BIOINFORMATICS ORIGINAL PAPER Vol. 24 no. 3 2008, pages 383–388
doi:10.1093/bioinformatics/btm621

Genetics and population analysis

A new algorithm for cluster analysis of genomic methylation: the *Helicobacter pylori* case

F. F. Vale[1,*], P. Encarnação[1] and J. M. B. Vítor[2]

[1]Engineering Faculty, Portuguese Catholic University, Estrada Octávio Pato, 2635-631 Rio de Mouro, Portugal and [2]CECF (iMed.UL), Faculty of Pharmacy, University of Lisbon, Av. Forças Armadas, 1649-003 Lisboa, Portugal

Received on October 19, 2007; revised and accepted on December 13, 2007

Advance Access publication December 16, 2007

Associate Editor: Martin Bishop

ABSTRACT

Motivation: The genomic methylation analysis is useful to type bacteria that have a high number of expressed type II methyltransferases. Methyltransferases are usually committed to Restriction and Modification (R-M) systems, in which the restriction endonuclease imposes high pressure on the expression of the cognate methyltransferase that hinder R-M system loss. Conventional cluster methods do not reflect this tendency. An algorithm was developed for dendrogram construction reflecting the propensity for conservation of R-M Type II systems.

Results: The new algorithm was applied to 52 *Helicobacter pylori* strains from different geographical regions and compared with conventional clustering methods. The algorithm works by first grouping strains that share a common minimum set of R-M systems and gradually adds strains according to the number of the R-M systems acquired. Dendrograms revealed a cluster of African strains, which suggest that R-M systems are present in *H.pylori* genome since its human host migrates from Africa.

Availability: The software files are available at http://www.ff.ul.pt/paginas/jvitor/Bioinformatics/MCRM_algorithm.zip

Contact: filipavale@fe.ucp.pt

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Typing methods are useful to understand the natural history, epidemiology, mode of transmission, reservoir and clinical implications of bacterial infection (Owen *et al*., 2001). The genomic methylation typing method is useful to type bacteria that have a high number of expressed Type II methyltransferases (MTase) (Vale and Vitor, 2007). Most of the bacterial MTases are members of Restriction and Modification (R-M) systems. A Type II R-M system is defined by the association of at least two genes: one codes for a restriction endonuclease (REase) that recognizes a specific DNA sequence and cuts both strands; the other gene codes for a cognate MTase that methylates the same DNA sequence, thus protecting it from being cleaved by the companion REase (Roberts *et al*., 2003). The physical separation of the restriction and modification activities on different proteins contrasts with the tight linkage of their genes. The linkage of their genes allows for simultaneous loss of R and M genes, while physical separation of their gene products allows for hydrolysis of the genomic DNA by residual REase present in daughter cells, and leads to postsegregational killing. This happens because the cell loses the ability to protectively methylate all recognition sites in the newly synthesized chromosome, while unmethylated sites are cut by the residual REase still present in the bacterial cytoplasm. Considering this, Type II R-M systems have been classified as selfish genetic elements that in some instances are not easily lost from their host cell. Due to the pressure of the REase on the expression of the cognate MTase, Type II R-M systems tend to be maintained after acquisition (Kusano *et al*., 1995; Naito *et al*., 1995; Nakayama and Kobayashi, 1998). According to the selfish gene theory, the only way to escape cell death is the prior inactivation, or elimination, of the R gene, followed by inactivation or deletion of M gene, a few generations later (Naito *et al*., 1995).

To perform genomic methylation typing, the expression of methyltransferases is determined by digesting the genomic DNA with the cognate REase (Vale and Vitor, 2007). However, cluster analysis by conventional methods does not consider the propensity for R-M systems conservation after acquisition.

The genomic methylation typing method may be applied to the small group of species with a large number of R-M systems, like *Campylobacter upsaliensis*, *Neisseria gonorrhoeae* FA 1090, and *Helicobacter pylori* with 30, 17 and 26 putative methyltransferases genes, respectively (Roberts *et al*., 2007).

In this work we developed a new clustering algorithm that takes into account the pressure of REases on MTases, and that is based on the hypothesis that each strain evolves by acquiring new R-M systems without loosing acquired RM systems. Type II R-M systems capacity to act as selfish genetic elements may contribute to dissemination and conservation in host genome, and its different GC content is compatible with horizontal gene transfer (Kusano *et al*., 1995; Naito *et al*., 1995).

We use *H.pylori* to benchmark the clustering algorithm with real data from genomic methylation typing. *H.pylori* is mainly transmitted intrafamilial from person to person

*To whom correspondence should be addressed.

(Perez-Perez *et al*., 2004; Raymond *et al*., 2004), and most likely has coevolved with its human host (Covacci *et al*., 1999; Linz *et al*., 2007). Investigation on *H.pylori* is of particular importance since the bacteria colonize the mucosa of human stomach, causing several diseases such as chronic gastritis, peptic ulcer or gastric cancer (Dunn *et al*., 1997; Kusters *et al*., 2006).

2 SYSTEMS AND METHODS

2.1 *H.pylori* strains

We worked with 52 *H.pylori* strains (Table S1), 50 random selected from different geographic regions, and two of the complete sequenced strains, 26695 (Tomb *et al*., 1997) and J99 (Alm *et al*., 1999). *H.pylori* culture and DNA extraction was performed as described elsewhere (Megraud, 1996; Vale and Vitor, 2007). Every chromosomal DNA preparation was incubated with the four New England Biolabs buffers, 1 h at 37°C, to verify if the DNA was nuclease free.

2.2 Genomic methylation typing

To identify the expression of the cognate methyltransferase, the *H.pylori* genomic DNA was digested with 27 REases (AciI, AseI, BseRI, BssHII, BstUI, DdeI, DpnI, DpnII, DraI, EagI, FauI, Fnu4HI, FokI, HaeIII, HhaI, Hpy188I, Hpy188III, Hpy99I, HpyCH4III, HpyCH4IV, HpyCH4V, MspI, NaeI, NlaIII, Sau96I, ScrFI, and TaqI). Digestions were performed according to the manufacturer's instructions (New England Biolabs, USA), in a reaction volume of 50 µl. An excess of REase units was always used to assure complete cleavage of the DNA. After electrophoresis through 0.7% agarose gel (Agarose LE Seakem, FMC) the results were coded on a binary matrix, where '0' indicates digestion observed (DNA is unmethylated), and '1' indicates no digestion, suggesting an active methyltransferase (Vale and Vitor, 2007).

2.3 Minimum common restriction modification (MCRM) algorithm development

The algorithm aims at constructing a dendrogram that reflects the selfish behavior of R-M systems, i.e. the loss of Type II R-M gene complexes inhibits the propagation of a cell population and causes chromosome breakage (Handa and Kobayashi, 1999).

The new algorithm works as follows. Consider a set of strains with different methylation status. The algorithm starts by choosing one of the strains in the set that contains less R-M systems, and if two or more strains have exactly the same blueprint, they are grouped and treated as one. For future reference, let it be strain A. In case of several strains sharing the same minimum number of R-M systems, the first of them in the set is chosen. The strain with less R-M systems is hypothesized to be the one that has the core set of the most abundant R-M systems expressed among the typed strains. We consider the hypothesis that these core set of R-M systems was the first to be acquired by *H.pylori*, so that they exhibit a large dissemination (expression) among several daughter strains. Then, the R-M system A_i in strain A that is shared by the largest number of strains is selected (again, in case of tie, the first feature is selected). The selected R-M system A_i is hypothetically considered as the first to be acquired by an ancestral strain and is also the one that is most spread among strains. The set of strains is now divided into two groups: the group X of the strains having the A_i feature in common with strain A, and the group Y of those who have not. If the strain A contained only the R-M system A_i, it is removed from the X group since it cannot be grouped with any other strain. The strain A position in the dendrogram is settled at a similarity level of $1/nM$, where

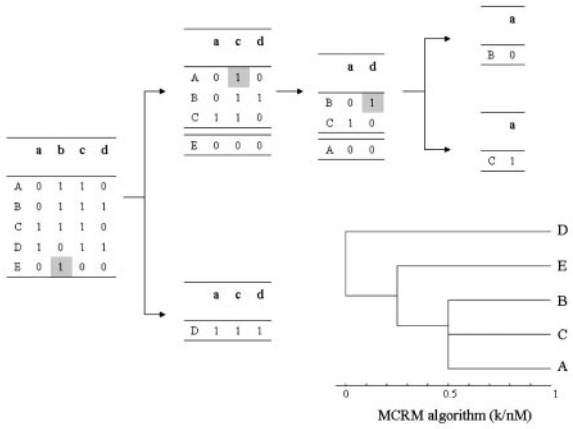

Fig. 1. MCRM algorithm application example. Strains are represented in capital letters and R-M systems in small caps. The selected R-M system in the selected strain is highlighted at each algorithm step. Strain E is first selected because it has the smallest number of R-M systems (only R-M system b). Position of strain D is found in the dendrogram at 0% (it does not share any gene with strain E), and strain E at 25%. Strain E and feature b are removed. The last three strains diverge at 50%.

nM is the total expressed R-M systems among all strains. If more than one R-M system is common to all tested strains (let us consider this k), then the similarity level would be k/nM. In both groups X and Y, the feature A_i is removed since it was already been taken into account in the strain grouping. Therefore, the X and Y groups now have strains with one feature less than the strains in the original set. The algorithm then runs recursively taking as inputs groups X and Y, one at a time, choosing the strain with less R-M systems in each group, then the R-M system that is shared by the largest amount of strains in the group, and creating new X and Y groups. It ends when the input list has one or none strains. This way, the recursive algorithm produces a tree (dendrogram) where the branches bifurcate each time the algorithm is run over a set of strains, one of the bifurcated branch containing the subset of the strains that share the chosen feature, and the other containing the subset of strains that do not have that feature. The tree leaves are the individual strains (or group of strains in case there were strains with exactly the same blueprint in the original set). Figure 1 shows an application example of MCRM algorithm.

The displayed dendrogram depends on the particular choice of the A strain and/or the A_i R-M system in A when a tie occurs. Different choices may lead to different dendrograms, which is also common in conventional algorithms (Sneath and Sokal, 1973). The algorithm implemented can produce different dendrograms if prompted by the user. It does so by rearranging the strains and the features in a randomized way. Since the candidate strain (or candidate feature) selected by the algorithm is the first one, different orders may lead to different dendrograms. However, this simple strategy of strain and feature rearrangements does not warrant that the new dendrogram is different from the precedent. In practice, several dendrograms will be similar, but it is possible that distinct dendrograms are generated since different choices of strain or R-M system at ties may result in different clustering. Thus, the user should generate a high number of different dendrograms using the MCRM algorithm in order to gain confidence in the clustering results.

2.4 Algorithm implementation

The Minimum Common Restriction Modification algorithm was implemented using Matlab® R12 taking advantage of the built in functions

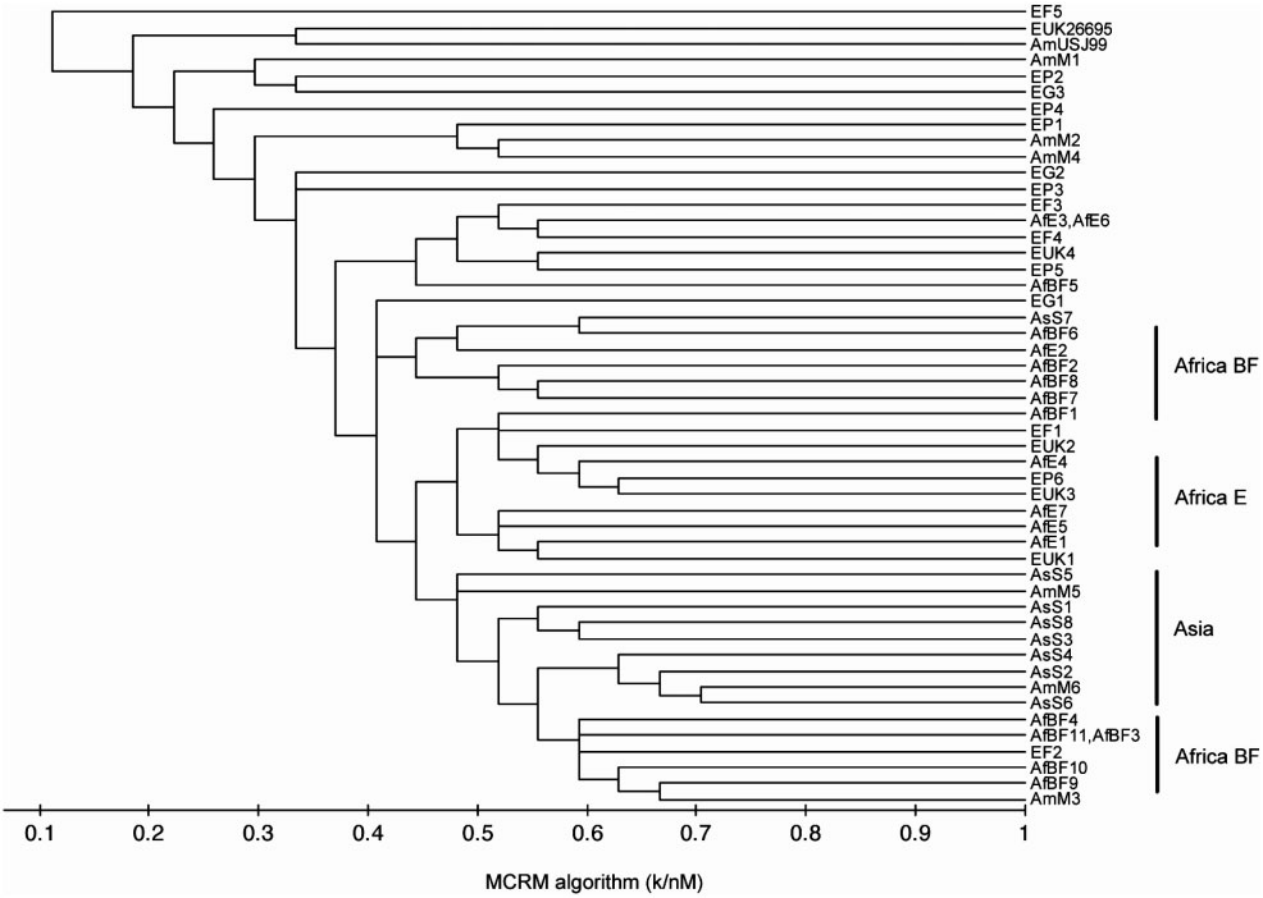

Fig. 2. Dendrogram produced by MCRM algorithm (k/nM), where K is the number of R-M systems common to all tested strains and nM the total number of R-M systems screened. The vertical bars point out the Asian (Singapore) and African (BF, Burkina Faso; E, Egypt) clusters.

for array manipulation. The set of strains is imported from a MSExcel® worksheet and stored in an array where each line contains a strain and each column corresponds to a different feature. The first line should list the R-M system's designations and the first column the strains names. Figure S1 displays the flowchart of the algorithm implemented.

2.5 Comparison with other cluster algorithms

In order to compare the newly suggested algorithm we analyze the same data by conventional methods previously used for genomic methylation typing (Vale and Vitor, 2007). A dendrogram was constructed using the unweighted pair-group method for arithmetic averages (UPGMA) method and the Jaccard coefficient, using the software NTSYS v.2.0 (Exeter Software). The Jaccard coefficient reflects the percentage of common methyltransferases between two strains, and the UPGMA method assumes that each branch has the same evolution rate of MTases expression.

3 IMPLEMENTATION

3.1 MCRM algorithm applied to *H.pylori* typing

The genomic methylation typing data of the 52 *H.pylori* strains were analyzed with the proposed MCRM algorithm. The resulting dendrogram is shown in Figure 2. The option to display more than one dendrogram was used and several dendrograms were obtained (not shown). The similarity level represents number of the shared expressed methytransferases at each node. The similarity level of all tested strains is 11.1% (3/27). This value is the rate of k/nM, where k is the number of common methyltransferases expressed by all strains and nM is the total number of methyltransferases screened for. In this application we have three common methyltransferases transversal expressed and a total of 27 MTases screened. All common MTases found were previously described as possibly conserved in *H.pylori*, which are M.NlaIII, also described as *iceA* or hpyIM (Xu *et al.*, 1997), M.HhaI and M.NaeI (Vale and Vitor, 2007).

To evaluate the dendrogram we determined the discriminatory method capacity, which is the strain frequency per cluster, and should be <5% ($nj/N < 0.05$); and the Simpson's index of diversity that reflects the capacity of the method to distinguish unrelated strains (Hunter and Gaston, 1988; Maslow *et al.*, 1993; Priest and Austin, 1993).

The Simpson's index of diversity was 99.8%, meaning that the discriminatory power has the average probability, that the typing method will assign as different types, two unrelated strains randomly sampled in the microbial population; the

frequency of each group for a similarity level of 100% was <5% (nj/N = 0.04%) as suggested by Maslow *et al*. (Maslow *et al*., 1993). The typeability, which is the proportion of strains that are assigned a type by the typing system, revealed that the 50 strains were considered different. The strains with exactly the same methylation profile are AfBF3 and AfBF11 and AfE3 and AfE6. The strains of each pair were isolated from the same country, Burkina Faso and Egypt, respectively.

Our data demonstrate that beside the common MTase to all tested strains, we have two others that are present in all but one strain, which are M.BssHII and M.BseRI (Table S2). In this particular case both MTases are absent in the same strain, EF5. The algorithm randomly chooses one of these MTases. Let us consider that the analyzed MTase is M.BssHII. The MCRM algorithm generates two groups: the X group that expresses M.BssHII (all strains but EF5) and the Y group that do not expresses M.BssHII (strain EF5). The position of the strain EF5 determined at a similarity level of 11.1% with the rest of the strains. The strain EF5 is then removed from the group as well as the feature M.BssHII. All the remaining strains share a total of 5 expressed MTases (M.NlaIII, M.HhaI, M.NaeI, M.BssHII and M.BseRI). All strains but EF5 diverge now at 18.5% (k/nM = 5/27) similarity level (Fig. S2). Then, the MCRM algorithm recursively runs trough all data eventually producing one, or several, dendrogram(s).

The visual analysis of the dendrogram suggests that there are several clusters and sub-clusters, clearly associated with different geographic regions (discussed below). In a total of 100 produced dendrograms with MCRM algorithm we have observed clusters of African strains in all dendrograms. The relative frequency of an Asian cluster (Singapore strains) in these 100 dendrograms was 53%. We have considered as an Asian cluster at least half of the tested Singapore strains present in the same cluster, and the Singapore strains as the majority strains present in that cluster. No other cluster (from Europe or America), besides the cluster of African strains that is common to all dendrograms, appear in these set of 100 dendrograms. We consider this value (53%) as a measure of repeatability of the dendrograms obtained, and not as a measure of the true *H.pylori* geographic distribution.

Dendrograms are useful to determine the population structure of a bacterial population. *H.pylori* population structure has been classified either as panmictic (Salaun *et al*., 1998) or as family clonal (Drumm *et al*., 1990; Suerbaum and Achtman, 2004). Due to the frequent genome rearrangement among *H.pylori* strains (Alm *et al*., 1999), the classification of the population structure is highly influenced by the typing method selected. As the number of analyzed strains increases the distinction of clear clusters, bacterial population structure may not be so obvious to determine. The dendrogram obtained with the newly developed algorithm appears to show clusters, but the shape of the dendrogram for the more distant strains is typical of a panmictic population.

3.2 Comparison with conventional algorithms

The data were also analyzed with a conventional method previously used for genomic methylation typing (Vale and Vitor, 2007), i.e. UPGMA method and Jaccard Coefficient (Fig. S2). Both dendrograms produced either by MCRM algorithm or by a conventional method present some similarities. Obviously the strains with the same genomic methylation status are the same; The Simpson's index of diversity for a similarity level of 100% was the same (99.8%); the frequency of each group for a similarity level of 100% was also the same (3.8%). The cophenetic correlation coefficient, which represents the goodness of fit of the dendrogram to the initial data, revealed a very good fit to the data matrix ($r = 0.86$). A cophenetic correlation of $r = 1$ would be ideal, but dendrograms are bidimensional representations that introduce distortion to the multidimensional existence structure (Priest and Austin, 1993). The lines connecting strains in the hierarchical dendrogram reveal that the most distant strain is once again EF5, the same observed with MCRM algorithm. The observation of the dendrogram (Fig. S2) suggests that clusters and several sub-clusters are present. Once more it seams that strains are grouped by geographic areas. Nevertheless, the dendrogram shape is not of a clearly clonal population (with defined clusters), but there is evidence of clonal proliferation for at least some of the strains (observe in Fig. 1 strains from Burkina Faso, Egypt, and Singapore). Even though the similarity found between both dendrograms, we must recall that the UPGMA method assumes that the same rate of evolution applies to each branch, and that the Jaccard coefficient is an association coefficient representing the proportion of variables that match, excluding those that both strains lack. For genomic methylation typing the UPGMA method assumes that each branch exhibit the same rate of evolution of methyltransferase expression, whether this change is gain or lost; while the Jaccard coefficient reflects the percentage of common methylases between two strains. Globally this method does not reflect the selfish gene concept for R-M systems.

4 DISCUSSION

The dendrogram produced with MCRM algorithm (Fig. 2) confirms the good discrimination power of the genomic methylation typing method (Vale and Vitor, 2007), and respects the concept of postsegregational killing if a Type II R-M system is lost. In fact, according to the selfish gene concept, the descendants of cells that lose a restriction-modification gene complex are unable to modify a sufficient number of recognition sites in their genomes to protect them from lethal attack by the remaining molecules of restriction enzyme (Kobayashi, 1998). The conservation of R-M systems after acquisition is not reflected by any of the conventional cluster methods. We have developed a novel algorithm for cluster analysis that is based on the concept of preservation of R-M systems. The MCRM algorithm proposed solves the typing data and produces a dendrogram with excellent Simpson's index of diversity and discriminatory method capacity. In fact, in the *H.pylori* application example presented in this work these values are exactly the same as those obtained by the conventional method tested for comparison, which indicates the hierarchical cluster capacity of MCRM algorithm.

Conventional clustering methods produce an intermediate similarly or distance matrix that is used for dendrogram construction. The Mantel test evaluates the null hypothesis of

no relationship between two dissimilarity (distance) or similarity matrices. The first matrix is the one used to produce the dendrogram, and the second is a matrix produced from the dendrogram. These two matrices normally are not exactly the same, as the dendrogram faces the problem of ties, which resolution generates several dendrograms for one data set. The result of Mantel test is reflected by the cophenetic correlation coefficient, which represents the goodness of fit of the dendrogram to the initial data (Priest and Austin, 1993; Sneath and Sokal, 1973). As the dendrogram produced by MCRM algorithm is not obtained after generation of an intermediate similarity matrix, we are not able to evaluate the reliability of the dendrogram preservation to original data through the cophenetic correlation coefficient calculation.

H.pylori infection is present in ~50% of the human population (Kusters *et al*., 2006). This large percentage of infection has permitted to compare the human genetic diversity described by Cavalli-Sforza (Cavalli-Sforza, 2001) with the bacteria genetic diversity. The pattern of geographic distribution of *H.pylori* and man is surprisingly similar, which allowed speculating for simultaneous coevolution of human and *H.pylori* (Covacci *et al*., 1999). Recently a simulation predicted that *H.pylori* has spread from East Africa over the same time scale (58 000 years ago) as anatomically modern humans (Linz *et al*., 2007). Both dendrograms here presented, obtained by the new developed algorithm and by conventional methods reveal that clusters and subclusters get together by geographic source of analyzed strains. Strains with African origin form two particular clusters (Burkina Faso, strains coded AfBF; and Egypt, strains coded AfE) and from these core clusters all others appears to diverge. Therefore our data are in agreement with the old association between bacteria and man before the out of Africa modern human migration previously described (Covacci *et al*., 1999; Linz *et al*., 2007). The other major cluster is formed by strains whose source is Asia (coded AsS), more precisely Singapore, which emerge together with minor exceptions. It is worth to mention that in some of the dendrograms generated by the MCRM, corresponding to different tie solutions, Singapore strains are not clustered together. Nevertheless, in the majority of the produced dendrograms (53%), these strains are grouped together. European and American originated strains come out mixed in all the 100 analyzed dendrograms, which is also in agreement with the human recent history of America colonization by Europeans. Notably in the other 47%, we found half of the Asian strains in a cluster of 18 strains whose main origin is Africa (9 strains), which is also in agreement with a migration event out of Africa to Asia (data not shown).

In Figure 2 the clustering of some strains appears to contradict this association of population distribution between human and bacteria. Indeed, some strains (like AfE4, AfBF1 or AsS7, Fig. 2) are outside the clusters of strains originated from Africa and Asia (Singapore). It is important to recall that we just know the geographic origin of each strain, but we ignore the history of each particular human host. It is thus possible that the human host has been influenced by recent migration events. Additionally, we also cannot rule out the hypothesis of *H.pylori* recent horizontal transmission from a human source with a diverse genetic background. Moreover, in established human population, *H.pylori* presents elevated horizontal gene transfer, as this bacteria is natural competent (Hofreuter *et al*., 2000), either from mixed strains stomach infection, or from free bacteria DNA presents derived from food (Torres *et al*., 2000), which may originate the genetic shift of these *H.pylori* strains and consequently the discrepancies in dendrogram strain position. Indeed, restriction and modification genes may be acquired by horizontal gene transfer (Alm *et al*., 1999; Lin *et al*., 2001). This contradiction may simply be a matter of chance in strain selection.

The presence of clusters of strains with geographic source in Africa and Asia (Singapore) is in agreement with the fact that the modern humans appeared first in Africa, then in Asia, and from this continent settled its three appendices: Oceania, Europe, and America (Cavalli-Sforza, 2001). These strain specific majority genes (restriction and modification genes) appear to be present in *H.pylori* genome since the out of Africa human diaspora, since the genomic methylation typing seems to reproduce the traces of human migrations.

The advantage of using MCRM algorithm arrives from the fact that it is based on the hypothesis of conservation of R-M systems (Kusano *et al*., 1995; Naito *et al*., 1995). Indeed, a follow up study of 10 patients during a period of 6 months to 4 years revealed substantial conservation (94,5%) in the MTase expression. The minor differences in expression were mainly acquisition (Vale and Vitor, 2007), which is in agreement with the selfish gene concept (Kusano *et al*., 1995; Naito *et al*., 1995). Taken together we suggest that the approach of the MCRM algorithm displays a dendrogram that reflects with more accuracy the *H.pylori* strain distribution. Future work will include a systematic comparison between the results here described and the ones reported in other studies like (Linz *et al*., 2007), and also between MCRM algorithm and other clustering methods (for example MLST). Another goal is to develop an automatic method to compare dendrograms in order to show only the dendrograms that are generated from distinct tie solutions that are indeed different from each other.

5 CONCLUSION

In summary, we observe that strain typing by the genomic methylation method analyzed either by conventional cluster methods, or by the new proposed methodology is consistent with the hypothesis of coevolution of *H.pylori* and its human host. The new algorithm possibly reflects with more accuracy the history of migrations and current geographic distribution of *H.pylori*, since it takes in consideration the conservation of R-M systems after acquisition. Our data suggest that if the clusters associated with geographic distribution of *H.pylori* are compatible with those of humans, then R-M systems may be present in *H.pylori* chromosome before the modern humans left Africa. On the other hand, this also supports the conservation of R-M systems after acquisition; otherwise this method would not predict this coevolution.

ACKNOWLEDGEMENTS

We thank Francis Mégraud, Lurdes Monteiro, and Sebastian Suerbaum for the *H.pylori* strains; Nuno Taveira for critical

reviewing of the manuscript; and three anonymous reviewers for their valuable comments and suggestions. This work was partially funded by New England Biolabs, Inc. (USA).

Conflict of Interest: none declared.

REFERENCES

Alm,R.A. *et al*. (1999) Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. *Nature*, **397**, 176–180.

Cavalli-Sforza,L.L. (2001) *Genes, Peoples and Languages*. Penguin Books, London.

Covacci,A. *et al*. (1999) *Helicobacter pylori* virulence and genetic geography. *Science*, **284**, 1328–1333.

Drumm,B. *et al*. (1990) Intrafamilial clustering of *Helicobacter pylori* infection. *N. Engl. J. Med.*, **322**, 359–363.

Dunn,B.E. *et al*. (1997) *Helicobacter pylori*. *Clin. Microbiol. Rev.*, **10**, 720–741.

Handa,N. and Kobayashi,I. (1999) Post-segregational killing by restriction modification gene complexes: observations of individual cell deaths. *Biochimie*, **81**, 931–938.

Hofreuter,D. *et al*. (2000) Genetic competence in *Helicobacter pylori*: mechanisms and biological implications. *Res. Microbiol.*, **151**, 487–491.

Hunter,P.R. and Gaston,M.A. (1988) Numerical index of the discriminatory ability of typing systems: an application of Simpson's index of diversity. *J. Clin. Microbiol.*, **26**, 2465–2466.

Kobayashi,I. (1998) Selfishness and death: raison d'etre of restriction, recombination and mitochondria. *Trends Genet.*, **14**, 368–374.

Kusano,K. *et al*. (1995) Restriction-modification systems as genomic parasites in competition for specific sequences. *Proc. Natl Acad. Sci. USA*, **92**, 11095–11099.

Kusters,J.G. *et al*. (2006) Pathogenesis of *Helicobacter pylori* infection. *Clin. Microbiol. Rev.*, **19**, 449–490.

Lin,L.F. *et al*. (2001) Comparative genomics of the restriction-modification systems in *Helicobacter pylori*. *Proc. Natl Acad. Sci. USA*, **98**, 2740–2745.

Linz,B. *et al*. (2007) An African origin for the intimate association between humans and *Helicobacter pylori*. *Nature*, **445**, 915–918.

Maslow,J.N. *et al*. (1993) Molecular epidemiology: application of contemporary techniques to the typing of microorganisms. *Clin. Infect. Dis.*, **17**, 153–162.

Megraud,F. (1996) Diagnostic bactériologique standart de l'infection à *Helicobacter pylori*. In Megraud,F. and Lamouliatte,H. (eds.) *Helicobacter pylori*. Elsevier, Amsterdam, pp. 249–266.

Naito,T. *et al*. (1995) Selfish behavior of restriction-modification systems. *Science*, **267**, 897–899.

Nakayama,Y. and Kobayashi,I. (1998) Restriction-modification gene complexes as selfish gene entities: roles of a regulatory system in their establishment, maintenance, and apoptotic mutual exclusion. *Proc. Natl Acad. Sci. USA*, **95**, 6442–6447.

Owen,R.J. *et al*. (2001) Heterogeneity and subtyping. In Mobley,H.L., Mendz,G.L. and Hazell,S.L. (eds.) *Helicobacter pylori physiology and genetics*. ASM, Washington, DC, pp. 363–378.

Perez-Perez,G.I. *et al*. (2004) Epidemiology of *Helicobacter pylori* infection. *Helicobacter*, **9** (Suppl. 1), 1–6.

Priest,F. and Austin,B. (1993) *Modern Bacterial Taxonomy*. Chapman & Hall, London.

Raymond,J. *et al*. (2004) Genetic and transmission analysis of *Helicobacter pylori* strains within a family. *Emerg. Infect. Dis.*, **10**, 1816–1821.

Roberts,R.J. *et al*. (2003) A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. *Nucleic Acids Res.*, **31**, 1805–1812.

Roberts,R.J. *et al*. (2007) REBASE–enzymes and genes for DNA restriction and modification. *Nucleic Acids Res.*, **35**, D269–D270.

Salaun,L. *et al*. (1998) Panmictic structure of *Helicobacter pylori* demonstrated by the comparative study of six genetic markers. *FEMS Microbiol. Lett.*, **161**, 231–239.

Sneath,P.H.A. and Sokal,R.R. (1973) *Numerical taxonomy: the Principles and Practice of Numerical Classification*. Freeman, San Francisco.

Suerbaum,S. and Achtman,M. (2004) *Helicobacter pylori*: recombination, population structure and human migrations. *Int. J. Med. Microbiol.*, **294**, 133–139.

Tomb,J.F. *et al*. (1997) The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature*, **388**, 539–547.

Torres,J. *et al*. (2000) A comprehensive review of the natural history of *Helicobacter pylori* infection in children. *Arch. Med. Res.*, **31**, 431–469.

Vale,F.F. and Vitor,J.M. (2007) Genomic methylation: a tool for typing *Helicobacter pylori* isolates. *Appl. Environ. Microbiol.*, **73**, 4243–4249.

Xu,Q. *et al*. (1997) The *Helicobacter pylori* genome is modified at CATG by the product of hpyIM. *J. Bacteriol.*, **179**, 6807–6815.