Graph-theoretic modelling of the domain chaining problem[⋆]

Poly H. da Silva[1], Simone Dantas[1], Chunfang Zheng[2] and David Sankoff[2]

[1] Institute of Mathematics and Statistics - Fluminense Federal University, Brazil
[2] Department of Mathematics and Statistics - University of Ottawa, Canada

Abstract. Methods for the clustering of genes into homologous families (sets of genes descending from a single gene in an ancestral organism) are susceptible to the inappropriate merging of unrelated families, called domain chaining. We give formal criteria for the chaining effect by defining multiple alternative clique relaxation and path relaxation models and the relationships among them, involving different graph characteristics. We implement these definitions and apply them to 45 flowering plant genomes in order to compare the Markov Cluster Algorithm (MCL) and Soft Cliques with Backbones (SCWiB) clustering method. In the process we note the extreme behavior of the *Amborella trichopoda* genome.

1 Introduction

A gene family is a set of genes, in one genome or several, that includes all descendants of a single gene in an ancestral organism. The genes in a family are called homologous. The goal of gene family classification is to partition a set of sequences into homologous families. In practice, gene families are constructed on the basis of DNA or protein sequence similarities under the assumption that genes in the same family will retain more sequence similarity than unrelated genes. Many methods are currently available for the clustering of genes into families. However these methods are susceptible to the inappropriate merger of unrelated families, due to the multiple domain structure of many proteins. Some domains, more or less lengthy sequence fragments, recur in many different families with largely distinct histories and functions, blurring the boundaries between these families. This problem is called the *domain chaining effect* [4, 5]; it stems from the evolutionary acquisition of widespread protein modules that may help in the binding or movement of the protein but generally not its specific primary enzymatic, synthetic, signaling or regulatory function.

Gene family classification has often been studied using graph concepts [3, 13, 14]. A theoretical model of this problem can be obtained in the following way: each gene is identified with a vertex v of undirected graph $G_S = (V, E)$, where there exists a edge $\{u, v\} \in E$ between two vertices u and v if the pair of genes exceeds a threshold similarity score. Graph G_S is called a *similarity graph*;

[⋆] Partially supported by FAPERJ, CNPq, CAPES and NSERC. DS holds the Canada Research Chair in Mathematical Genomics.

ideally each maximal clique in G_S corresponds to a single gene family, and vice versa, and a long path may represent a chaining effect.

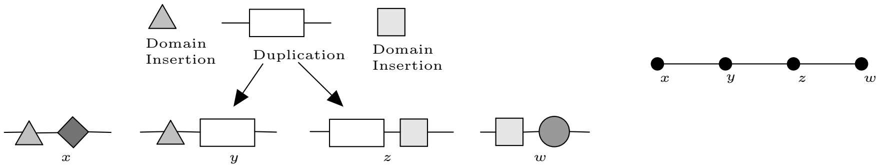

Fig. 1. The evolutionary history of a hypothetical multidomain family showing both gene duplication and domain insertions. Genes y and z share a common ancestor but do not have identical domain composition. Genes x and w share homologous domains with genes y and z, respectively, but there is no gene that is ancestral in all of x, y, z and w.

There is a large literature on clique relaxation models, where not all the elements are "directly connected" to each other. These models are useful in applied contexts where connections between members of a group need not be direct and could be meaningfully accomplished through intermediaries. Clique relaxation models are obtained by allowing the clique property to be relaxed in various ways. Some examples are: *s-clique* — where the distance between vertices within the group must be at most s [6]; *s-club* — where the diameter of the graph induced by the group must be at most s [2]; *s-plex* — where the number of non-neighbors among elements of the group is bounded [11]; *k-core* where a certain minimum number k of neighbors within the group is guaranteed [10]; *s-defective clique* — which differs from a clique by at most s missing edges [12]; γ*-quasi-clique* — ensures a certain minimum ratio γ of the number of existing links to the maximum possible number of links within the group [1].Although clique relaxation models have proved useful in many applications, there has been no formal study of domain chaining associated to these methods.

In this present paper, we also give formal criteria for the chaining effect in terms of a number of alternative *path* relaxation models and the relationships among them, involving different graph characteristics. We define the α*-quasi-path* that ensures a certain minimum ratio α of the diameter to number of existing links; *k-chain* that is relative to average degree; and (x, y)*-damaged-path* that takes into account the number of missing and extra edges to be dealt with in order to turn the graph into a path.

We use these cluster and path relaxation definitions in comparing two methods for generating gene families: *Markov Cluster Algorithm (MCL)* [3], one of the most widely used procedures for inferring gene families; and *Soft Cliques With Backbones (SCWiB)* [14], a new method that ensures that clusters satisfy a tolerant edge-density criterion that takes into account cluster size. We perform the comparisons on 45 published angiosperm genome sequences.

In Section 2 we give some basic graph theory definitions and formalize our proposed new path relaxation definitions. In Section 3 we implement the definitions and apply them to 45 genomes in order to compare the MCL and SCWiB methods. Finally, in Section 4, we summarize our results.

2 Definitions and notations

We denote a simple undirected graph by $G = (V, E)$, where V is the set of vertices and E is the set of edges. Given a vertex $v \in V$, we denote the degree of v by $d(v)$ and the minimum degree of G by $\delta(G)$. A clique of a graph is a set of mutually adjacent vertices. A path of length r between vertices u and v in G is a subgraph of G defined by an alternating sequence of distinct vertices and edges $u \equiv v_0, v_1, ..., v_{r-1}, v_r \equiv v$ such that $\{v_i, v_{i+1}\} \in E$ for all $1 \leq i \leq r-1$. We denote by P_n the graph that is a path with n vertices. The distance $dist(u, v)$ between two vertices u and v of a connected graph is the length of the shortest path connecting them. The eccentricity of a vertex v, denoted by $\varepsilon(v)$, in a connected graph G, is defined to be the maximum distance between v and any other vertex u of G. Then we say the diameter of G, $diam(G)$, is the maximum value of $\varepsilon(v)$ over all vertices $v \in V$. The density $\rho(G)$ of G is the ratio of the number of edges to the total number of possible edges, i.e., $\rho = \frac{2|E|}{|V|(|V|-1)}$.

Next, some of clique relaxation models, which were already mentioned in the previous section, are formally defined. We assume that $G = (V, E)$ is connected, $\gamma \in (0, 1]$ is real and the constant s is positive integer.

Definition 1. (s-plex)*. G is an s-plex if* $\delta(G) \geq |V| - s$.

Definition 2. (γ-quasi-clique)*. G is a* γ*-quasi-clique if* $\rho(G) \geq \gamma$.

Definition 3. (s-defective-clique)*. G is a s-defective-clique if G contains at least* $\frac{|V|(|V|-1)}{2} - s$ *edges.*

In order to study the domain chaining effect, we introduce some path relaxation definitions, each of which measures, in some sense, how close a graph is to being a path. We assume that $G = (V, E)$ is connected, constants k, x and y are positive integers and $\alpha \in (0, 1]$ is real.

Definition 4. (α-quasi-path)*. G is an* α*-quasi-path if* $\frac{diam(G)}{|E|} \geq \alpha$.

Definition 4 ensures a certain minimum ratio α of the diameter to the number of edges. Note that this definition is more pertinent to chaining than the definition of an s-club, since in addition to considering the diameter, it also considers the total number of edges and the ratio between them.

Observe that for $\alpha = 1$, G is an α-quasi-path if and only if it is a path, and that the minimum value for α is $\frac{2}{|V|(|V|-1)}$, which occurs when G is a complete graph. If G contains an α-quasi-path, where α is close to 1, the graph is highly chained, i.e., similar in structure to a path.

Definition 5. (k-chain). *G is a k-chain if k is the smallest integer such that $k \geq \frac{\sum_{v\in V\setminus\{u,w\}} d(v)}{|V|-2} = \frac{2|E|-d(u)-d(w)}{|V|-2}$, where u and w are the vertices of smallest degree and $d(u) \leq d(w) < k$.*

Definition 5 ensures that the average degree of a graph G, without the two vertices of smallest degree, is less than or equal to k. Furthermore, as $\delta(G) < k$ if G is a k-chain, all k-chain graphs are at most $(k-1)$-connected. Note that all p-regular graphs are $p+1$-chains. In particular a complete graph is a $|V|$-chain. Thus $2 \leq k \leq |V|$, and when k approaches 2, the graph is highly chained (it is structurally similar to a tree).

Definition 6. ((x, y)-damaged-path). *G is an (x, y)-damage-path if $x = |V| - 1 - diam(G)$ and $y = |E| - diam(G)$.*

Definition 6 involves two parameters (x, y). The former is the difference between the length of $P_{|V|}$ and the diameter of G; the latter is the difference between the number of edges and the diameter of G. Here the idea is, given a graph G and a path P in G with length equal to the diameter of G, y represents the number of edges that are not in P (extra edges), and x indicates the number of edges needed to complete P (missing edges) in order to obtain a path of length $|V| - 1$. Note that x and y could be defined in different ways, as long as the difference $y - x = |E| - |V| + 1$ always remained the same; it is convenient here to define x and y in terms of the diameter of G so that $x = y = 0$ if and only if G is a path. In this case, we have $0 \leq x \leq |V| - 2$ and $0 \leq y \leq \frac{|V|(|V|-1)}{2} - 1$. Note that $x = |V| - 2$ and $y = \frac{|V|(|V|-1)}{2} - 1$ if G is a complete graph. Thus, given two graphs G_1 and G_2, such that G_1 is (x_1, y_1)-damaged-path and G_2 is (x_2, y_2)-damaged-path, we say that G_1 is more chained than G_2 if $y_1 - x_1 < y_2 - x_2$, or $y_1 - x_1 = y_2 - x_2$ and $x_1 < x_2$. Some examples are depicted in Figure 2.

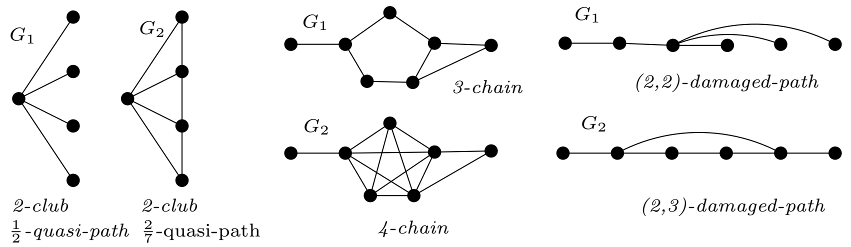

Fig. 2. Examples of path relaxation models, where G_1 is more chained than G_2 according to the respective definition.

We will compare two methods for generating gene families, SCWiB and MCL, in terms of the parameters we have defined.

The SCWiB method ensures that a gene family is determined by strong similarities connecting each of its members, by setting a high similarity threshold U, and requiring that a cluster be connected, in the graph theoretical sense, solely in terms of similarities exceeding U. Also to control chaining, this method sets a less stringent threshold W, and requires that the elements in the cluster form an s-plex in terms of similarities exceeding W.

MCL is one of the most widely used methods for inferring gene families. Its basic principle is the iteration of a procedure that strengthens certain heavily weighted edges and weakens those with lesser weight. With appropriate parameter settings, MCL and SCWiB can produce very similar distributions of cluster sizes. The lack of any cluster quality criterion influencing the MCL process, however, results in many of its clusters, including some of the largest ones, having very few internal edges, while the SCWiB construction explicitly prohibits this.

3 Results

Data source In order to compare the SCWiB and MCL methods, we calculate the parameters introduced in Section 2 in 45 genomes. We extracted the data on these genomes from the CoGe database [7, 8]. We require genomes to be published, publicly available, and have associated structural gene annotations. The genomes include *Amborella*, sacred lotus, rice, *Brachypodium*, maize, sorghum, millet, banana, duckweed, date palm, grape, eucalyptus, clementine, sweet orange, cacao, papaya, *Arabidopsis thaliana*, *Arabidopsis lyrata*, turnip, *Capsella rubella*, *Leavenworthia alabamica*, *Sisymbrium irio*, *Aethionema arabicum*, *Thellungiella parvula*, *Eutrema parvulum*, watermelon, cucumber, peach, strawberry, lotus, common bean, pigeonpea, soybean, coffee, poplar, flax, cassava, *Ricinus communis*, kiwifruit, tomato, potato, pepper, *Utricularia*, *Mimulus* and *Medicago* [15-56].

We analyze these two methods for each genome individually, calculating the average of each parameter (diameter, α, k, (x, y)) separately, for clusters with $|V|$ in each the following bins: 2, 3-4, 5-8, 9-16, 17-24, 25+. We consider each parameter as a function of bin size.

Comparison of clustering methods In Figures 3 and 4, we note that, in general, the diameter is a non-decreasing function for both methods and, in the SCWiB method it is always bounded above by 2 (by definition, given our choice of parameters). Furthermore, for all the genomes analyzed the average diameter of SCWiB clusters with the same $|V|$ is always less than that of the corresponding MCL clusters.

For α-quasi-paths, we observe that α is a decreasing function and, starting at bin 5-8, α decreases faster for SCWiB than for MCL.

For k-chains, k is a increasing function and, starting at bin 5-8, k increases in the SCWiB clusters faster than in those obtained by MCL. Therefore for the

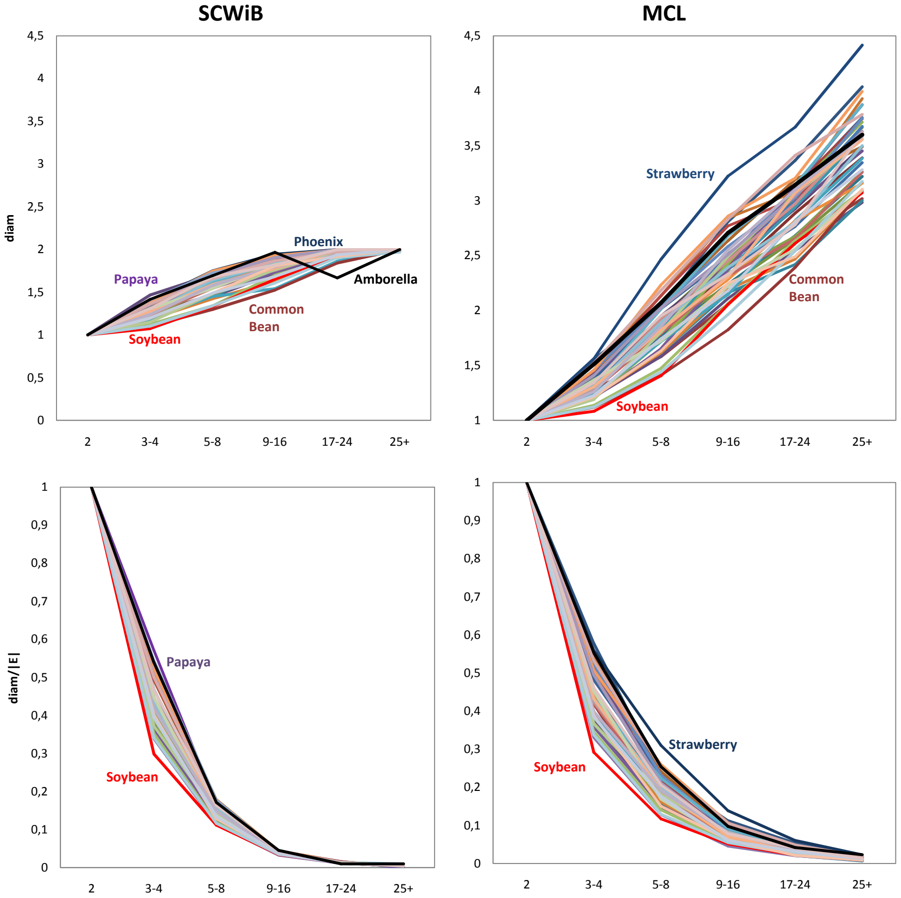

Fig. 3. Average of diameter (top) and α(bottom) separately for clusters with $|V|$ in each the following bins: 2, 3-4, 5-8, 9-16, 17-24, 25+. Singletons not included. Left: the SCWiB. Right: MCL. *Amborella* results highlighted in black.

parameters "diameter", α and k, we find that MCL leads to more chaining than the SCWiB method.

Turning to the relaxation clique criteria, γ-quasi-clique and s-defective-clique, we also compare the similarity to cliques of the gene families generated by both the two clustering processes. We observe that SCWiB clusters are denser, and more uniform in density, than those of MCL, and the former have fewer missing edges than the latter in comparison with a complete graph. The range, across all genomes, of the average number of missing edges in SCWiB-generated families in every bin, starting at bin 9-16, is less than and actually disjoint from that for MCL method. SCWiB clusters are thus more clique-like than MCL clusters.

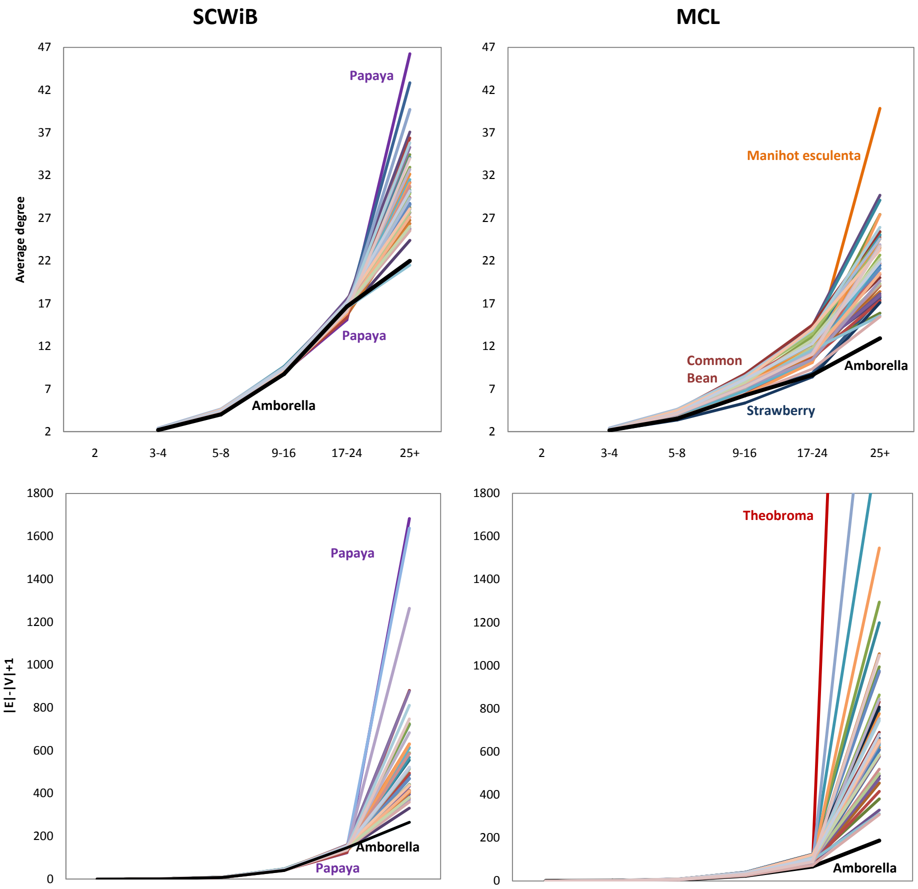

Fig. 4. Average of k (top) and $y - x$(bottom) separately for clusters with $|V|$ in each the following bins: 2, 3-4, 5-8, 9-16, 17-24, 25+. Singletons not included. On the left the SCWiB clusters and in right the clusters from MCL. *Amborella* always highlighted in black.

Comparison of genomes We previously observed that *Amborella trichopeda* has an anomalously small number of moderate and large-sized clusters [14]. Here, in comparing all 45 genomes by both clustering methods, we observe in Figs. 3,4 and 5 that *Amborella* also demonstrates extreme behaviour with respect to high levels of chaining and non-clique-like families. Only papaya and strawberry have comparable behavior, but less consistently. The common bean, soybean and *Theobroma* are at the other extreme, with little chaining and more clique-like clusters.

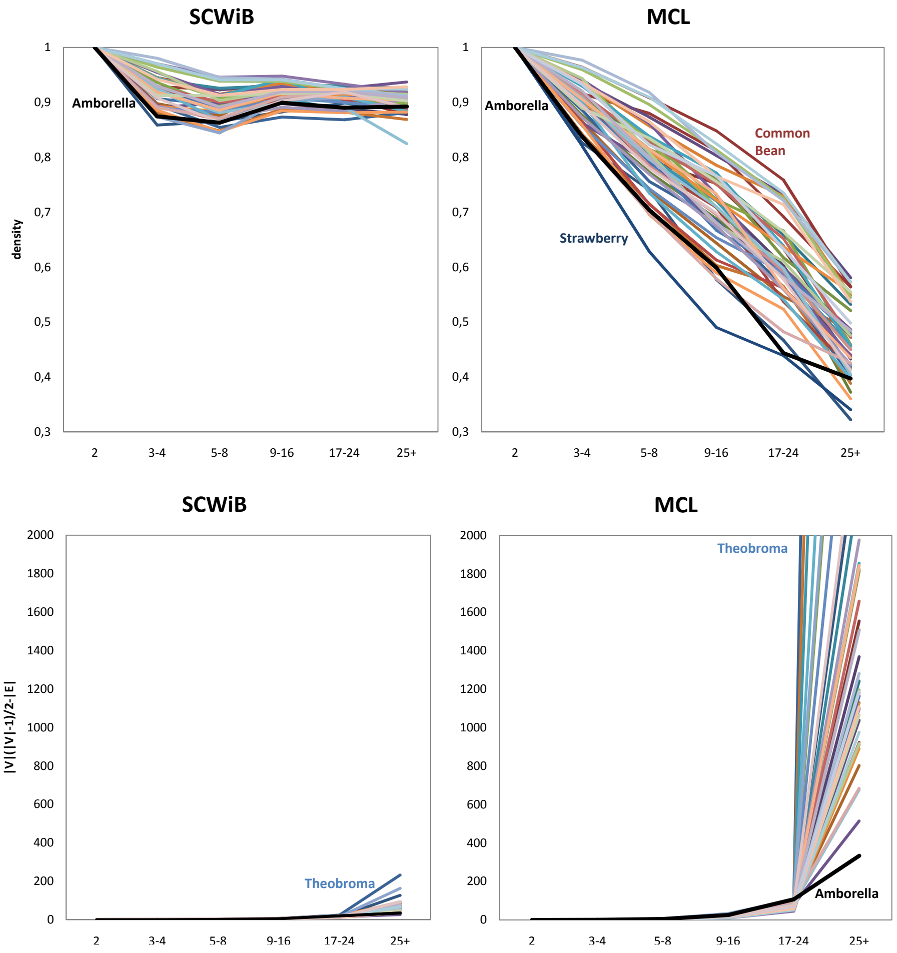

Fig. 5. Average of density (top) and average of missing edges to be complete graph (bottom) separately for clusters with $|V|$ in each the following bins: 2, 3-4, 5-8, 9-16, 17-24, 25+. Singletons not included. On the left the SCWiB clusters and in right the clusters from MCL. *Amborella* always highlighted in black.

Comparison of criteria We computed the Pearson correlation coefficient among the parameters from an array of average values over all 45 genomes, for each bin (Table 1). We did this separately for the SCWiB clusters and the MCL clusters. All pairs of parameters are significantly correlated, either positively and negatively. To find the overall pattern, we submitted the correlations to a multidimensional scaling (MDS) procedure using the XLSTAT package from Addinsoft™. For coherence, we used $-\alpha$ and $-\rho$ instead of α and ρ, so that all the correlations would have the same sign, and larger values of all parameters would indicate increased chaining.

Table 1. Pearson correlation coefficient among: diameter, α, k, $|y-x|$, density ρ and s missing edges.

SCWiB	*diam*	α	k	$\lvert y-x\rvert$	ρ	s	*MCL*	*diam*	α	k	$\lvert y-x\rvert$	ρ	s
diam	1						*diam*	1					
α	-0.94	1					α	-0.84	1				
k	0.84	-0.68	1				k	0.95	-0.75	1			
$\lvert y-x\rvert$	0.64	-0.46	0.96	1			$\lvert y-x\rvert$	0.75	-0.41	0.93	1		
ρ	-0.76	0.93	-0.28	-0.35	1		ρ	-0.99	0.86	-0.97	-0.76	1	
s	0.64	-0.45	0.96	0.99	-0.35	1	s	0.69	-0.37	0.89	0.99	-0.72	1

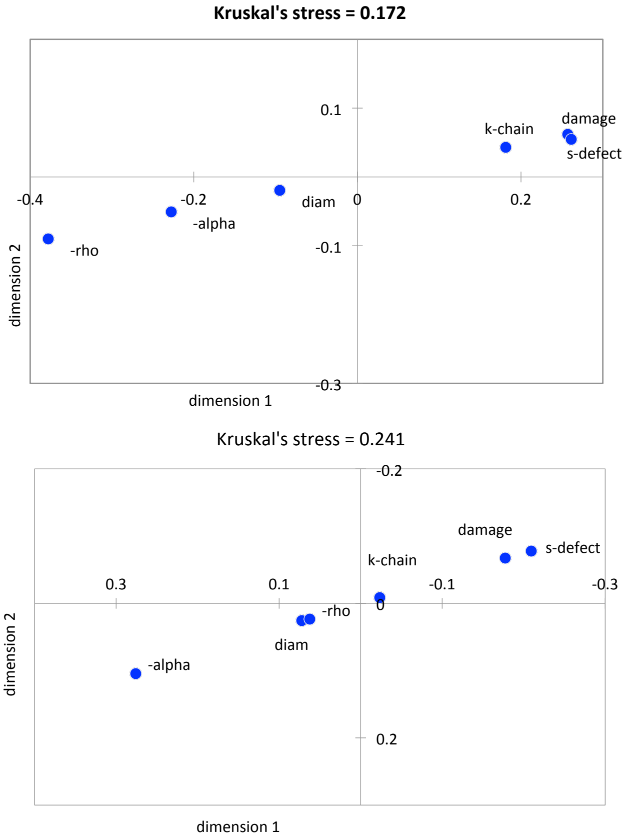

Fig. 6. MDS analysis of correlations among clique- and path-relaxation criteria for chaining. Top: SCWiB clusters. Bottom: MCL clusters.

Fig. 6 shows that the the parameters are largely disposed in a single dimension, although a two-dimensional space was specified in the scaling settings. $-\alpha$, $-\rho$ and "diameter" are at opposite sides from s and $|y-x|$, while k is intermediate. There are clear differences between SCWiB and MCL, in the ordering of $-\alpha$, $-\rho$ and "diameter", for example, reflecting the constraints on these quantities in their SCWiB definitions.

4 Conclusion

Our application to plant genomes shows that *Amborella* clusters show less chaining than other flowering plants, extending the previous discovery, in the same set of angiosperm genomes surveyed here, of the special nature of this basal flowering plant [14] in having exceptionally few large and moderate-size gene families.

The different uses of clustering in genomics suggest that no one definition is universally useful. For partitioning the set of genes into disjoint gene families, as we have done here, allowing minor relaxation from a clique is probably more appropriate. On the other hand, for the investigation of the evolution of genes through the accumulation, loss and exchange of protein domains, it may be interesting to balance clique-like behaviour with a degree of chaining. For functional studies of gene networks, still other concepts and definitions of cluster shape may be preferable.

Although "chaining" is a generally understood concept in cluster analysis and domain chaining is familiar to all who work on automated gene family construction, there is no one single formal definition of chaining. We have suggested a range of formalizations that turn out to differ (empirically) along an axis measuring relaxation from a clique at one extreme, to relaxation from a path at the other. Our tests show that in general SCWiB yields clusters with less chaining than MCL, not only according to the clique relaxation criteria, but also by the path relaxation ones.

That the clustering criteria are disposed in an almost one-dimensional subspace when applied to our database is more than just an artifact of the clustering method is confirmed by similar results with the two methods. It is also unlikely to reflect only the properties of our database, but this should be confirmed by simulation studies. These observations reinforce our suggestion of more general research into how to operationalize various concepts of cluster shape. We could hope that eventually this would lead to an understanding of the statistical nature of the evolution of gene families.

References

1. J. Abello, M. Resende and S. Sudarsky, Massive quasi-clique detection, In: Rajsbaum, S. (Ed.), LATIN 2002: Theoretical Informatics. Springer-Verlag, London, (2002) pp. 598–612.
2. R.D. Alba, A graph theoretic definition of a sociometric clique, Journal of Mathematical Sociology 3.1 (1973) 113–126.
3. A.J. Enright, S. Van Dongen and C.A. Ouzounis, An efficient algorithm for large-scale detection of protein families, Nucleic Acids Research 30, (2002) 1575–1584.

4. J.M. Joseph, On the Identification and Investigation of Homologous Gene Families, with Particular Emphasis on the Accuracy of Multidomain Families. Lane Center for Computational Biology Technical Report CMU-CB-12-103.pdf (2012).
5. J.M. Joseph, D. Durand, Family classification without domain chaining, Bioinformatics 25.12 (2009) i45–i53.
6. R. Luce, Connectivity and generalized cliques in sociometric group structure, Psychometrika 15, (1950) 169–190.
7. E. Lyons, et al., Finding and comparing syntenic regions among Arabidopsis and the outgroups papaya, poplar and grape: CoGe with rosids, Plant Physiology, 148, (2008) 1772–1781.
8. E. Lyons and M. Freeling, How to usefully compare homologous plant genes and chromosomes as DNA sequences, The Plant Journal, 53, (2008) 661–673.
9. J. Pattillo, N. Youssef and S. Butenko, On clique relaxation models in network analysis, European Journal of Operational Research, 226.1 (2013) 9–18.
10. S.B. Seidman, Network structure and minimum degree, Social Networks 5, (1983) 269–287.
11. S.B. Seidman and B.L. Foster, A graph theoretic generalization of the clique concept, Journal of Mathematical Sociology 6, (1978) 139–154.
12. H. Yu, A. Paccanaro, V. Trifonov and M. Gerstein, Predicting interactions in protein networks by completing defective cliques, Bioinformatics 22, (2006) 823–829.
13. C. Zahn, Graph-theoretical methods for detecting and describing gestalt clusters, IEEE Transactions on Computers C-20 (1971) 68–86.
14. C. Zheng, A. Kononenko, J. Leebens-Mack, E. Lyons and D. Sankoff, Gene families as soft cliques with backbones: Amborella contrasted with other flowering plants, BMC Genomics 2014 15(Suppl 6):S8.

Reference for genome data

15. Amborella-Genome-Project: The *Amborella* genome and the evolution of flowering plants. Science 342(6165), 1241089 (2013)
16. Ming, R., et al.: Genome of the long-living sacred lotus (*Nelumbo nucifera Gaertn.*). Genome Biology 14(5), 41 (2013)
17. Yu, J., et al.: A draft sequence of the rice genome (*Oryza sativa L. ssp. indica*). Science 296(5565), 79–92 (2002)
18. Vogel, J.P., et al.: Genome sequencing and analysis of the model grass *Brachypodium distachyon*. Nature 463(7282), 763–768 (2010)
19. Schnable, P.S., et al.: The b73 maize genome: complexity, diversity, and dynamics. Science 326(5956), 1112–1115 (2009)
20. Paterson, A.H., et al.: The *Sorghum bicolor* genome and the diversification of grasses. Nature 457(7229), 551–556 (2009)
21. Bennetzen, J.L., et al.: Reference genome sequence of the model plant *Setaria*. Nature Biotechnology 30(6),555–561 (2012)
22. D'Hont, A., et al.: The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. Nature 488(7410), 213–217 (2012)
23. Wang, W., et al.: The *Spirodela polyrhiza* genome reveals insights into its neotenous reduction fast growth and aquatic lifestyle. Nature Communications 5 (2014)
24. Al-Dous, E.K., et al.: De novo genome sequencing and comparative genomics of date palm (*Phoenix dactylifera*). Nature Biotechnology 29(6), 521–527 (2011)
25. Jaillon, O., et al.: The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449(7161), 463–467 (2007)

26. Myburg, A., et al.: The *Eucalyptus grandis* genome project: Genome and transcriptome resources for comparative analysis of woody plant biology. In: BMC Proceedings, vol. 5, p. 20 (2011).
27. Wu, G.A., et al.: Sequencing of diverse mandarin, pummelo and orange genomes reveals complex history of admixture during citrus domestication. Nature Biotechnology (2014)
28. Xu, Q., et al.: The draft genome of sweet orange (*Citrus sinensis*). Nature Genetics 45(1), 59–66 (2013)
29. Argout, X., et al.: The genome of *Theobroma cacao*. Nature Genetics 43(2), 101–108 (2011)
30. Ming, R., et al.: The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya Linnaeus*). Nature 452(7190), 991–996 (2008)
31. Arabidopsis-Genome-Initiative, et al.: Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408(6814), 796 (2000)
32. Hu, T.T., et al.: The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. Nature Genetics 43(5), 476–481 (2011)
33. Wang, X., et al.: The genome of the mesopolyploid crop species *Brassica rapa*. Nature Genetics 43(10), 1035–1039 (2011)
34. Slotte, T., et al.: The *Capsella rubella* genome and the genomic consequences of rapid mating system evolution. Nature Genetics 45(7), 831–835 (2013)
35. Haudry, A., et al.: An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. Nature Genetics 45(8), 891–898 (2013)
36. Dassanayake, M., et al.: The genome of the extremophile crucifer *Thellungiella parvula*. Nature Genetics 43(9), 913–918 (2011)
37. Yang, R., et al.: The reference genome of the halophytic plant *Eutrema salsugineum*. Frontiers in Plant Science 4 (2013)
38. Guo, S., et al.: The draft genome of watermelon (*Citrullus lanatus*) and resequencing of 20 diverse accessions. Nature Genetics 45(1), 51–58 (2013)
39. Huang, S., et al.: The genome of the cucumber, *Cucumis sativus L.* Nature Genetics 41(12), 1275–1281 (2009)
40. Verde, I., et al.: The high-quality draft genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution. Nature Genetics 45(5), 487–494 (2013)
41. Shulaev, V., et al.: The genome of woodland strawberry (*Fragaria vesca*). Nature Genetics 43(2), 109–116 (2011)
42. Sato, S., et al.: Genome structure of the legume, *Lotus japonicus*. DNA Research 15(4), 227–239 (2008)
43. Schmutz, J., et al.: A reference genome for common bean and genome-wide analysis of dual domestications. Nature Genetics (2014)
44. Varshney, R.K., et al.: Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. Nature Biotechnology 30(1), 83–89 (2012)
45. Schmutz, J., et al.: Genome sequence of the palaeopolyploid soybean. Nature 463(7278), 178–183 (2010)
46. Tuskan, G.A., et al.: The genome of black cottonwood, *Populus trichocarpa (torr. & gray)*. Science 313(5793), 1596–1604 (2006)
47. Wang, Z., et al.: The genome of flax (*Linum usitatissimum*) assembled de novo from short shotgun sequence reads. The Plant Journal 72(3), 461–473 (2012)
48. Prochnik, S., et al.: The cassava genome: current progress, future directions. Tropical Plant Biology 5(1), 88–94 (2012)

49. Chan, A.P., et al.: Draft genome sequence of the oilseed species *Ricinus communis*. Nature Biotechnology 28(9), 951–956 (2010)
50. Huang, S., et al.: Draft genome of the kiwifruit *Actinidia chinensis*. Nature Communications 4 (2013)
51. Tomato-Genome-Consortium, et al.: The tomato genome sequence provides insights into fleshy fruit evolution. Nature 485(7400), 635–641 (2012)
52. Potato-Genome-Sequencing-Consortium, et al.: Genome sequence and analysis of the tuber crop potato. Nature 475(7355), 189–195 (2011)
53. Qin, C., et al.: Whole-genome sequencing of cultivated and wild peppers provides insights into *Capsicum* domestication and specialization. Proceedings of the National Academy of Sciences 111(14), 5135–5140 (2014)
54. Ibarra-Laclette, E., et al.: Architecture and evolution of a minute plant genome. Nature 498(7452), 94–98 (2013)
55. Hellsten, U., et al.: Fine-scale variation in meiotic recombination in *Mimulus* inferred from population shotgun sequencing. Proceedings of the National Academy of Sciences 110(48), 19478–19482 (2013)
56. Young, N.D., et al.: The *Medicago* genome provides insight into the evolution of rhizobial symbioses. Nature 480(7378), 520–524 (2011)