Aggregating Homologous Protein Families in Evolutionary Reconstructions of Herpesviruses *

Boris Mirkin[†], Renata Camargo, Trevor Fenner, George Loizou[‡] and Paul Kellam[§]

Abstract–Protein families can be used to reconstruct evolutionary histories of organisms. The accuracy of protein assignment to such families is critical for the success of such studies. Here we investigate the automatic aggregation of motif-defined homologous protein families for further reconstruction of their evolutionary histories. We propose a method that utilises only parameters that can be adjusted by using the data. The building blocks of the method include: (a) a majority rule for combining protein homologous neighbourhood lists into that for a family, and (b) a robust clustering procedure whose only parameter, the similarity shift, can be estimated from information on proteins with known function. The method is applied to a herpesvirus protein dataset leading to insights into the composition of ancestors of herpesvirus superfamilies. Comparison of the computational reconstructions with more comprehensive analyses also show how alignment-based between-protein similarity scoring can be improved by using data on gene arrangements.

1 Introduction

Reconstructing evolutionary histories of organisms using information on the total protein coding content of each genome has provided new insights into genome evolution [10, 14, 17]. Gene gain and loss events have also been applied to reconstructing virus evolutionary history [1]. Herpesviruses infect a wide variety of animal species and eight herpesviruses are known to infect humans. Herpesviruses cause important diseases in both animals and man. Despite their pathogenesis, herpesviruses are highly adapted to sustain a lifelong infection of their host after the primary infection. This adaptation to a mutual co-existence has involved the evolution of the herpesvirus genomes with the result that the species in existence today capture an extensive evolutionary history, within which important insights into host and pathogen interactions are buried. Many herpesviruses have had their complete genomes sequenced, and this shows that herpesviruses maintain some common genes but also undergo extensive gene gain and loss. The availability of high-level genomic data creates an opportunity to accurately map the history of individual genes onto a phylogeny. Such a mapping presents a unique way of visualising the evolution of the pathogen's functions. In addition, different herpesvirus genomes have different gene contents, suggesting active gain and loss of genes. The problem of retracing individual gene histories over the evolution of herpesvirus as a family requires a guide phylogenetic tree. The evolution of herpesvirus family has been mapped to such phylogenetic trees rather robustly using single genes or subsets of conserved herpesvirus genes, all producing trees with very similar topologies. Such a tree can be used as the target for parsimoniously mapping "phyletic" patterns of genes onto it. The pattern of gene presence/absence for the extant species can be plausibly extended over the entire tree in such a way that an evolutionary scenario is built by annotating the phylogenetic tree with events corresponding to gene birth, horizontal transfer and loss.

Currently the basic units of the mapping to the trees are homologous protein families representing aggregations of related individual genes. Assignment to protein families is often determined with a large manual component because the degree of similarity between proteins within an alignment of protein sequences is not always sufficient to automatically identify the families. Significant protein similarity over the full length of the protein is often insufficient to group proteins into families, especially for rapidly evolving organisms such as bacteria and viruses. This makes the application of phyletic gene/protein gain and loss mapping difficult to perform routinely. Computational methods for identifying fragments of high sequence similarity within proteins followed by the classification of these proteins into homologous groups can help identify distant functional relationships in proteins, such as for example the PROSITE database. However, motif identification can lead to arbitrarily fragmented protein families. Identifying conserved regions within large sets of proteins is now achieved through iteration and aggregation of sequences using alignment methods such as PSI-BLAST or hidden Markov chain profiles of

*The authors thank the Wellcome Trust for its support under Grant 072831/Z/03/Z to Birkbeck University of London. We are grateful to E. Koonin for his helpful comments and advice. We also thank the referees for their comments, which helped us to improve the presentation.

[†]For communications, use e-mail address: mirkin at dcs.bbk.ac.uk

[‡]All from School of Computer Science and Information Systems, Birkbeck, University of London, UK

[§]Centre for Virology, Department of Infection, University College London, UK

the close families.

In this paper, we describe a different strategy to improve the sensitivity and specificity of computational methods for grouping proteins into families. This strategy involves a two-stage process: (1) building "fragmented" motif based homologous protein families (HPFs) such as those developed in VIDA database [2]; (2) aggregating the HPFs using their sequence based similarity estimates. The similarity between two HPFs is estimated by comparing their "homologous neighbours" (HN) sets, that are derived from sets of homologous neighbours of individual proteins in the families. An individual protein sequence is assigned with HN set by using an alignment based tool such as PSI-BLAST. Having the HN sets defined, we measure similarity between HPFs as the similarity between their HN sets. To aggregate the HPFs, we adapt a method for incomplete clustering, that is derived from data recovery approaches [13]. This method involves only one parameter to distinguish between functionally similar and dissimilar HPFs. This parameter, the threshold similarity shift value, is estimated by comparing two distributions of similarity values: one related to lists with clearly different functions and the other related to lists with clearly similar functions.

The aggregate protein families are mapped onto an evolutionary tree of herpesviruses according to both the maximum parsimony and the maximum likelihood principles. We consider 740 HPFs residing in 30 herpesvirus genomes and analyse the reconstructed composition of the herpesvirus universal common ancestor (HUCA) as well as those of the herpesvirus superfamilies' ancestors.

2 Methods

2.1 Measuring similarity between HPFs

There are many examples of proteins, especially virus encoded proteins, whose pair-wise similarity is low, but which are known to be functionally related and which have many common homologues. For example the glycoprotein H like protein of murine herpesvirus 4 (gi: 1246777) and the UL22 protein of Bovine herpesvirus 1 (gi: 1491636) have minimal sequence identity (15%, identified on the second PSI-BLAST iteration), and have been initially assigned to separate HPFs within the VIDA database, namely HPFs 12 and 42 [2]. However, their sets of homologous protein neighbours (with 20% or greater sequence identity), contain 25 and 20 sequences, respectively, and have 14 common proteins, making the overlap between the homologous protein lists quite significant: the average relative overlap is 63% (14/25=56% in one of the sets and 14/20=70% in the other). To alleviate the issue, PSI-BLAST runs are conventionally reiterated for accruing distantly related proteins into families. This, however, may import irrelevant proteins or proteins that are not within the organism group under investigation. An HPF obtained in this way requires manual curation, but the overlap between the neighbourhood lists suggests a different computational strategy.

Given a query protein sequence p, we utilise the PSI-BLAST program [3] to sort all protein sequences under consideration (we use those in the NCBI Entrez web site [16]) by their similarity to the query sequence. An initial fragment of this sorted list, defined by a contrasting cut-off similarity value, is identified. The list of all those proteins similar to this fragment that are also identified in our collection of herpesvirus genome protein sequences makes the homology neighbourhood of p, denoted by $l(p)$.

Given a protein family h consisting of m proteins $p1$, $p2$,...,pm, with herpesvirus constrained HN sets $l(p1)$, $l(p2)$, ...,$l(pm)$ assigned to each of them, we aggregate these sets by using the majority rule. Let us assign a membership score $s(p)$ to each sequence; $s(p)$ being defined as the proportion of the HN sets $l(p1)$,..,$l(pm)$ to which p belongs; this is 1 if p belongs to all m of the sets.

Given $a > 0$, the a-majority list $M_a(h)$ is defined as the set of those p for which $s(p) \geq a$. For $a = 1/2$, $M_{1/2}(h)$ is the so-called *simple* majority list. As a decreases, the size of $M_a(h)$ can only increase, so that for $a \leq 1/m$ the a-majority list $M_a(h)$ is the set-theoretic union of the sets $l(pi)$ for all $pi \in h$.

To determine an appropriate value for a, we accept the following view: the proteins in an HPF have developed over a period of time; thus, the longer the time period spanned by the a-majority list proteins, the smaller should be the value chosen for a.

To measure similarity between two HPFs represented by their HN sets of protein sequences, L1 and L2, one should rely on the quantities involved: the size of the overlap between L1 and L2, denoted by a, the number of elements in L1 denoted by a1, and the number of elements in L2 denoted by a2. To take into account the relative size of the overlap, we use the average proportion of the overlap, $mbc = \frac{1}{2}(\frac{a}{a1} + \frac{a}{a2})$, known as the Maryland Bridge coefficient [15]. This index is co-monotone with the popular Jaccard coefficient $J = \frac{a}{a1+a2-a}$, but does not suffer from the intrinsic flaw of the Jaccard coefficient, which systematically underestimates the similarity [15].

2.2 Clustering HPFs with similarity data

As is well-known, no known clustering algorithm reliably determines the number of clusters. We therefore employ an approach that finds high-density clusters one-by-one. Such methods are becoming increasingly popular in bioinformatics; see, for instance, the algorithm CAST [4], which is similar to our algorithm ADDI-S [12] described below. A criterion for finding a high-density cluster should combine

two conflicting principles: (i) a cluster should contain only highly similar entities, and (ii) the cluster size should be as large as possible. We utilise a function $A(S)$ of a cluster S, that incorporates both of these principles. Specifically, $A(S)$ is the product of the within-cluster average similarity $b(S)$ and the cluster size N_S, i.e. $A(S) = b(S)N_S$. The choice of this criterion fits well into several frameworks: (a) maximum density subgraphs, (b) spectral clustering, and (c) the data recovery approach, in which $A(S)$ emerges in a least-squares context [12, 13].

A shift in the origin of the similarity measure may affect the clustering results, in the manner similar to that of threshold graphs drawn at different thresholds, which can be of advantage in contrasting within- and between- cluster similarities.

To maximise $A(S)$, we utilise the following method, ADDI-S, introduced in [12], which extracts clusters one-by-one.

Denote the similarity data after the similarity shift by $B = (b_{ij})$, $i, j \in I$, where I is the set of all HPFs under consideration. Take an arbitrary $i^* \in I$ and find j^*, such that $b_{i^*j^*}$ is maximised over all $j \in I$. If $b_{i^*j^*} \leq 0$, the computation stops: S must be a singleton consisting of just i^*. Otherwise, put both i^* and j^* into S. S is updated as follows. Given the current S, $b(i, S)$, the average similarity between i and all S members, is calculated for all $i \in I$. Then the within-cluster average $b(S)$ is calculated and the threshold $\pi = b(S)/2$ is used to select those $i \notin S$ that satisfy $b(i, S) > \pi$ and those $i \in S$ that satisfy $b(i, S) < \pi$. If there are such i's, put one of them into S or out of S, if $i \notin S$ or $i \in S$, respectively. If there are none, stop. This procedure is applied $|I|$ times, starting from every $i^* \in I$, and the densest cluster, according to $A(S)$, is selected.

ADDI-S is a local search algorithm for maximizing $A(S)$. The resulting cluster is mathematically proven to be well separated from the rest: its "attraction" coefficients $\beta(i, S) = b(i, S) - b(S)/2$ are positive for within-cluster elements $i \in S$ and negative for out-of-cluster elements $i \notin S$ [13].

A clustering is produced by the repeated application of algorithm ADDI-S to those entities that remain unclustered. The process stops when the remainder manifests no positive similarities between its elements. The result is a set of clusters $\{S_t\}$, $t = 1, ..., T$, each assigned with its contribution value, $A(S_t)^2$, and the remaining unclustered part.

Although the ADDI-S method utilises no ad hoc parameters, the clustering results do depend on the similarity shift value that must be defined by the user. However, this value can be chosen based on biologically inspired considerations as explained in section 4.2.

2.3 Mapping HPFs to the evolutionary tree

Given an evolutionary tree over genomes together with the phylogenetic profile of an HPF in the extant species (leaf nodes of the tree), the problem that arises is to generate the most plausible evolutionary scenario that would lead to this phylogenetic profile. Such a scenario may involve the evolutionary events of emergence, inheritance, horizontal transfer and loss. Since, at this level of resolution, we cannot distinguish between emergence and horizontal transfer of a gene, we refer to either of these events as a gain. The total number of loss and gain events in a scenario shows the extent of incompatibility between the evolutionary history of the given gene and the species tree. Among all possible scenarios, we select the most parsimonious, i.e. requiring the minimum number of events to explain the observed phylogenetic profile, or the most likely, i.e. those for which the probability of the observed phylogenetic profile is maximised. We consider first the criterion of maximum parsimony and then that of maximum likelihood.

2.3.1 Maximum Parsimony reconstruction

Since the likelihoods of loss and gain events are likely to differ, we may need to weight them differently. This is achieved by introducing event penalties l and g; the loss penalty l is normally assigned the value 1, whereas the gain penalty g can differ from 1. Then a parsimonious scenario should minimise the total weighted score; this is the inconsistency of the HPF. Recently, a number of approaches have been proposed for this problem [17, 14, 10], of which only that by the authors [14] involves no additional parameters or constraints. This method proceeds by recursively building a parsimonious scenario for each parent node from parsimonious scenarios for its children. At each node of the tree, sets of loss and gain events are maintained under both the assumption that the gene has been inherited at the node and the assumption that it has not been inherited. It is necessary to distinguish these two cases since, clearly, it is only meaningful to consider the loss of a gene at a node if it was inherited at that node, or the gain of a gene if it was not inherited. We denote the number of events under the inheritance and non-inheritance assumptions by e_i and e_n, respectively, where gains are weighted by the gain penalty g. An evolutionary scenario at a given node is defined by a pair of sets (G, L), representing the gains and losses in the subtree rooted at the node. We use (G_i, L_i) and (G_n, L_n) to denote scenarios under the inheritance and non-inheritance assumptions, respectively. As shown in [14], in a parsimonious scenario, the parental inconsistency score can be derived from those of its children (indicated by subscripts 1 and 2) as $e_i = \min(e_{n1} + e_{n2} + l,\ e_{i1} + e_{i2})$ or $e_n = \min(e_{i1} + e_{i2} + g,\ e_{n1} + e_{n2})$, under the inheritance or non-inheritance assumption, respectively. These lead to a tree accumulation algorithm PARS [14] for computing parsimonious scenarios for parental nodes. At a leaf node the four sets G_i, L_i, G_n and L_n are empty, except that

$G_n = \{a\}$ if gene a is present in the given leaf or $L_i = \{a\}$ if a is not present.

2.3.2 Maximum Likelihood reconstruction

The algorithm PARS can be refined from using only the heuristic principle of parsimony to incorporating a maximum likelihood approach, using probabilities of loss and gain of genes at each node of the tree. These probabilities can be estimated from the totality of parsimonious scenarios produced by PARS for all HPFs under consideration, taking account of the total numbers of gains and losses in reconstructed scenarios as well as which HPFs could be potentially gained or lost. Given the probabilities, λ of a loss and γ of a gain, respectively, a probabilistic scoring function can be developed for utilisation in a modified version of the algorithm PARS for producing a maximum likelihood scenario for any given phyletic profile. For an internal node of the tree, under the assumption that the gene has been inherited from the node's parent, the probability of a scenario leading to the observed phyletic profile in the subtree rooted at the node is equal to either $\lambda p_{n1} p_{n2}$ or $(1-\lambda) p_{i1} p_{i2}$, where p_{n1} and p_{n2} (p_{i1} and p_{i2}) are the probabilities of the scenarios for the child subtrees under the assumption that the gene was not inherited (inherited) by the children. The first of these expressions, $\lambda p_{n1} p_{n2}$, is the probability of the scenario in the case in which an inherited gene is lost at the node, and thus not inherited by its children. This assumes the stochastic independence of evolutionary events in disjoint subtrees and that the Markov property holds. The second expression is the corresponding probability in the case in which the inherited gene is not lost at the node. We select the scenario for which the probability is maximum. Analogous formulae can be derived for a non-inherited gene; in this case, γ is used instead of λ. The modified algorithm PARS can be used to compute the maximum likelihood scenario for the whole tree. This modified algorithm, MALS, differs from PARS by both the scoring function and the event sets at the leaves.

It is easy to see that MALS can also be applied when loss and gain probabilities are node-specific.

After MALS is applied to the totality of all HPFs, the resulting scenarios and thus the numbers of gains and losses can change from those found with PARS, which were used for defining MALS' loss and gain probabilities. This could lead to different posterior gain and loss probabilities. These posterior probabilities can be used as input for iterating MALS, until convergence is achieved. We have no formal proof that such iterated MALS will always converge, but in all our experiments both with the VIDA and COG [18] databases this was the case.

The maximum likelihood approach above differs from the conventional ones in that we determine a single scenario for a pre-specified evolutionary tree rather than a distribution of probabilities over all possible scenarios and trees.

Table 1: List of the 30 herpesvirus genomes under consideration.

#	VIDA Ref.	Genome	GenBank Ref.
		Alphaherpesvirinae	
01	CeHV-1	Cercopithecine hv 1	NC_004812
02	HHV-1	Human hv 1/simplex 1	NC_001806
03	HHV-2	Human hv 2/simplex 2	NC_001798
04	EHV-4	Equid hv 4	NC_001844
05	EHV-1	Equid hv 1	NC_001491
06	BoHV-1	Bovine hv 1	NC_001847
07	BoHV-5	Bovine hv 5	NC_005261
08	CeHV-7	Cercopithecine hV 7	NC_002686
09	HHV-3	Human hv 3/varicella-zoster	NC_001348
10	MeHV-1	Meleagrid hv 1	NC_002641
11	GaHV-2	Gallid hv 2/Marek's disease	NC_002229
12	GaHV-3	Gallid hv 3	NC_002577
13	PsHV-1	Psittacid hv 1	NC_005264
		Betaherpesvirinae	
14	HHV-6	Human hv 6	NC_001664
15	HHV-7	Human hv 7	NC_001716
16	HHV-5	Human hv 5/cytomegalovirus	NC_006273
17	ChCMV	Chimpanzee cytomegalovirus	NC_003521
18	MuHV-2	Murid hv 2/rat cytomegalovirus	NC_002512
19	TuHV	Tupaiid hv	NC_002794
		Gammaherpesvirinae	
20	HVS-2	Saimiriine hv 2	NC_001350
21	AtHV-3	Ateline hv 3	NC_001987
22	EHV-2	Equid hv 2	NC_001650
23	BoHV-4	Bovine hv 4	NC_002665
24	MuHV-4	Murid hv 4/murine hv 68	NC_001826
25	RRV-17577	Macaca mulatta rhadinovirus	NC_003401
26	HHV-8	Human hv 8/Kaposi's sarcoma	NC_003409
27	AIHV-1	Alcelaphine hv 1	NC_002531
28	CeHV-15	Cercopithecine hv 15	NC_006146
29	HHV-4	Human hv 4/Epstein-Barr	NC_001345
30	CaHV-3	Callitrichine hv 3	NC_004367

3 Data

3.1 Evolutionary tree

A set of 30 complete herpesvirus genomes covering the α, β and γ herpesvirus superfamilies (see Table 1) has been extracted from the herpesvirus database VIDA, release 3 [2]; and an evolutionary tree has been built over the genomes for the conserved DNA polymerase gene using the PHYLIP package [7] (see Figure 1). This tree agrees well with previously published instances of herpesvirus phylogenies.

Additional support for this being a suitable tree is that very similar trees have been constructed by us using more comprehensive data, such as the genome phylogenetic profiles formed by the 257 HPFs that are present in two or more of the genomes [2]. Second, our tree is similar to that published in [11] and reproduced in [5] on the subset of 25 herpesvirus genomes common to our tree. The latter tree is based on comprehensive data including expert knowledge.

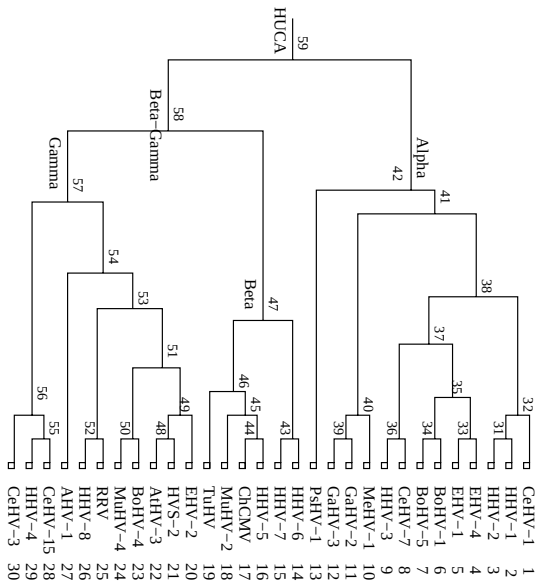

Figure 1: Herpesvirus evolutionary tree.

Our tree in Figure 1 differs from the tree in [11, 5] on two clusters rooted at nodes 43 (four β genomes) and 53 (seven γ genomes), respectively. This difference, however, falls within the margin of uncertainty of the tree topology indicated in [11, 5]. All subsequent computations for phyletic assessment of gene gain and loss events were performed on both trees, that on Figure 1 and that with the topology of subtrees rooted at nodes 43 and 53 changed according to the reconstruction in [11, 5]. The mapping results reported in section 5 below are the same for both tree topologies.

3.2 Homologous Protein Families (HPFs)

A set of 740 homologous protein families (HPFs) residing in 30 genomes were extracted from the VIDA database [2]. Each HPF is defined by a conserved fragment in the proteins constituting the HPF; these were computed using the algorithm XDOM [8, 2]. In this way, each HPF is proposed to represent basal functional grouping, whose origin can be mapped to the evolutionary tree under the assumption that the function is inherited according to the tree topology. As discussed, such motif based protein family assignment can suffer from fragmentation of protein families and from the non-assignment of proteins to a family due to lack of pairwise similarity. We therefore used our method of generating and assessing similarity between lists of homologous proteins to check the validity of the starting HPFs and merge HPFs that were artificially fragmented into aggregate protein families (APFs).

4 Tuning Methods to Data

There are two places in our method that require fitting the computational parameters to the data: (1) selection of parameter a in the majority rule, and (2) selection of the similarity shift at the ADDI-S clustering. These will be described in this section.

4.1 Selection of the majority threshold

We considered values $a = 2/3, 1/2, 1/3, 1/4, 1/5, 1/6$. Obviously, there is no need to take a between these values, since they would produce the same majority lists. The majority lists at $a = 1/6$ coincide with the set-theoretic unions for all HPFs comprising six or less proteins. At any specified a, the mbc similarity coefficients, the average percentage of the overlap, have been computed between a-majority lists obtained for individual HPFs. Then the obtained HPF-to-HPF similarity matrix was processed with the ADDI-S clustering algorithm at different similarity shift values, from $b = 0$ incremented by 0.1 to $b = 0.9$ and, a greater similarity shift value, $b = 0.97$.

To compare two different clusterings, we use the same mbc coefficient. We apply a flexible rule to identify clusters S_{1i} and S_{2j} as similar when they differ by just one or two elements.

The majority threshold has been set at the level of 20%, i.e. $a = 1/5$, because:

1. Clusterings produced at different similarity shift levels differ minimally. The median mbc similarity value between clusterings corresponding to "neighbouring" majority thresholds 1/6 and 1/5 is 0.98; 1/5 and 1/4, 1.00; 1/4 and 1/3, 0/99; 1/3 and 1/2, 0.96. The similarity between clusterings at non "neighbouring" thresholds slightly decreases. The average mbc similarity value varies similarly, taking its maximum at the majority thresholds 1/5 and 1/4. The sets of unclustered entities behave similarly.

2. The clustering found over 20%-majority lists is "central" among other clusterings; it is more similar to the other clusterings than at any other of the considered majority thresholds.

3. The clustering found over 20%-majority lists is more similar than the others to clusterings produced with the homology lists obtained with the iterated PSI-BLAST search [3], starting from a random protein in an HPF. Repeated PSI-BLAST search, over an averaged profile of the first search results, allows one to catch more distant homologues to the query sequence [3]. The median similarity between the clustering at 20%-majority lists and the clustering found at HPF neighbourhood lists of the first iteration is 0.91; lists of the

second iteration, 0.82; and lists of the third iteration, 0.50.

4.2 Choosing the similarity shift

At the chosen 1/5-majority neighbourhood lists, the summary number of obtained clusters corresponding to aggregate functions changes from 99 at no similarity shift at all to 29 at $b = 0.97$. The number of HPFs remaining unclustered changes from 430 to 681, respectively, leading to the total numbers of "aggregate" functions from 529 to 710. Note that the latter number corresponds to the situation when the HN sets of the clustered HPFs are practically the same sets of proteins: to overlap at the level of 97% or higher, majority lists of less than 30 elements (that is, almost all) must be equal to each other.

To choose an appropriate similarity shift value, we compare the values of similarity between the 1/5-majority neighbourhood sets of two types of pairs of HPFs: those that are synonymous and those that are not. Two proteins are considered synonymous if they are consistently named between the herpesvirus genomes and/or they share the same known function. Such proteins should therefore belong in the same aggregate protein family. Two proteins are considered non-synomymous if they have different functions and thus should belong to different protein families.

Out of the 287 available pairs of HPFs with known function and positive similarity value, no non-synonymous pair has a greater *mbc* similarity than 0.66, which should imply that the shift value $b = 0.67$ confers specificity for the production of APFs.

Unfortunately, the situation is less clear cut for synonymous proteins. There are 24 out of 86 synonymous pairs (28%) that have their mutual similarity value less than 0.67. Thus, accepting the similarity shift at 0.67, 28% of the synonymous pairs would not be identified suggesting that at this similarity shift the method lacks sensitivity. To choose a similarity shift that minimises the error in assigning negative and positive similarity values, one needs to compare the distribution of similarity values in the set of synonymous pairs with that in the set of non-synonymous pairs. As Figure 2 shows, the graphs intersect when the similarity value *mbc* is 0.42. The number of synonymous pairs whose similarity falls into the wrong side of 0.42 (that is, less than 0.42) falls to 11, whereas the number of non-synonymous pairs whose similarity is higher than 0.42 increases to 7 (from 0 at 0.67), which leads to the minimum summary error rate of 16%, at $b = 0.42$.

Thus two possible similarity shift values are indicated: (a) $b = 0.67$ to guarantee specificity in that non-synonymous HPFs are not be clustered, and (b) $b = 0.42$ to ensure the minimum misclassification error rate.

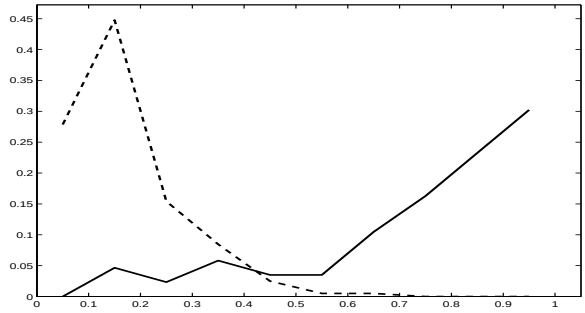

Figure 2: Empirical percentage frequency functions (y-values) for the sets of synonymous pairs (solid line) and non-synonymous pairs (dashed line). The x-values represent the *mbc* similarity.

5 Reconstruction of APF histories

The two similarity shift values indicated, i.e. $b = 0.67$ and $b = 0.42$, lead to somewhat different but rather compatible clusterings of the set of 740 HPFs under consideration. There are 80 APF clusters comprising an original 180 HPFs and leaving 560 HPFs unclustered at $b = 0.67$. There are 102 APF clusters, over an original 249 HPFs, and 491 HPFs unclustered at $b = 0.42$.

The first 80 clusters extracted at similarity shift $b = 0.42$ correspond one-to-one to the 80 clusters obtained at $b = 0.67$. All 22 of the additional clusters extracted at $b = 0.42$ are doublets with similarity values of 0.50 to 0.62.

The aggregation found at $b = 0.67$ suggests $560 + 80 = 640$ APFs altogether whereas $b = 0.42$ leads to a smaller total, $491 + 102 = 593$. We analysed reconstructions of histories of APFs at each of the two aggregations and found minimal differences at the ancestor nodes related to ancestors of superfamilies α, β and γ as well as the more universal common ancestors, HUCA and $\beta\gamma$. Moreover, application of the iterated MALS algorithm, starting at PARS results at gain penalty values from 1 to 3, did not lead to any changes of the reconstructed HUCA, consistent with the observation that most of herpesvirus HPFs follow the topology of the evolutionary tree.

The gains and losses reconstructed in HUCA and its immediate descendants, α, $\beta\gamma$, β and γ, differ minimally between the aggregations found at similarity shifts $b = 0.67$ and $b = 0.42$. We present results at the more conservative level $b = 0.67$ and then comment on the only difference of notice that comes at $b = 0.42$.

Of the four ancestors, α, $\beta\gamma$, β and γ, only the contents of the α superfamily is relatively well studied. Of its 33 gained HPFs (plus the inherited HUCA contents) only 9 are of unknown function.

This pattern is not repeated in the $\beta\gamma$ ancestor, with 10 gains (plus the inherited HUCA) of which only 2 are of known function. Similarly, of 31 additional gains at β-ancestor, only 10 have known function and of 32 additional gains at the γ-ancestor, the function is known for only 9. To-

Table 2: Comparison between a previously determined herpesvirus common ancestor D-HUCA's [5, 6] list of functions in the herpesvirus ancestor (two columns on the right) versus the results from the mapping of HPF/APFs (first four columns), with function descriptions taken from VIDA.

Mapping	A/HPF	Function	Description	HSV-1 Gene	D-HUCA
					Peripheral Enzymes
HUCA	8	Nucleotide repair/ metabolism	uracil-DNA glycosylase, HHV-1 UL2	UL2	Uracil-DNA glycose
HUCA	24	Nucleotide repair metabolism	RNA reductase large subunit, HSV-1 UL39	UL39	RNA reductase; large subunit
HUCA	33	Nucleotide repair metabolism	RNA reductase small subunit, HHV-1 UL40	UL40	RNA reductase small subunit
HUCA	APF 10			UL23	Thymidine Kinase
	2	*Nucleotide repair/ metabolism*	*thymidine kinase*		
	27	*"*	*thymidine kinase*		
HUCA	43	Nucleotide repair/ metabolism	dUTPase, HHV-8 ORF54	UL50	dUTPase
					Surface and Membrane
HUCA	20	Membrane glycoprotein	glycoprotein M, HHV-1 UL10	UL10	Glycoprotein M; complexed with glycoprotein N
HUCA	3	Membrane glycoprotein	glycoprotein B, HHV-1 UL27	UL27	Glycoprotein B
HUCA	APF 3			UL22	Glycoprotein H; complexed with glycoprotein L
	42	*Membrane/ glycoprotein*	*glycoprotein H, HHV-1 UL22*		
	12	*"*	*glycoprotein H, HHV-8 ORF22*		
	531	*"*	*glycoprotein H, HHV-8 ORF22*		
Node 32	267	Virion protein	envelope protein, HHV-1 UL49A	UL49A	Glycoprotein N; complexed with glycoprotein M
ALPHA	47	Membrane glycoprotein	glycoprotein L, HHV-1 UL1	UL1	Glycoprotein L; complexed with glycoprotein H
BETA	50	"	glycoprotein L, HHV-5 UL115		
GAMMA	114	"	glycoprotein L, HHV-8 ORF47		
GAMMA	296	"	glycoprotein L, MuHV-4 ORF47		

gether, these three ancestors, $\beta\gamma$, β and γ, received 73 gains of which 52, more than 70%, are of unknown function.

The reconstructions of the ancestors with APFs found at the similarity shift $b = 0.42$ are essentially the same. The only exception is the ancestor of the α superfamily, which gains three more APFs at $b = 0.42$. These are APF81 comprised of HPFs 9 and 504, both of glycoprotein C; APF82 comprised of HPF 38 and HPF 736, both of glycoprotein I; and APF84 comprised of HPF 47 and HPF 205, both of glycoprotein L.

The common ancestor of herpesviruses, HUCA, according to our reconstruction, should comprise 45 HPFs aggregated to 29 APFs, i.e. 29 protein families. These are well studied proteins with only three of the participating families, HPFs 17, 23 and 107, of no known function. Our HUCA is consistent with the work of others, D-HUCA[5, 6], but does not include all the protein families assigned by Davidson et al. This concurs with our view that our approach, relying only on sequence similarity alone, is conservative.

Typical relations between our mapping results and D-HUCA are illustrated in Table 2

In some cases, it is clear that the fragmented HPFs fail to aggregate at that level of moving from the α, β and γ-ancestor into HUCA because of almost zero sequence similarity between them. For example, Table 2 shows how a difference between the reconstructed HUCA and D-HUCA can emerge. All three ancestors, of each α-, β-, and γ families, have a glycoprotein L. However, the corresponding HPFs, 47, 50 and 296, have no significant sequence similarity and, thus, cannot be combined together, even in terms of the neighbourhood lists. Still, at the genome organisation level, illustrated on Figure 3, each of the glycoprotein L genes always exactly precedes the corresponding Uracil-DNA glycosylase gene, which is mapped into HUCA. This suggests these are common ancestral genes indeed; just they have undergone sequence change to a level where sequence similarity is no longer sufficient to assign homology. Putting the corresponding gene UL2 into D-HUCA has been based on experimental evidence that in the α-, β-, and γ families, glycoprotein L sequences in HPF 47, 50 and 296 functionally complex with glycoprotein H [6].

6 Discussion

Reconstructing evolutionary relationships using whole genome gene content provides novel insights into the gene gain and loss events that have shaped the evolution of extant organisms. To achieve such reconstructions, similarity between proteins must be established to allow their correct place-

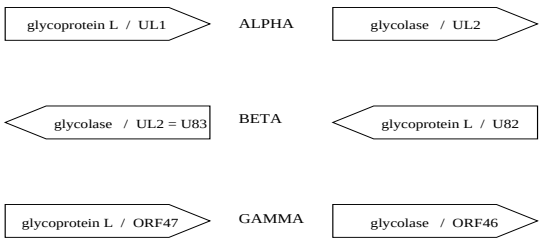

Figure 3: Positional homology between glycoprotein L sites in the herpesvirus superfamilies α, β and γ. The homology suggests that the glycoprotein L gene co-functions with the glycosylase gene and thus the former, like the latter, should be mapped to HUCA.

ment within a phylogenetic tree. Here we present a method that first aggregates homology lists of individual proteins into motif based families, and then finds family clusters based on similarity between the HN sets. This provides an efficient and accurate way to identify protein families for such studies. Conventional clustering algorithms that partition the dataset into a pre-specified numbers of clusters are not of great help here because one needs not to partition but rather identify a relatively few aggregations of protein families leaving the rest unclustered (incomplete clustering). Importantly, the true number of clusters is unknown and must be assessed through parameters adjustment. These parameters, namely, the majority threshold and similarity shift, can be reasonably determined from the data.

We have applied this method to the phyletic reconstruction of herpesvirus phylogeny and the results support the validity of the method. We have successfully reconstructed a herpesvirus universal common ancestor (HUCA) and the most likely common ancestors of the α, β and γ herpesviruses. The method is still limited by the requirement for sequence similarity but consistent with current herpesvirus genome annotation. We show that inclusion of gene position information to this analysis can help in identifying functionally homologous sequences with minimal protein identity.

References

[1] M.M. Alba, R. Das, C. Orengo, and P. Kellam, "Genomewide function conservation and phylogeny in the herpeviridae," *Genome Research*, 11, 43-53, 2001.

[2] M.M. Alba, D. Lee, F.M. Pearl, A.J. Shepherd, N. Martin, C. Orengo, and P. Kellam, "VIDA: A virus database system for the organisation of animal virus genome open reading frames," *Nucleic Acids Research*, 29, 133-136, 2001.

[3] S.F. Altschul, T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Research*, 25, 3389-3402, 1997.

[4] A. Ben-Dor, R. Shamir, Z. Yakhini, "Clustering gene expression patterns," *Journal of Computational Biology*, **6**, 281-297, 1999.

[5] A.J. Davison, "Evolution of the herpesviruses," *Veterinary Microbiology*, 86, 69-88, 2002.

[6] A.J. Davison, D.J. Dargan, and N.D. Stow, "Fundamental and accessory systems in herpesvirus: Review," *Antiviral Research*, 56, 1-11, 2002.

[7] J. Felsenstein, *PHYLIP 3.6: Phylogeny Inference Package*, http://evolution.genetics. washington.edu/phylip/, 2001.

[8] J. Gouzy, P. Eugene, E.A. Greene, D. Khan, and F. Corpet, "XDOM, a graphical tool to analyse domain arrangements in any set of protein sequences," *Comput. Appl. Biosciences*, 13, 601-608, 1997.

[9] R. Holzerlandt, C. Orengo, P. Kellam, and M.M. Alba, "Identification of new herpesvirus gene homologs in the human genome," *Genome Research*, 12, 1739-1748, 2002.

[10] V. Kunin and C.A. Ouzounis, "GeneTRACE – reconstruction of gene content of ancestral species," *Bioinformatics*, 19, 1412-1416, 2003.

[11] D.J. McGeoch, A. Dolan, and A.C. Ralph, "Toward a comprehensive phylogeny for mammalian and avian herpesviruses," *Journal of Virology*, 74, 10401-10406, 2000.

[12] B. Mirkin, "Additive clustering and qualitative factor analysis methods for similarity matrices," *Journal of Classification, 4*, 7-31, 1987.

[13] B. Mirkin, *Clustering for Data Mining: A Data Recovery Approach*, Chapman and Hall, Boca Raton, 2005.

[14] B. Mirkin, T. Fenner, M. Galperin, and E. Koonin, "Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes," *BMC Evolutionary Biology*, 3:2, 2003 (www.biomedcentral.com/1471-2148/3/2/).

[15] B. Mirkin and E. Koonin, "A top-down method for building genome classification trees with linear binary hierarchies," in *Bioconsensus*, M. Janowitz, J.-F. Lapointe, F. McMorris, B. Mirkin, and F. Roberts, Eds. DIMACS Series, Vol. 61, Providence: AMS, 97-112, 2003.

[16] *NCBI GenBank/Entrez web site*, http://www.ncbi.nlm.nih.gov/entrez, 2006.

[17] B. Snel, P. Bork, and M.A. Huynen, "Genomes in flux: The evolution of archaeal and proteobacterial gene content," *Genome Research*, 12, 17-25, 2002.

[18] R.L. Tatusov, M.Y. Galperin, D.A. Natale, and E.V. Koonin, "The COG database: a tool for genome-scale analysis of protein function and evolution," *Nucleic Acids Research*, 28, 33-36, 2000.