BMC Bioinformatics

Methodology article

A hybrid clustering approach to recognition of protein families in 114 microbial genomes

Timothy J Harlow[1,2], J Peter Gogarten[3,4] and Mark A Ragan*[1,2,4]

Address: [1]Institute for Molecular Bioscience, The University of Queensland, Brisbane, Qld 4072, Australia, [2]Australian Research Council (ARC) Centre in Bioinformatics, Australia, [3]Molecular and Cell Biology, University of Connecticut, Storrs, CT 06269-3044 USA and [4]Canadian Institute for Advanced Research, Program in Evolutionary Biology, Canada

Email: Timothy J Harlow - t.harlow@imb.uq.edu.au; J Peter Gogarten - gogarten@uconn.edu; Mark A Ragan* - m.ragan@imb.uq.edu.au

* Corresponding author

Published: 29 April 2004

BMC Bioinformatics 2004, **5**:45

Received: 23 October 2003
Accepted: 29 April 2004

This article is available from: http://www.biomedcentral.com/1471-2105/5/45

Abstract

Background: Grouping proteins into sequence-based clusters is a fundamental step in many bioinformatic analyses (*e.g.*, homology-based prediction of structure or function). Standard clustering methods such as single-linkage clustering capture a history of cluster topologies as a function of threshold, but in practice their usefulness is limited because unrelated sequences join clusters before biologically meaningful families are fully constituted, *e.g.* as the result of matches to so-called promiscuous domains. Use of the Markov Cluster algorithm avoids this non-specificity, but does not preserve topological or threshold information about protein families.

Results: We describe a hybrid approach to sequence-based clustering of proteins that combines the advantages of standard and Markov clustering. We have implemented this hybrid approach over a relational database environment, and describe its application to clustering a large subset of PDB, and to 328577 proteins from 114 fully sequenced microbial genomes. To demonstrate utility with difficult problems, we show that hybrid clustering allows us to constitute the paralogous family of ATP synthase F1 rotary motor subunits into a single, biologically interpretable hierarchical grouping that was not accessible using either single-linkage or Markov clustering alone. We describe validation of this method by hybrid clustering of PDB and mapping SCOP families and domains onto the resulting clusters.

Conclusion: Hybrid (Markov followed by single-linkage) clustering combines the advantages of the Markov Cluster algorithm (avoidance of non-specific clusters resulting from matches to promiscuous domains) and single-linkage clustering (preservation of topological information as a function of threshold). Within the individual Markov clusters, single-linkage clustering is a more-precise instrument, discerning sub-clusters of biological relevance. Our hybrid approach thus provides a computationally efficient approach to the automated recognition of protein families for phylogenomic analysis.

Background

The comprehensive classification of proteins into similarity groups is an important but difficult challenge in post-genomic bioinformatics. These similarity groups might be based on *e.g.* common sequence, structure, or function. We are studying the evolutionary diversification of microbial genomes [1,2] and therefore wish to group proteins into families based on sequence similarity for subsequent

multiple alignment and inference of phylogenetic trees. The ongoing rapid appearance of new microbial genome sequences makes it imperative that this clustering be rapid, scalable, and automated to the extent possible.

Sets of protein sequences linked by pairwise matches can be clustered, and the resulting clusters interpreted as families [3,4]. This has proven successful for protein domains (ADDA [5], DIVCLUS [6], PRODOM [7]) and for complete protein sequences (ProtoMap [8], SYSTERS [9]). Sequence similarity is first assessed pairwise, typically using BLAST [10], following which single-linkage clustering [11] is used to generate a hierarchy of internal nodes or subtrees. Sometimes FASTA [12] or another statistically based comparison tool is utilised in place of BLAST, or another criterion in place of single-linkage (SL) clustering; although our discussion below focuses on BLAST and SL clustering, the argument applies more generally to other pairwise comparison tools and clustering criteria.

This approach to recognition of clusters offers several important advantages. Protein family groups built up in this way are not simply unstructured lists, but natively possess an internal structure (topology) readily interpretable in the formalism of graphs, with vertices representing protein sequences, and edges representing pairwise matches. Each edge can be assigned a length that is inversely proportional to the strength of the corresponding pairwise BLAST match. Membership in these protein families can be made more, or less, stringent by adjusting a single threshold that defines when an edge is recognised to be present. This threshold can be based in a straightforward, intuitive way on the BLAST output (*e.g.* bit score or *e*-value). As the stringency of this threshold is (for example) increased, a given paralogous family will often be resolved into its constituent orthologous families.

There is, however, one major drawback to this naïve approach. At threshold values that are sometimes more stringent than required to constitute all orthologous (much less all paralogous) protein families, xenologous (evolutionarily unrelated) proteins begin to be drawn into these clusters, undermining their biological interpretability. This occurs for several reasons. At higher stringencies, BLAST may recognise so-called "promiscuous domains" common to otherwise unrelated proteins [13]. At lower stringencies, BLAST may report weakly convergent motifs [14] or chance similarities. These non-specific clusters grow explosively in size with decreasing stringency, thereby preventing the useful extension of standard clustering down to the threshold values required to fully constitute most evolutionarily related protein families.

Alternative approaches are of course available, but are not necessarily appropriate for the problem at hand. Approaches based on machine learning, *e.g.* [15,16], identify putative homologs even at low levels of sequence identify, but can be computationally very expensive. Putative remote homologs can likewise be recognised based on similarities in folded structure. However, inclusion of remote homologs is likely to be counterproductive for us, for three reasons: rigorous multiple alignment and inference of phylogenetic trees scale exponentially with number of sequences; alignment of weakly similar sequences is problematic; and weakly supported branches contribute little or no biological information to our analyses. For this and similar applications, it is therefore far better that remote homologs be excluded from the analysis pipeline at the clustering stage, rather than later.

Recently, an alternative approach based on the Markov Cluster algorithm (MCL: [17]) has been introduced to comparative genomics [4]. The MCL algorithm simulates random walks through a graph (*e.g.* of sequences as vertices, and edges as pairwise matches). By iteratively re-computing random walks and favouring those with higher probability (which tend to be intra-cluster walks) over those with low probability (which tend to be inter-cluster walks), the algorithm partitions the graph into segments that can be interpreted as clusters [4,17]. Computation is rapid and, in application to molecular sequences, the Markov Cluster algorithm produces clusters that resist contamination by promiscuous domains. However, Markov clusters are unstructured lists without internal topology, and as such do not yield information useful to many biologists, *e.g.* a hierarchical ordering of orthologs. Information about edge lengths (strength of BLAST matches) is lost (transformed into stochastic Markov probabilities), making it very difficult to conceptualise these clusters in terms familiar to biologists. How aggressively the Markov clusters find membership (*i.e.*, the resulting granularity) can be adjusted only *via* operators that may have limited useful dynamic range, and are not intuitive to most biologists.

Here we present a hybrid approach to recognizing protein families among very large (multi-genome) datasets. Our hybrid approach preserves the advantages of single-linkage (SL) clustering identified above, but captures the power of the Markov Cluster algorithm to avoid indiscriminate cluster membership. As quality control, we apply our approach to a manually curated database, Protein Data Bank [18], and report how the Structural Classification of Proteins (SCOP) database families and domains [19] map onto the Markov clusters. We demonstrate the application of hybrid clustering to a problem that cannot be usefully addressed by either SL or the Markov Cluster algorithm alone, recognition of orthologs and paralogs of rotary motor ATP synthase F1 subunit proteins [20], and to a 114-genome dataset of protein

Table 1: Markov clustering of PDB (as of 25 February 2003) at selected inflation (I) values. This version of PDB contains 2147 SCOP families and 4526 SCOP domains. There are 1340 SCOP families among the 6435 entries/IDs annotated with one SCOP family each, and 2621 SCOP domains among the 6430 entries/IDs annotated with one SCOP domain each.

Inflation =	1.00	1.10	2.00	3.00	4.00	5.00
Number of clusters ($N \geq 2$)	761	773	906	952	976	998
pure clusters by SCOP family	427	433	424	404	393	384
with 1 SCOP family	380	383	386	369	360	353
with 2 SCOP families	36	39	35	33	31	29
with ≥ 2 SCOP families	11	11	3	2	2	2
most families in 1 cluster	8	7	3	3	3	3
Number of clusters ($N \geq 2$)	761	773	906	952	976	998
pure clusters by SCOP domain	681	693	787	796	793	793
with 1 SCOP domain	358	359	445	460	467	474
with 2 SCOP domains	176	181	194	202	197	195
with ≥ 2 SCOP domains	147	153	148	134	129	124
most domains in 1 cluster	21	21	15	22	23	23
Number of families in all clusters ($N \geq 2$)	831	831	817	814	814	812
with all members in 1 cluster	573	572	511	483	469	455
with all members in 2 clusters	150	146	169	174	177	181
with all members in 3 clusters	50	54	61	70	77	76
with all members in ≥ 3 clusters	58	59	76	87	91	100
most clusters 1 family occurs in	27	27	27	27	27	27
Number of domains in all clusters ($N \geq 2$)	1852	1849	1796	1781	1771	1763
with all instances in 1 cluster	1743	1740	1665	1625	1597	1565
with all instances in 2 clusters	84	84	102	122	139	158
with all instances in ≥ 2 clusters	25	25	29	34	35	40
most clusters 1 domain occurs in	10	10	10	10	10	10

sequences. Finally, we describe how clusters with non-redundant genome coverage ("maximally representative clusters", or MRCs) can be selected automatically from the output of our hybrid method, for subsequent analysis *e.g.* in a phylogenomic pipeline.

Results and discussion

PDB

In order to characterise the behaviour of our hybrid method with a well-understood dataset before application to multi-genome data, we used MCL [17] to cluster PDB at a range of granularities, then mapped SCOP families (fa) and domains (dm) onto the Markov clusters. We assess the resulting mapping from the viewpoints of both PDB (cluster purity) and SCOP (distribution of families or domains over multiple clusters) (Table 1; see Methods for definitions of purity and distribution). Recall that SCOP domains are more compact than SCOP families; one SCOP family can contain one or more SCOP domains.

The MCL inflation parameter *I*, which alters the relative probabilities of within-cluster and between-cluster random walks, is the main parameter by which users can adjust cluster granularity [4,17]. At inflation value I = 1.00, 89.5% of all $n \geq 2$ clusters are "pure" by SCOP domain, *i.e.* contain all instances of any SCOP domain represented in that cluster. Markov clusters found at I = 1.10 are slightly more pure by SCOP domain (89.7%) but purity diminishes somewhat thereafter with increasing graininess, to 79.5% at I = 5.00 (Table 1). Looking instead at SCOP families, the Markov clusters are less pure, ranging from 56.1% at I = 1.00 to 38.5% at I = 5.00. These percentages reflect the relative granularity SCOP families and SCOP domains within this subset of PDB: at I = 1.00, for example, the 761 Markov clusters of $N \geq 2$ contain 1852 SCOP domains but only 831 SCOP families.

Among the $n \geq 2$ Markov clusters that are pure by SCOP family, 88–92% (depending on granularity) contain only a single SCOP family, and >97% contain either 1 or 2 SCOP families. 51–60% of these clusters contain a single SCOP domain, and 77–85% either 1 or 2 SCOP domains. Cluster purity tends to increase slightly with increasing granularity: higher inflation values yield more and cleaner clusters (Table 1).

As clusters become finer-grained, individual SCOP families tend to become distributed among more clusters: the proportion fully contained within a single cluster drops from 69% (I = 1.00) to 56% (I = 5.00). Nevertheless, most (87–93%) families have all their members in ≤ 3 clusters throughout this range of inflation values. SCOP domains

show a similar but less-pronounced trend, decreasing from 94% (I = 1.00) to 88% (I = 5.00) within a single Markov cluster. Most (97–98%) domains have all their members in ≤ 2 clusters.

We carried out single-linkage clustering (*i.e.* completed our hybrid method) on the pure Markov clusters that have ≥ 2 SCOP families each. At I = 1.00, for instance, there are 427 - 380 = 47 such clusters. By raising the clustering threshold S'_{norm} (see Methods) we cleanly resolve 30 of these into constituent families (each in its own pure cluster); in the other 17 clusters, at least one family fragments (is not resolved into a pure cluster). Similarly, among 323 pure Markov clusters with ≥ 2 SCOP domains each, hybrid clustering resolves 292 cleanly, while 31 exhibit domain fragmentation (results not shown). The numbers of fragmenting families and domains decrease at higher inflation values (results not shown).

Multi-genome data

Single-linkage clustering of the 328577 proteins in these 114 completely sequenced microbial genomes yields, at maximum, 14440 clusters of size $n \geq 4$ (Figure 1a). This maximum is reached at S'_{norm} 0.47, at which point 157540 proteins are included in an $n \geq 4$ cluster (Figure 1b). The number of proteins included in clusters of size $n \geq 4$ continues to increase with further decrease in S'_{norm} threshold (Figure 1b), but the number of clusters decreases precipitously (Figure 1a) because existing clusters are progressively and quickly swallowed up into a single large non-specific cluster ("blob") that eventually encompasses 286109 proteins, more than 87% of the total (Figure 1c). We focus on n = 4 as a minimum cluster size because phylogenetic trees become interesting only for $n \geq 4$. For $n \geq 2$ (all non-singular graphs) at I = 1.10, the maximum number of clusters (53120) was at S'_{norm} 0.67, at which point 183119 proteins are members of a $n \geq 2$ cluster (results not shown).

Markov clustering at I = 1.10 yields 4797 clusters ($n \geq 4$). Projecting these onto the BLASTp data followed by SL clustering within (but disallowed between) Markov clusters yields, at S'_{norm} 0.47, 14403 clusters, almost (99.74%) as many as found by SL clustering alone (Figure 2a). As before, as the S'_{norm} threshold is decreased further, the number of proteins in clusters increases (Figure 2b) and the number of clusters decreases (Figure 2a); but disallowing edges that link proteins in different Markov clusters prevents the formation of a non-specific "blob" (Figure 2c). Consolidation within Markov clusters is complete by S'_{norm} 0.02 (Figure 2a), at which point 4802 hybrid protein-family clusters of $n \geq 4$ remain. In this way we estimate that the number of phylogenetically interesting ($n \geq 4$) protein families in these genomes is between about 4802 (the number at S'_{norm} 0.01, where paralogous families might be expected to dominate) and about 14403 (the maximum number observed, where orthologous families, some perhaps not fully consolidated, are presumably more numerous). Similar behaviour is observed at other inflation values and with clusters of size $n \geq 2$, although of course with different numbers of families and of proteins within these families, and with different inflection points.

The size distribution of all hybrid clusters obtained at Markov inflation values I = 1.10, 1.20, 2.00, 3.00, 4.00 and 5.00 is shown in Figure 3. Small and medium-sized clusters of a given size tend to be less numerous as I decreases across this range.

Characterisation of hybrid clusters from multi-genome protein data

In the context of our research on the evolutionary diversification of microbial genomes [1,2], the purpose of protein clustering is to generate sets of orthologs [21] that can be taken forward into subsequent analysis steps (multiple sequence alignment, phylogenetic inference, and topological comparison of subtrees). Protein families represented exactly once in each genome are promising candidates for being both ancient and orthologous. Over the entire hybrid cluster space (*i.e.* all clusters at all thresholds examined, from 1.01 through 0.01) for the 328577 proteins in these 114 microbial genomes, there are 18 clusters ($n \geq 4$) of size 114 in which all 114 genomes are represented; 5 clusters of size 113 in which 113 genomes are represented; and 3 of size 112 representing 112 genomes. Twenty-four of these 26 clusters are ribosomal proteins (the other two are phenylalanyl-tRNA synthetase α chain, and an *O*-sialoglycoprotein endopeptidase).

We define a *representative cluster* (RC) as one in which each protein represents a different genome, and a *maximally representative cluster* (MRC) as an RC that cannot grow further (as the S'_{norm} threshold is incremented toward zero) without incurring multiple representation of one of the genomes. The complete distribution of MRCs over these 114 genomes is shown in Figure 4, and the representation of bacterial "phyla" (second-level NCBI categories: [1]) in these MRCs in Figure 5. Not surprisingly, many of the MRCs – particularly the smaller ones – are of relatively limited distribution across bacterial phyla. The range of S'_{norm} values through which an RC is maximal is its *range of maximality*, and is bounded above by its *maximum threshold* and below by its *minimum threshold*. More precisely, the maximum threshold is the lesser of the maximum possible value of S'_{norm} (1.00 in the absence of rounding errors) and the increment just below that at which the cluster ceases to be maximal (because all edges linking one or more of its proteins no longer satisfy the threshold criterion). The minimum threshold is the greater of the minimum possible threshold (here 0.01)

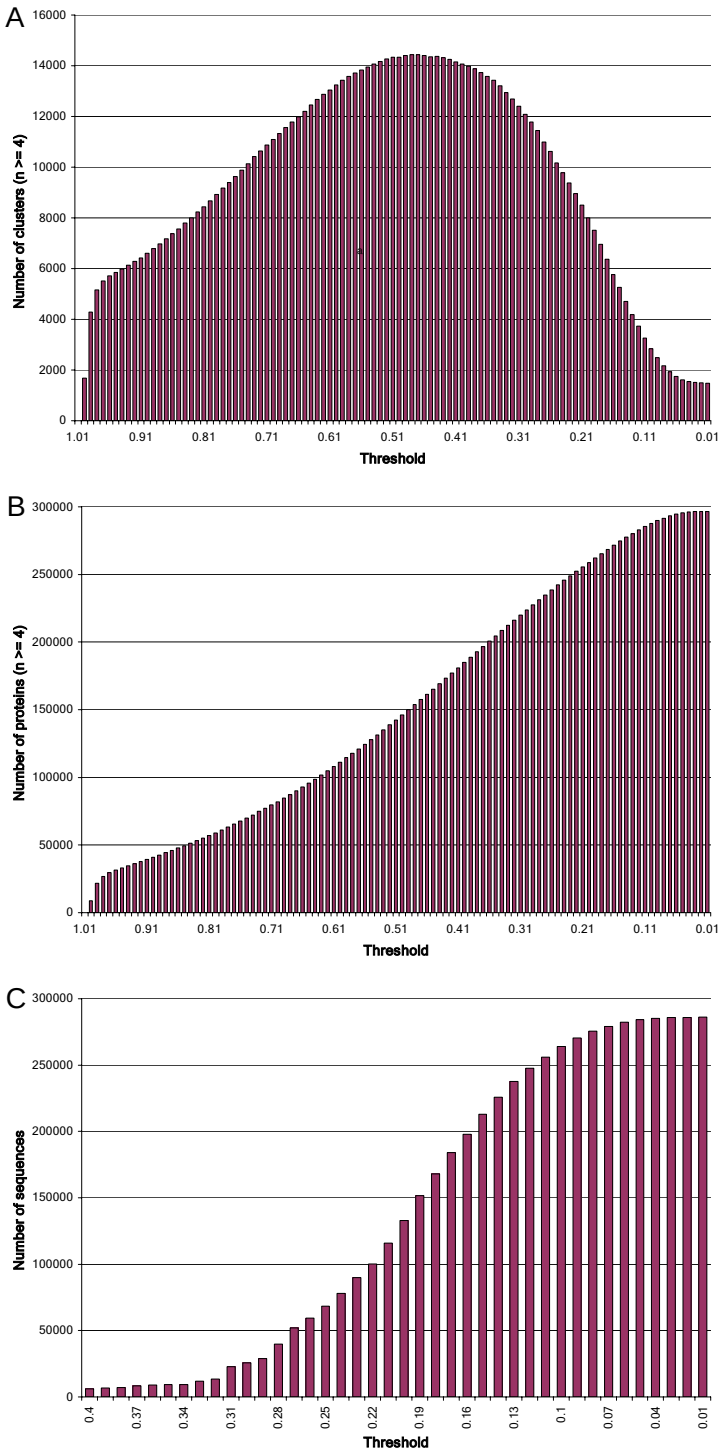

Figure 1

Single-linkage clustering of multi-genome data (a) Number of clusters of $n \geq 4$ members each produced by single-linkage clustering of proteins in 114 microbial genomes (without prior Markov clustering), as a function of S'_{norm} threshold; (b) number of proteins in single-linkage clusters ($n \geq 4$), as a function of threshold; (c) number of proteins in the largest single-linkage cluster, as a function of threshold.

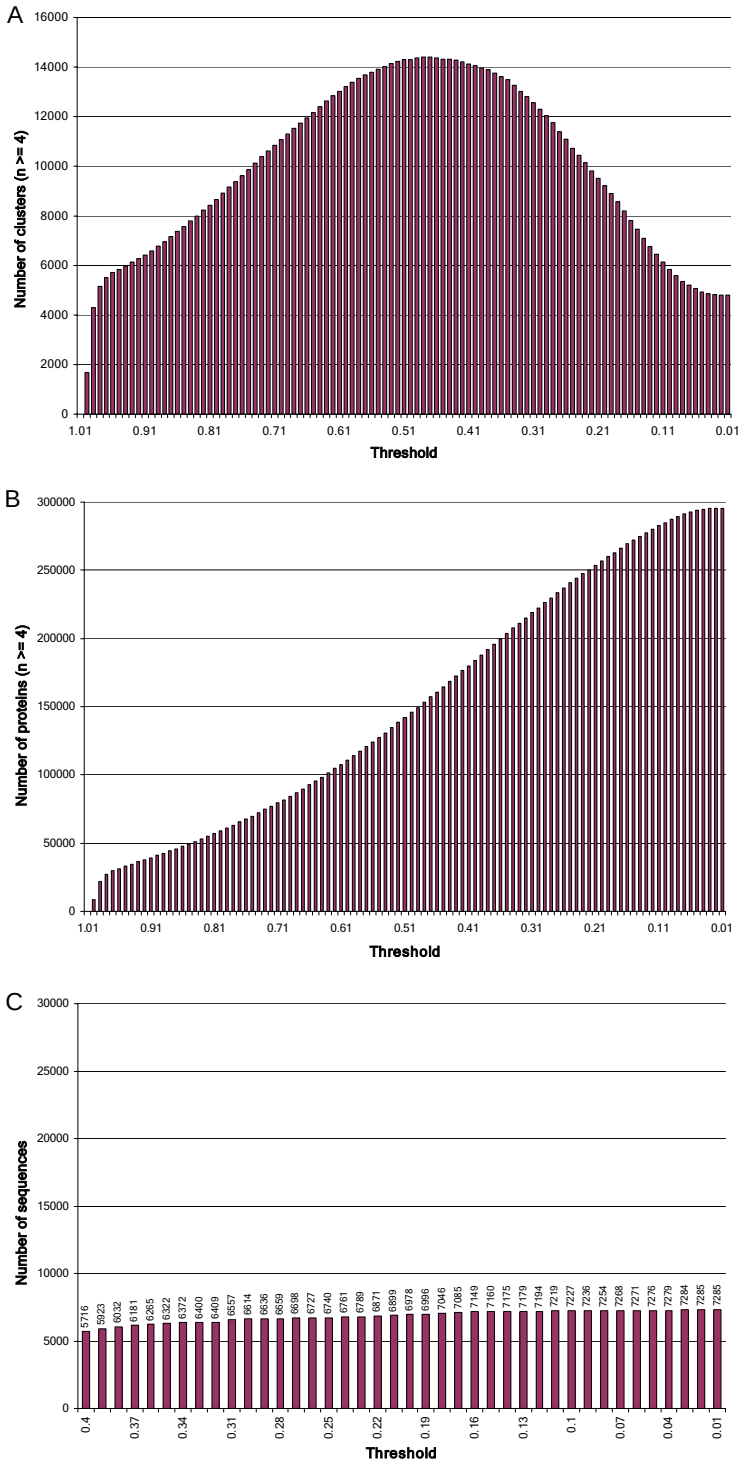

Figure 2

Hybrid clustering of multi-genome data (a) Number of clusters of $n \geq 4$ members each produced by hybrid (Markov followed by single-linkage) clustering of proteins in 114 microbial genomes, as a function of S'_{norm} threshold. Compare the value at the right-most point on the distribution (S'_{norm} 0.01) with that in Figure 1 to see the effect of the prior Markov clustering step; (b) number of proteins in hybrid clusters ($n \geq 4$), as a function of threshold; (c) number of proteins in the largest hybrid cluster, as a function of threshold. Note that the vertical axis is scaled differently than in Figure 1c.

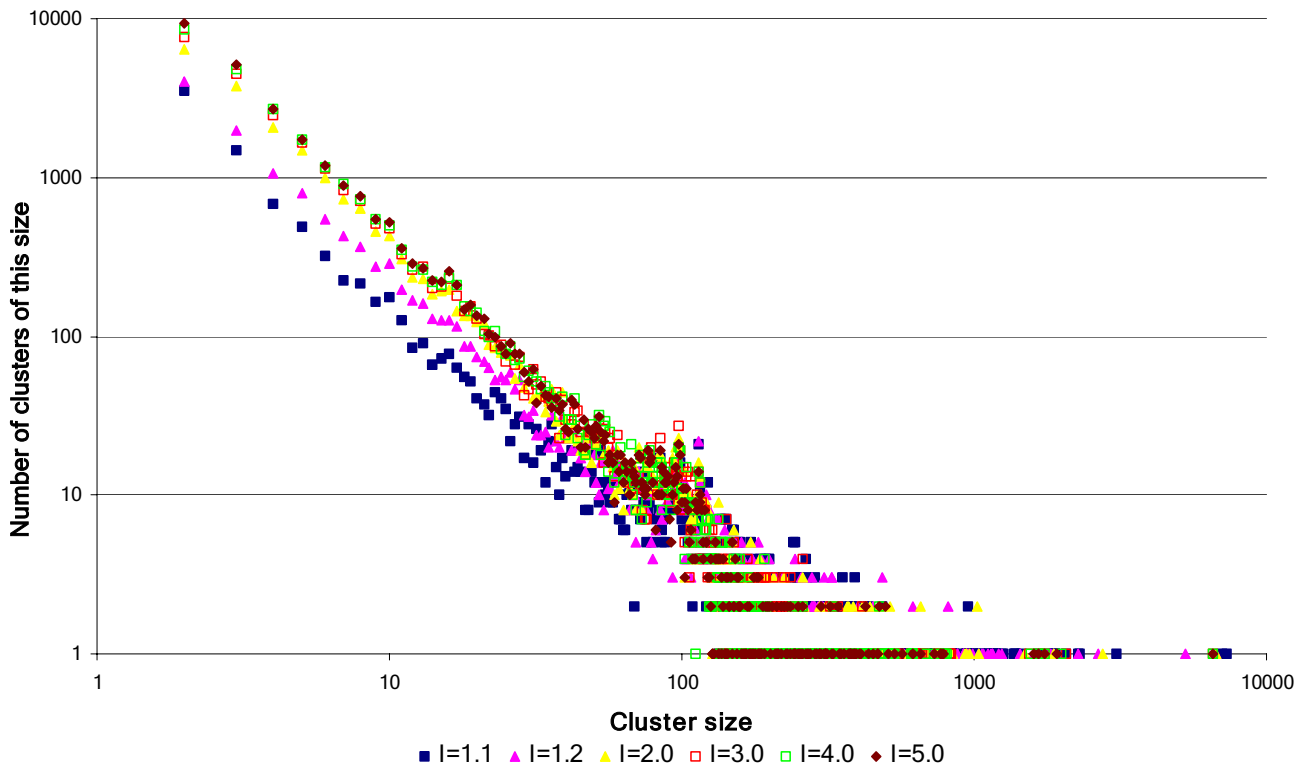

Figure 3
Markov clusters as a function of inflation value Markov clustering of proteins in 114 microbial genomes at six Markov inflation values, showing numbers of clusters as a function of number of proteins per cluster.

and the increment just greater than that at which one or more genomes becomes represented more than once in the cluster. Maximum and minimum thresholds for the MRCs are shown in Figures 6a and 6b respectively, and the distribution of ranges of maximality in Figure 6c. The 13 clusters that are maximally representative over 0.99 threshold units (and represent at least four genomes) include leader and attenuator peptides, a streptolysin S-associated protein, the *Streptococcus* lantibiotic precursor protein, and hypothetical proteins in *Escherichia coli, Salmonella enterica, Shigella flexneri* and *Streptococcus pyogenes*.

Because our cluster data are stored under a relational data model (in our case, in Oracle), these and even more-complex queries – for example, involving successive relaxation of the admittedly strict criterion used here for recognizing an MRC – can easily be made using very short SQL scripts.

Example: ATP synthase paralogous protein families

The superfamily of ATP synthase F1 rotary motor subunit paralogs comprises six families: the α, β, A, B, ρ, and flagellar subunits. Using only SL clustering, each of these six families is fully constituted by S'_{norm} 0.41 and remains intact until S'_{norm} 0.25, when the β and flagellar families join together. The B subunits join them at S'_{norm} 0.24, forming a specific cluster that persists through S'_{norm} 0.22. This transient cluster (from S'_{norm} 0.25 through 0.22) of, at most, three families (Figure 7), represents the full extent of specific clustering among ATP synthase F1 subunit paralogs under SL clustering alone. At S'_{norm} 0.23, the α subunit family is swallowed up into a large (89,840-member) nonspecific cluster ("blob") that grows further to 100,187 members by S'_{norm} 0.22. The ρ subunit cluster accrues 63 unrelated members at 0.22, then at S'_{norm} 0.21, together with the β+flagellar+B cluster and 5590 other proteins, is swallowed up into this blob. The A subunit cluster picks up 130 unrelated proteins at S'_{norm} 0.21 and 0.20, then at S'_{norm} 0.19 is swallowed up into the same blob, which

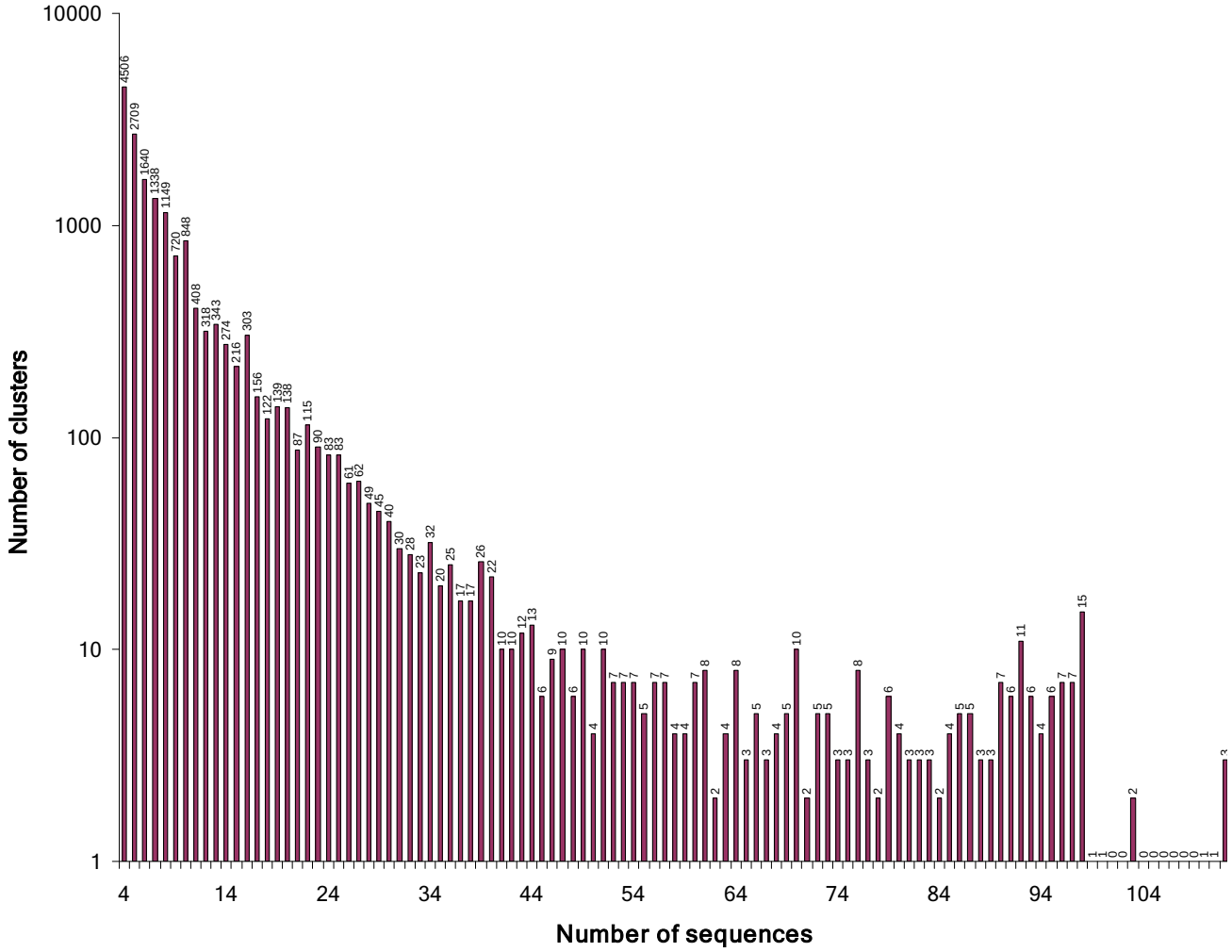

Figure 4
Genome representation in MRCs Numbers of maximally representative clusters of size 4 (the minimum cluster size considered in this work) to 114 (the number of genomes analysed).

then contains the entire ATP synthase F1 superfamily plus 151,112 other proteins (results not shown). Obviously this large, extremely heterogeneous cluster is not useful for phylogenetic inference, or indeed for any other analysis of biological relevance.

Under our hybrid (Markov Clustering plus SL) approach at I = 1.10, the cluster history is identical to that described immediately above for the SL approach down through S'_{norm} 0.24. The 96 α subunits and (separately) the *Methanosarcina acetivorans* C2A predicted protein gi|20090748 join them at 0.21, the 38 A subunits at 0.18, and the 80 ρ subunits at 0.13. The *Bradyrhizobium japonicum* unannotated protein gi|27377965 joins at S'_{norm} 0.06, and the *Agrobacterium tumefaciens* C58 hypothetical protein gi|17938963 at 0.04. This 433-member paralogous cluster is thus fully constituted at S'_{norm} 0.04 (or at 0.13, if the two outliers are spurious matches), and remains cohesive until at least S'_{norm} 0.01, the lowest threshold we examined. The size of the paralogous cluster depends on granularity (*i.e.* on parameterisation of I), and the exact threshold value at which a sequence or group of sequences joins a cluster is to some extent data-dependent; but we observed little or no non-specificity within the range of inflation parameter settings we examined for these data.

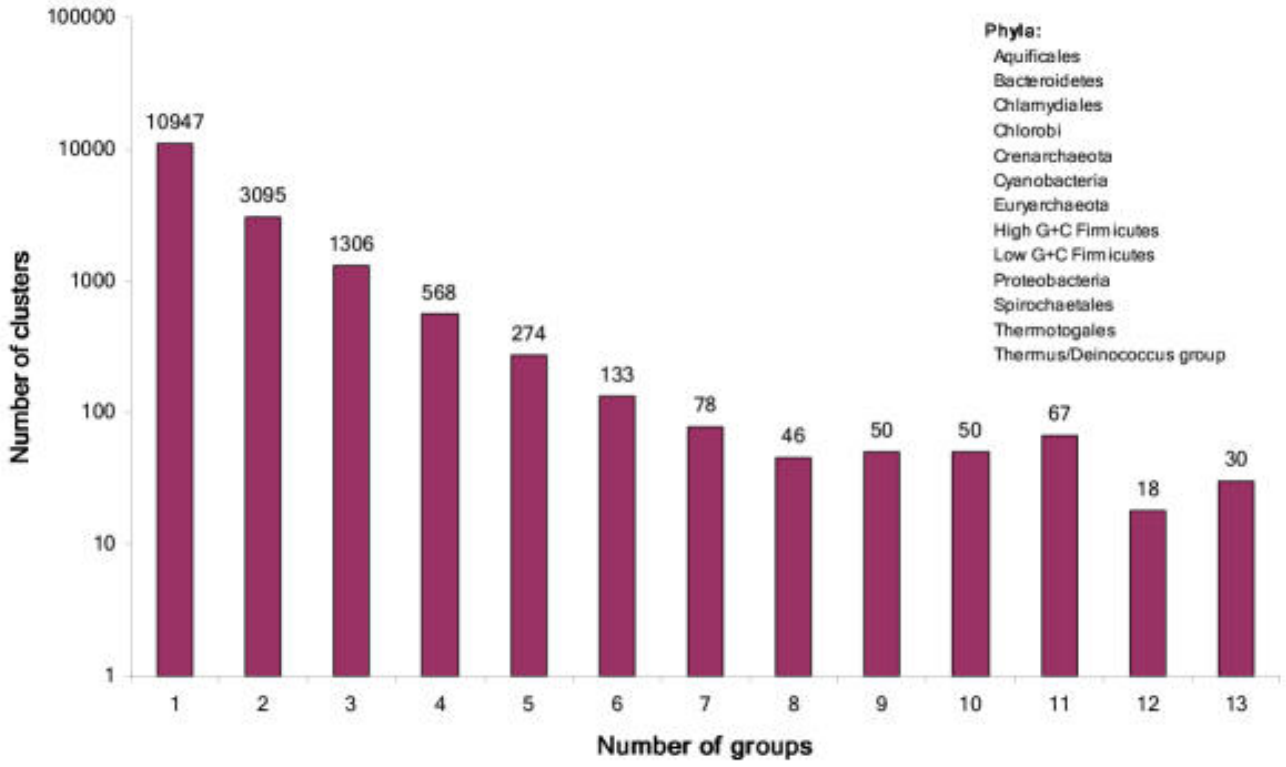

Figure 5
Bacterial phylum representation in MRCs Numbers of maximally representative clusters ($n \geq 4$) as a function of number of bacterial "phyla" (second-order NCBI classications, *e.g.* Aquificales, Bacteriodetes, etc.) represented in each.

Conclusions

The results reported above demonstrate that our hybrid (Markov followed by single-linkage) clustering approach efficiently sorts protein sequences into biologically meaningful clusters that are not accessible by SL clustering alone. The hybrid clusters retain intuitively meaningful topological and ordered edge-length information not immediately available using only MCL, as illustrated specifically by our recovery of all six ATP synthase F1 rotary motor subunit paralogous families, and more globally by the extent of SCOP-to-PDB mapping for protein sequences encoded by 114 microbial genomes.

For the well-behaved subset of PDB we examined, MCL preserves most SCOP domains intact, although the (intrinsically larger) SCOP families can be distributed among several Markov clusters within this range of granularities. Most pure clusters (as defined above) contain a single SCOP family. A very substantial majority of domains within multi-domain clusters, and most families within multi-family clusters, are cleanly resolved during the SL stage of our hybrid method. As the SCOP classification of proteins is based neither on evolutionary principles nor on primary-sequence similarity, it is unrealistic to expect perfect concordance between our hybrid clusters and SCOP families or domains.

Enright *et al.* [4] used a somewhat different approach to test the effectiveness of MCL in clustering protein sequences. Their analysis of the full PDB yielded 1167 families at I = 1.10 and 1761 families at I = 5.00, *i.e.* 50.9% increase, whereas for our subset of PDB we found 831 and 812 $n \geq 2$ families at these inflation values respectively; this suggests that the extra graininess observed by Enright *et al.* reflects the formation of many single-member clusters, which although important in some contexts would not be useful for phylogenetic analysis. These authors reported that 79–87% (depending on the MCL inflation value) of proteins were clustered consistently, as assessed against annotation by domain or domain combination. This validation statistic is not easily commensurate with the (in our opinion) more transparent statistics we present herein (above and Table 1), but was interpreted by Enright *et al.* [4] as indicating that MCL

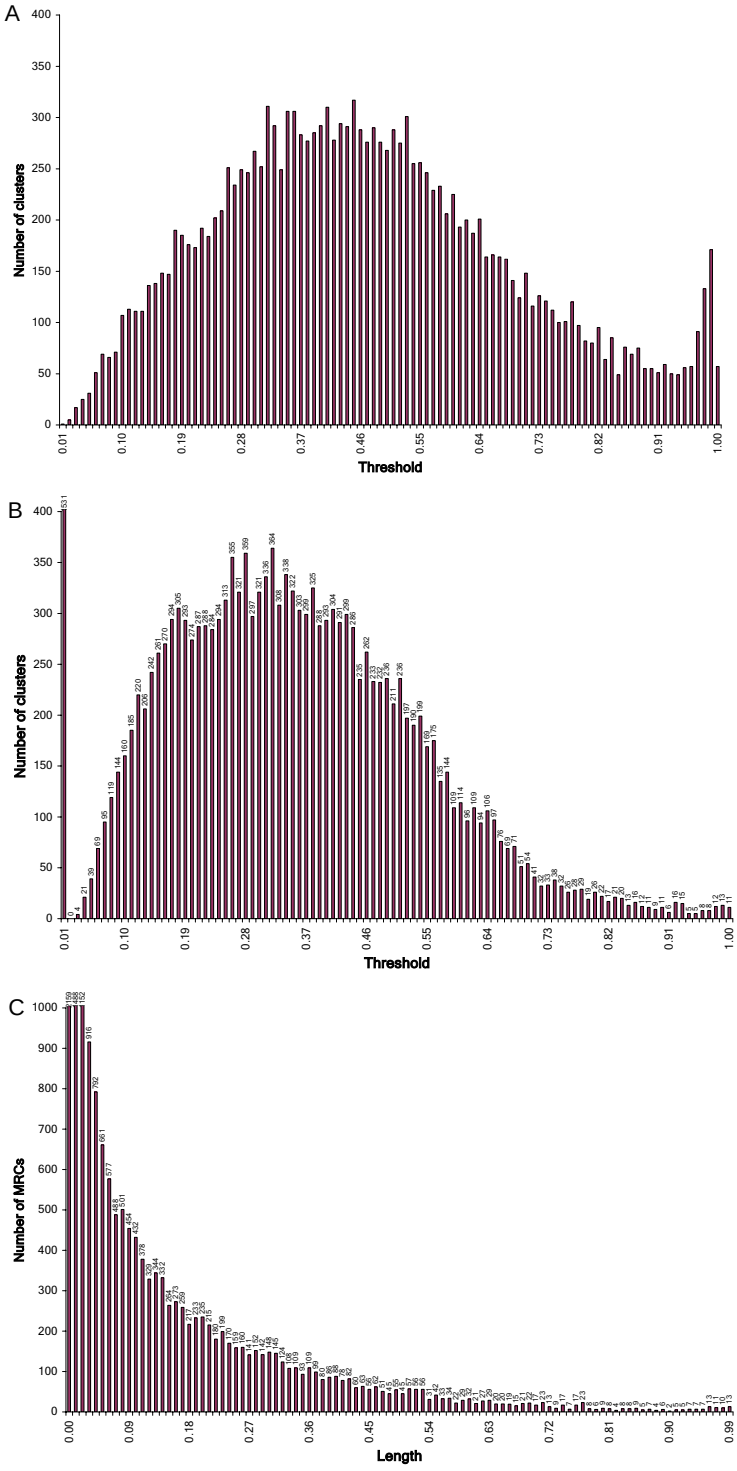

Figure 6

Threshold and range distributions of MRCs (a) Numbers of maximally representative clusters ($n \geq 4$), as a function of maximum threshold expressed as S'_{norm}; (b) numbers of maximally representative clusters ($n \geq 4$), as a function of minimum threshold expressed as S'_{norm}. Note the 1531 MRCs at $S'_{norm} = 0.01$; (c) numbers of maximally representative clusters ($n \geq 4$), as a function of range of maximality (extent along S'_{norm}). The range of maximality of a maximal cluster is the length of the internal edge immediately subtending it.

Figure 7

Clustering of ATP synthase F1 paralog sequences Membership in the ATP synthase F1 cluster, as a function of S'_{norm} threshold. Single-linkage and hybrid clustering gave identical results at $S'_{norm} \geq 0.22$; cluster structure below S'_{norm} 0.22 is for our hybrid method only (see text). NCBI gi numbers are displayed across the top for all F1β subunit sequences, and for three singleton sequences that group with this paralogous family. Large adjacent dots depict clusters at S'_{norm} 1.00, and small adjacent dots show singleton sequences at S'_{norm} 1.00 that are clustered at 0.99.

"accurately and consistently assign(s) proteins into families, despite the fact that this classification relies on structural similarities, which are not all readily detectable at the sequence level". We argue similarly for our results.

For both the 114-genome protein data set in general, and ATP synthase F1 subunits in particular, SL clustering alone fails progressively at thresholds below about S'_{norm} 0.50. This failure is due to non-specific attraction into a large, non-specific cluster. Depending on the data, granularity and probably other factors, more than one such large non-specific cluster can exist fleetingly, but all are soon attracted into a single large "blob" that grows extremely rapidly and eventually takes in most proteins in these genomes. By not allowing edges to be recognised between proteins in different Markov clusters, we prevent the formation of this blob.

Individual Markov clusters may, of course, contain paralogous or, possibly, even non-related sequences; these are resolved into families during the SL step. With the MCL software, the inclusiveness of Markov clusters is determined by the value of the inflation parameter *I*. The range of *I* values accepted by MCL [17] produces only a limited dynamic range of cluster sizes (Figure 3). We hypothesise (based on *e.g.* the ATP synthase example) that as the mean cluster size increases, at least the larger clusters are more likely to contain paralogous proteins. This will be tested

by iterative clustering, alignment and tree inference when our pipeline is in place. We also intend to examine the extent to which the hierarchical course of cluster formation with decreasing S'_{norm} threshold approximates the inferred phylogeny, and whether MRCs tend to be orthologous sets.

At a more algorithmic level, our results illustrate the complementarity between the Markov Clustering algorithm and hierarchical linkage-based clustering in application to these data. MCL operates not so much on the absolute differences among edge lengths, but rather on the overall density structure of the edge-length data [17]. The result is a partitioning of edge space that avoids the formation of non-specific (hence biologically meaningless) groupings, but does not produce the degree of resolution needed to resolve most orthologous protein families. Single-linkage clustering imposes an absolute view of edge-length differences that works with precision locally, but fails globally. MCL thus provides SL with the local environments to which it is best suited, and restricts it from the global context in which it can fail.

We have demonstrated that our hybrid approach can be implemented efficiently in conjunction with a relational database structure, with results saved automatically and queries conducted using generic SQL commands. The hybrid method is fast, and is appropriate for problems where remote homologs are not needed or wanted. It has already proven valuable as part of an automated inference pipeline for studying patterns of vertical and lateral gene transfer among microbial genomes.

Methods

PDB, SCOP families and SCOP domains

Protein Data Bank (25 February 2003) contains 17187 PDB IDs, of which 14548 have only one sequence entry and 2639 have ≥ 2 entries (*e.g.* the ribosome). PDB contains 13764 sequence entries, of which 11353 have only 1 ID each and 2411 have ≥ 2 IDs (*e.g.* lysozyme under different crystallisation conditions). There are 8180 entries (and IDs) in the intersection of the 14548-entry and 11353-ID sets above, and 7555 of these have a SCOP annotation. Of these 7555, there are 6435 annotated with exactly one SCOP family each, and 6430 with exactly one SCOP domain each; the 6430 are a fully contained subset of the 6435. When using SCOP families (clusters) to interpret the clustering of PDB, we consider only these 6435 (or 6430) to ensure one-to-one mappings among PDB entries, PDB IDs, and SCOP families (or domains). Family and domain information was obtained from the parseable file dir.cla.scop.txt version 1.63 available at http://scop.mrc-lmb.cam.ac.uk/scop/parse/index.html.

Multi-genome database

Our genomic dataset consisted of 328577 proteins from the 114 fully sequenced microbial genomes publicly available (NCBI) in May 2003 [see Additional file 2]. Sequences were stored in an Oracle 9i database environment on a Sun E450 cluster, and indexed by gi number. Queries (including those returning all statistics presented herein) were presented to the databases using generic SQL.

All-vs-all BLASTp

All-*versus*-all BLASTp [10] used NHGRI blastall with default settings, including low-complexity filtering, except that word size was set to 2. All BLASTp output was saved, but only output (bit scores) corresponding to matches for which expectation $e \leq 10^{-3}$ was used throughout this work (*i.e.* for analyses of PDB, multi-genome data and individual protein datasets by single-linkage, Markov and hybrid clustering).

Edge lengths

Relatedness between each protein pair (*a*, *b*) was calculated as follows: with *a* as query and *b* as target, we normalised [22] the BLASTp bit score S'_{ab} by dividing by S'_{aa}; then for *b* as query and *a* target, we normalised S'_{ba} by dividing by S'_{bb}. The relatedness of *a* and *b* is defined as $\max(S'_{ab}/S'_{aa}, S'_{ba}/S'_{bb})$. Thus each edge recognised as present (*i.e.* having $e \leq 10^{-3}$ in at least one direction) is assigned a single scalar value (S'_{norm}) between 0 and 1. Due to rounding, scores can slightly exceed 1. Edge values were also stored in the Oracle database at three decimal places precision.

ATP synthase subset

We analysed the clustering of ATP synthase F1 subunit paralogs as a function of threshold [see Additional file 1] using the 114-genome dataset (Results, and Figure 7). Orthology and paralogy were assessed manually based on annotations available in public genome databases (*i.e.* as supplied by the various genome projects); a more rigorous approach would use phylogenetic analysis (the goal of our automated pipeline project).

We also used the ATP synthase F1 dataset to establish the default setting of the MCL inflation parameter *I*. Working with an earlier, 84-genome (235970 protein) version of this dataset, by string search of annotation lines we found the largest SL cluster that contained only proteins annotated as homologs of ATP synthase rotary motor subunits. This contained F1 ATPase α (64 proteins) and β subunits (66); archaeal/vacuolar ATPase A (32) and B (24); bacterial flagellar assembly ATPase subunits (53); and termination factor ρ (56) (total, 295 proteins). We added to this cluster all 11083 proteins that match one or more of these 295 at S'_{norm} 0.20, and carried out MCL at values of *I*

between 1.00 and 5.00. Based on these results, we selected a default value (I = 1.10) small enough to group all the paralogous subunits of ATP synthase F1 within a single cluster, but large enough to avoid the extremely long computation times required for inflation values nearer 1.00.

Single-linkage clustering in a relational environment

Single-linkage clustering was initiated with the most-similar matches (those with $S'_{norm} \geq 1.01$, as the maximum S'_{norm} observed was 1.012) and proceeded in 0.01 stepwise intervals to S'_{norm} 0.01. Clusters were recognised, and allowed to grow and/or merge with others, as relaxation of S'_{norm} threshold progressively caused additional proteins (vertices) to join through valid matches (edges). The S'_{norm} threshold values we refer to must be understood in context. Recall that we store normalised edge length data to three decimal places precision, but step through these data in increments of 0.01. If two proteins (or groups of proteins) are in the same cluster at (for example) S'_{norm} 0.465 but in different clusters at S'_{norm} 0.466, we could say that they "join at S'_{norm} 0.46" or that they "split at S'_{norm} 0.47". The sets of dual vertical lines in Figure 7 are intended to convey this nuance.

Our automated phylogeny inference pipeline is being implemented over a relational environment for reasons well beyond the scope of work presented in this paper. We therefore implemented SL clustering over Oracle 9i to facilitate data coordination with the larger project. We processed (at each threshold) the ordered S'_{norm} edge data, writing a series of new cluster-state information (membership) tables. We do not store topology tables *per se*, but reconstruct topologies from membership plus edge (S'_{norm}) data. After the list is processed, clusters (graphs) are labelled in ascending order of S'_{norm}. Graphs present at S'_{norm} 0.01 are arbitrarily numbered 1, 2, etc. and as these graphs fragment with stepwise increase in S'_{norm}, the labels are extended. Thus if the graph labelled 1 fragments at a higher value of S'_{norm}, the fragments (having $n \geq 2$ members) are arbitrarily labelled 1.1, 1.2 etc. Similarly, if 1.1 fragments further at a higher S'_{norm}, its daughters ($n \geq 2$) are labelled 1.1.1, 1.1.2, etc.

Markov clustering

The Markov Cluster algorithm was implemented using MCL (available from http://micans.org/mcl) and was applied with inflation parameter typically set to I = 1.10, and with other parameters at default values. Cluster membership was stored in ordered Oracle tables as described above. With PDB we clustered all sequences, but subsequently used only the 6435 (or 6430) entries/IDs identified above.

Hybrid clustering

Our hybrid clustering method was carried out in two stages, as follows. First, we processed the entire dataset (valid edges with $S'_{norm} \geq 0.01$) with MCL, yielding Markov clusters (ordered tables). We then conducted SL clustering as above, but on only those edges that have both ends (proteins) within *the same* Markov cluster; edges that span Markov clusters (*i.e.* link proteins in different Markov clusters) were ignored. We again wrote tables of clustered proteins at each threshold, and backtracked to label graphs.

Cluster purity, family/domain distribution, and family/domain splitting

We define a Markov cluster to be *pure* if it contains only SCOP families (or domains) in their entirety. A pure cluster can contain multiple SCOP families (or domains) so long as each family (or domain) is contained in its entirety. If any member of any of the families (or domains) is external to that cluster, the cluster is not pure. SCOP families (or domains) are *distributed over multiple clusters* if members of that family (or domain) are found in more than one cluster ($n \geq 2$). A SCOP family (or domain) is *split* if one or more, but not all, of its members occurs in a cluster ($n \geq 2$) together with at least one member of at least one different family (or domain). The non-included members of the split family (or domain) do not need to be included in a cluster.

Authors' contributions

JPG and MAR have longstanding interests in lateral gene transfer. JPG suggested the use of Markov clustering, and provided the initial ATP synthase F1 datasets. TJH conducted all computational analyses, wrote programs and scripts, managed our relational database, and initiated specific questions. MAR provided conceptual design and wrote the manuscript. All authors contributed to data analysis.

Acknowledgements

We thank the anonymous referees for constructive comments. We acknowledge support of ARC grants DP0342987 and CE0348221. JPG thanks The University of Queensland for a travel grant.

References

1. Ragan MA, Charlebois RL: **Distributional profiles of homologous open reading frames among bacterial phyla: implications for vertical and lateral transmission.** *Intl J Syst Evol Microbiol* 2002, **52:**777-787.
2. Raymond J, Zhaxybayeva O, Gogarten JP, Gerdes SY, Blankenship RE: **Whole-genome analysis of photosynthetic prokaryotes.** *Science* 2002, **298:**1616-1620.
3. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV: **The COG database: new developments in phylogenetic classification of proteins from complete genomes.** *Nucl Acids Res* 2001, **29:**22-28.
4. Enright AJ, Van Dongen S, Ouzounis CA: **An efficient algorithm for large-scale detection of protein families.** *Nucl Acids Res* 2002, **30:**1575-1584.

5. Heger A, Holm L: **Exhaustive enumeration of protein domain families.** *J Mol Biol* 2003, **328:**749-767.
6. Park J, Teichmann SA: **DIVCLUS: an automatic method in the GEANFAMMER package that finds homologous domains in single- and multi-domain proteins.** *Bioinformatics* 1998, **14:**144-150.
7. Servant F, Bru C, Carrere S, Courcelle E, Gouzy J, Peyruc D, Kahn D: **ProDom: automated clustering of homologous domains.** *Brief Bioinform* 2002, **3:**246-251.
8. Yona G, Linial N, Linial M: **ProtoMap: automatic classification of protein sequences and hierarchy of protein families.** *Nucl Acids Res* 2000, **28:**49-55.
9. Krause A, Stoye J, Vingron M: **The SYSTERS protein sequence cluster set.** *Nucl Acids Res* 2000, **28:**270-272.
10. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucl Acids Res* 1997, **25:**3389-3402.
11. Sokal RR, Sneath PHA: **Principles of Numerical Taxonomy.** *London: Freeman* 1963.
12. Pearson WR, Lipman DJ: **Improved tools for biological sequence analysis.** *Proc Natl Acad Sci USA* 1988, **85:**2444-2448.
13. Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D: **Detecting protein function and protein-protein interactions from genome sequences.** *Science* 1999, **285:**751-753.
14. Smith TF, Zhang X: **The challenges of genome sequence annotation or "The devil is in the details".** *Nat Biotechnol* 1997, **15:**1222-1223.
15. Karplus K, Barrett C, Hughey R: **Hidden Markov models for detecting remote protein homologies.** *Bioinformatics* 1998, **14:**846-856.
16. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer ELL: **The Pfam Protein Families Database.** *Nucl Acids Res* 2002, **30:**276-280.
17. van Dongen S: **Graph clustering by flow simulation.** *PhD thesis. University of Utrecht* 2000 [http://micans.org/mcl/lit/svdthesis.pdf.gz].
18. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank.** *Nucl Acids Res* 2000, **28:**235-242.
19. Lo Conte L, Brenner SE, Hubbard TJP, Chothia C, Murzin AG: **SCOP database in 2002: refinements accommodate structural genomes.** *Nucl Acids Res* 2002, **30:**264-267.
20. Stock D, Leslie AGW, Walker JE: **Molecular architecture of the rotary motor in ATP synthase.** *Science* 1999, **286:**1700-1705.
21. Fitch WM: **Aspects of molecular evolution.** *Annu Rev Genet* 1973, **7:**343-380.
22. Bansal AK, Bork P, Stuckey PJ: **Automated pair-wise comparisons of microbial genomes.** *Math Modelling Sci Comput* 1998, **9:**1-23.