings). After sorting the distance between all pairwise sequences comparisons, the program identified that no sequences differed by 69–107 nucleotide changes. Using

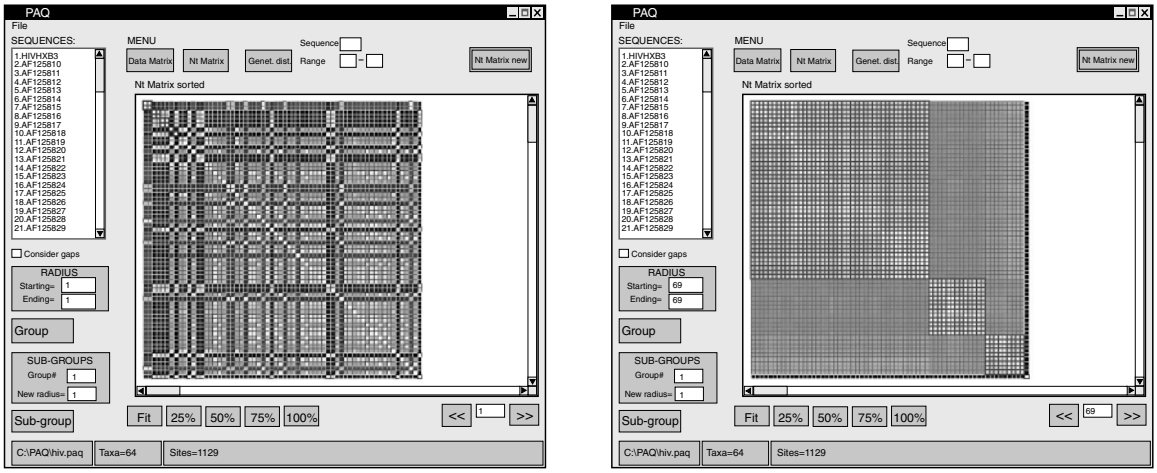


**Fig. 1.** Illustration of the windows-based PAQ program. The illustration on the left shows the original genetic distance matrix for the HIV-1 brain data, with the darker squares indicating large genetic distances. A radius value of 69 was used to group the sequences, and an illustration of the rearranged genetic distance matrix is shown on the right. In the rearranged distance matrix, the light squares indicate small genetic distances, and the gray squares indicate large genetic distances. The rearranged matrix shows three distinct, non-overlapping groups.

menu option 3 with a radius value of 69, three distinct, non-overlapping groups were identified and designated Groups A, B, and C. These groups contained all the sequences except the outgroup sequence. Group A contained 41 variants, while Groups B and C contained 13 and 9 variants, respectively. These results can be summarized by the rearranged genetic distance matrix, particularly using the graphical output of the windows-based PC version of the program (Figure 1). Groups A, B and C, identified by PAQ, corresponded to patients 144, 222, and 196, respectively, from which the envelope sequences were isolated.

Using menu option 4 with an initial radius of 69, we searched for possible sub-groups present in Groups A, B and C. In Group A, the program identified that no sequences differed by 26 or 27 nucleotide changes. Using a radius value of 26, four distinct, non-overlapping sub-groups were found that accounted for all but one variant within Group A. In Group B, a radius of 17 (identified by the program) found two distinct sub-groups, where two sequences belonged to both sub-groups. Two distinct, non-overlapping sub-groups were identified in Group C using a radius of 9 (identified by the program).

By simply using menu option 3 and the information about regions where no genetic distances were detected, three distinct, non-overlapping groups were found that accounted for all sequences except the outgroup sequence. Using menu option 4, eight distinct sub-groups were found within the three main groups. These eight sub-groups contained all but two variants, and two other sequences were shared by two of the sub-groups. This broad picture of the groups was further refined by utilizing menu options 5–9. For each group, smaller radius values were determined to maximize the compactness of the groups. The final result of the group analysis shows five sub-groups in Group A, three sub-groups in Group B, and two sub-groups in Group C (Figure 2). Each of the sub-groups corresponded to sequences that were isolated from the same region of the brain. Virus had been isolated from four different regions of the brain. In patient 144 (Group A), envelope sequences were amplified from all four regions. In the other patients, however, envelope sequences were recovered from only three regions in patient 222 (Group B) and from only two regions in patient 196 (Group C).

The genetic distance between the three large groups was approximated by the distance separating centers of their respective sub-groups. The genetic distance among Groups A, B, and C was greater than 100 nucleotide differences. The large genetic distances separating the three groups further supported the fact that the sequences in this data set had been isolated from three different patients. The five sub-group in Group A had radius values ranging from 7 to 20 nucleotide differences. In contrast, the range of radius values in the three sub-groups of Group B was 4–6, and both sub-groups of Group C had a radius value of 3. The genetic distance between sub-groups of the major groups was variable and was calculated as the distance between the center genotypes of each sub-group. The inter-group distance among sub-groups of Group A was generally around 50 nucleotide differences. In Group B, the inter-group

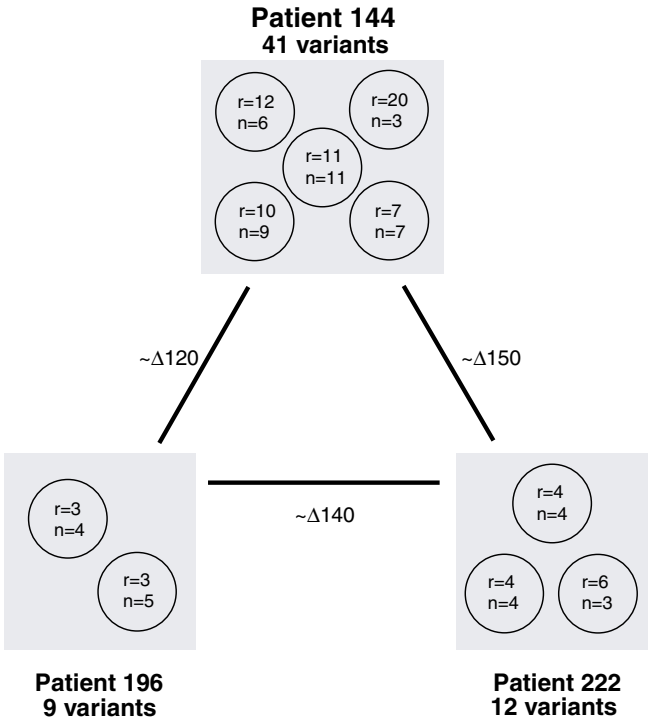


**Fig. 2.** Groups of HIV-1 *env* sequences isolated from the brain of infected patients. The ten groups, represented by the circles, segregate according to the patients from which the sequences were isolated. The genetic distance between groups in different patients (Δ) was much greater than the distance among groups within patients. For each patient, the number of variants isolated from the patient is indicated. The numbers within each circle indicates the radius of the group (*r*) and the number of sequences within the group (*n*).

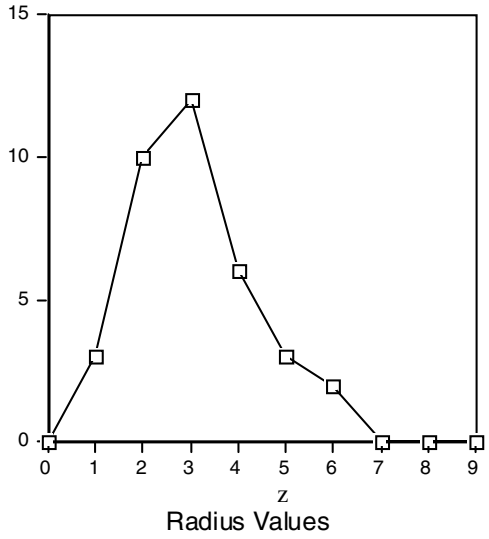


**Fig. 3.** Number of new neighbors in the group centered around variant #1. The range of radius values, 0–9 nucleotide differences, was used to define the group centered around variant #1. For each radius value, the number of new neighbors (variants not contained in the group defined by the previous radius value) is plotted.

distance of the sub-groups was around 25 nucleotide differences, and the two sub-groups of Group C were separated by 36 nucleotides. The genetic diversity within Group A was much greater than in the other two groups, as reflected by the large inter-group distances and radius values of the Group A sub-groups. This observation suggests that patient 144 (Group A) had been infected for a longer period relative to the other two patients. Shapshak also noted that the genetic divergence within brain regions was higher in patient 144 and came to the same conclusion about the relative age of infection in patient 144. Overall, the results of our PAQ analysis agreed well with Shapshak's phylogenetic analysis.

PAQ has also been used to study the viral quasispecies of equine infectious anemia virus (EIAV) within a serum sample commonly used to experimentally infect horses. EIAV is a lentivirus genetically (Kawakami *et al.*, 1987) and antigenically (Montelaro *et al.*, 1988) related to HIV. The data set was comprised of 37 sequences of *rev*, a regulatory gene absolutely required for viral replication (accession numbers AF314257-AF314402). Unlike the HIV data set, these sequences were more genetically similar, with only 9 nucleotide differences separating the most distant variants.

With no obvious radius values to begin the analysis, we used menu option 2 to examine the distinct groups defined by radius values from 1 to 9. The group centered around variant #1 was the largest distinct group for 6 of the 9 radius values. To determine which radius value to further examine, we plotted the number of new neighbors (variants not contained in the group defined by smaller radius values) within the group centered around variant #1 for the different radius values (Figure 3). A radius value of 3 nucleotide differences gave the maximum number of new neighbors, while larger radius values did not greatly include more variants. Thus, a radius of three nucleotide differences was used to examine the *rev* partitions.

Using a radius value of three nucleotide differences, variant #1 was the central genotype of the major group, which contained 25 neighbors. No other groups were mutually exclusive of the group centered around variant #1. To find a minor group, menu option 6 was used to display all the potential partitions defined by a radius of three nucleotide differences. To identify the minor group, all potential partitions were compared with respect to the number ($N$) and proportion ($P$) of variants within the partition that did not also belong to (overlap with) the major group. The partition centered around variant #28

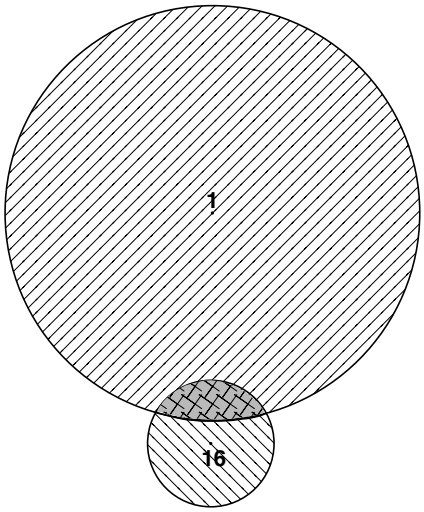


**Fig. 4.** The two groups of EIAV *rev* variants. The size of each group is proportional to the number of variants contained within it, and the numbers inside each group represents the central genotype. The larger group is centered around variant #1; the minor group is centered around variant #16, and the double-hatched region represents the overlap (sharing variants) between the two groups.

had the largest $N$-value, six, but the proportion was $P = 6/15$ (40%). There were 8 other partitions with $N = 5$, and the partition with the largest proportion, $P = 5/8$ (62.5%), was centered around variant #16. The only other partition with a higher proportion was centered around variant #24, $P = 2/3$ (66%), but the partition only had $N = 2$. Thus, the minor group was defined to be the group centered around variant #16. A third group was explored, but no groups had a value of $N$ greater than one. Therefore, the EIAV *rev* variants were characterized by a major group centered around variant #1, containing 26 variants, and a minor group centered around variant #16, containing 8 variants, with 3 variants shared by both groups (Figure 4).

## DISCUSSION AND CONCLUSION

PAQ represents a new tool for cluster analysis. PAQ differs from the majority of clustering methods because its non-hierarchical technique uses partitions rather than a dendogram to find groups of genetically similar variants. In comparison to the only notable non-hierarchical method, K-means, both methods strive to maximize the compactness of the distance within groups. The main difference between the two methods is that PAQ allows overlapping groups, while K-means forces variants into mutually exclusive groups. This difference was designed to handle viral quasispecies data, where variation between sequences may be large, as in the HIV example. In the EIAV example, however, the genetic variation in that data set was relatively small, and overlapping groups were used to describe the quasispecies.

The analysis offered by the PAQ program, as described in the previous section, is an intuitive means of grouping genetically similar variants. The genetic distance that separates sequences and groups also provides an understandable quantitative tool for researchers studying genetic data. By comparison, PAQ relies on fewer assumptions than phylogenetic analyses. Accordingly, PAQ has less power of inference than phylogenetic analyses. Implementation of PAQ on the HIV and EIAV data sets, however, exemplifies how the partition analyses can find groups of genetically similar sequences that can help researchers explore viral quasispecies to better understand the data. The fact that our partition analysis of the HIV data and the phylogenetic analysis gave similar results suggests that both analyses found some biologically significant results. Furthermore, analysis of the EIAV data by PAQ revealed a population structure that was not previously evident.

There are several layers of complexity involved with genetic data, and no single analytical tool can accurately describe the genetic and evolutionary relationships within the data. Multiple tools may be necessary to describe the genetic complexity. PAQ is an exploratory tool and is not meant to replace other genetic analysis tools, like phylogenetic reconstruction and multivariate statistical methods, but it should be used in conjunction with other tools to gain insight into how genetic sequences group together and how the variants are related. The main strengths of this program are its intuitive nature and its ability to explore several types of data, but specifically viral quasispecies data. These strengths may best be utilized with newly attained data that is not well-characterized and difficult to model. For example, PAQ has been used to analyze a longitudinal study that resulted from the experimental infection of a pony with the EIAV inoculum described above. Analysis of the partitioned groups present in sequential samples taken post-infection can examine their dynamic evolution over time (in preparation). Using PAQ to study how viral quasispecies exist and evolve may contribute greatly to our overall understanding of viruses.

## ACKNOWLEDGEMENTS

We greatly appreciate the discussions and critiques offered by Dean Adams, Gavin Naylor, and Xun Gu. This program was partially funded by the NSF grant DGE9972653 and by the USDA grant 96-358204-3847.

## REFERENCES

Charleston,M.A. (1998) Spectrum: spectral analysis of phylogenetic data. *Bioinformatics*, **14**, 98–99.

Domingo,E., Escarmis,C., Sevilla,N., Moya,A., Elena,S.F., Quer,J., Novella,I.S. and Holland,J.J. (1996) Basic concepts in RNA virus evolution. *FASEB J.*, **10**, 859–864.

Dopazo,J., Dress,A. and von Haeseler,A. (1993) Split decomposition: a technique to analyze viral evolution. *Proc. Natl. Acad. Sci. USA*, **90**, 10320–10324.

Eigen,M. (1993) Viral quasispecies. *Sci. Am.*, **269**, 42–49.

Felsenstein,J. (1993) PHYLIP: *Phylogeny Inference Package*, Version 3.572c, University of Washington, Seattle, WA.

Gower,J.C. (1966) Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, **53**, 325–333.

Guenoche,A. (1986) Graphical representation of a Boolean array. *Computers and the Humanities*, **20**, 277–281.

Higgins,D.G. (1992) Sequence ordinations: a multivariate analysis approach to analysing large sequence data sets. *Comput. Appl. Biosci.*, **8**, 15–22.

Holland,J.J., De La Torre,J.C. and Steinhauer,D.A. (1992) RNA virus populations as quasispecies. *Curr. Top. Microbiol. Immunol.*, **176**, 1–20.

Johnson,S.C. (1967) Hierarchical clustering schemes. *Psychometrika*, **32**, 241–254.

Kawakami,T., Sherman,L., Dahlbert,J., Gazit,A., Yaniv,A., Tronick,S.R. and Aaronson,S.A. (1987) Nucleotide sequence analysis of equine infectious anemia proviral DNA. *Virology*, **158**, 300–312.

Lance,G.N. and Williams,W.T. (1966) Computer programs for hierarchical polythetic classification. *Comput. J.*, **9**, 60–64.

MacQueen,J. (1965) Some methods for classification and analysis of multivariate observations. *Proc. 5th Berkeley Symp. Math. Stat. Probab.*, **1**, 281–297.

Mardia,K.V., Kent,J.T. and Bibby,J.M. (1979) *Multivariate Analysis*. Academic Press, New York.

Montelaro,R.C., Robey,W.G., West,M.D., Issel,C.J. and Fischinger,P.J. (1988) Characterization of the serological cross-reactivity between glycoproteins of the human immunodeficiency virus and equine infectious anemia virus. *J. Gen. Virol.*, **69**, 1711–1717.

Shapshak,P., Segal,D.M., Crandall,K.A., Fujimura,R.K., Zhang,B., Xin,K., Okuda,K., Petito,C.K., Eisdorfer,C. and Goodkin,K. (1999) Independent evolution of HIV type 1 in different brain regions. *AIDS Res. Human Retroviruses*, **15**, 811–820.

Sneath,P.H.A. (1957) The application of computers to taxonomy. *J. Gen. Microbiol.*, **17**, 201–226.

Swofford,D.L. (1999) PAUP*. *Phylogenetic Analysis Using Parsimony (* and other methods)* Version 4, Sinauer Associates, Sunderland, Massachusetts.

van Heel,M. (1991) A new family of powerful multivariate statistical sequence analysis techniques. *J. Mol. Biol.*, **220**, 877–887.

Ward,J.H. (1963) Hierarchical grouping to optimize an objective function. *J. Am. Statist. Ass.*, **58**, 236–244.