The Combinatorics of Finite Metric Spaces and the (Re-)Construction of Phylogenetic Trees.

Extended abstract and some relevant references

Andreas Dress,
CAS-MPG Partner Institute for Computational Biology,
Shanghai Institutes for Biological Sciences,
Chinese Academy of Sciences, Shanghai 200031, P.R. China
and Max-Planck-Institut fuer Mathematik in den Naturwissenschaften,
Inselstrasse 22-26, D-04103 Leipzig, Germany

2. Oktober 2007

Over the last thirty years, various combinatorial aspects of finite metric spaces have been studied in the context of phylogenetic-tree reconstruction. A starting point was the following observation:

Given a metric $D : X \times X \to \mathbb{R}$ representing the approximative genetic distances $D(x, y), \ldots$ between the members $x, y, \ldots$ of a finite collection X of taxa, it was shown in [15] that the following assertions relating to the *object of desire*, a "phylogenetic X-tree", all are equivalent:

(i) The "tight span"

$$T_D := \{f \in \mathbb{R}^X : \forall_{x \in X} f(x) = \sup_{y \in X}(D(x,y) - f(y))\}$$

of D is an $\mathbb{R}$-*tree*.

(ii) There exists an *edge-weighted X-tree* $(V, E; \ell)$ — i.e., a finite tree (V, E) with vertex set V and edge set E such that V contains X and every vertex in $V - X$ has degree at least 3, and a (positive) edge weighting

$\ell : E \rightarrow \mathbb{R}_{>0}$ that assigns a positive length $\ell(e)$ to every edge e in E — such that D is the restriction to X of the *shortest-path metric* induced by ℓ on V.

(iii) There exists a map $w : \mathcal{S}(X) \rightarrow \mathbb{R}_{\geq 0}$ from the set $\mathcal{S}(X)$ of all bipartitions or *splits* $S = \{A, B\}$ of X into the set $\mathbb{R}_{\geq 0}$ of non-negative real numbers such that

- given any two splits $S = \{A, B\}$ and $S' = \{A', B'\}$ in $\mathcal{S}(X)$ with $w(S), w(S') \neq 0$, at least one of the four intersections $A \cap A'$, $B \cap A'$, $A \cap B'$, and $B \cap B'$ is empty and
- $D(x, y) = \sum_{S \in \mathcal{S}(X:x \leftrightarrow y)} w(S)$ holds where

$$\mathcal{S}(X : x \leftrightarrow y) := \big\{\{A, B\} \in \mathcal{S}(X) : x \in A, y \in B\big\}$$

denotes the set of those splits $S = \{A, B\} \in \mathcal{S}(X)$ that *separate* x and y.

(iv) $D(x, y) + D(u, v) \leq \max\big(D(x, u) + D(y, v), D(x, v) + D(y, u)\big)$ holds for all $x, y, u, v \in X$

Moreover, the metric space T_D actually coincides in this case with the $\mathbb{R}$-tree that is canonically associated with an edge-weighted tree $(V, E; \ell)$.

This observation suggested to further investigate (1) the tight-span construction and (2) representations of metrics by weighted split systems with various specific properties, even if the metric in question would not satisfy the very special properties described above. These investigations have, in turn, given rise to a full-fledged research program dealing many diverse aspects of these two topics (see the list of references below).

In my lecture, I will focus on the rather new developments relating to *block decomposition* and *virtual cut points* of metric spaces reported in [28] to [31] that allow to canonically decompose any given finite metric space into a sum of pairwise compatible *block metrics* thus providing a far-reaching generalization of the result recalled above.

Literatur

[1] J. Apresjan, An algorithm for constructing clusters from a distance matrix. *Mashinnyi perevod: prikladnaja lingvistika* **9** (1966) 3-18.

[2] E. Baake, What can and cannot be inferred from pairwise sequence comparisons? *Math. Biosci.* **154** (1998) 1-21.

[3] H.-J. Bandelt, Recognition of tree metrics. *SIAM J. Disc. Math.* **3** (1990) 1-6.

[4] H.-J. Bandelt and A. Dress. Weak hierarchies associated with similarity measures – an additive clustering technique, *Bul. Math. Biol.* **51** (1989) 133-166.

[5] H.-J. Bandelt and A. Dress. A canonical split decomposition theory for metrics on a finite set, *AiM* **92** (1992) 47-105.

[6] H.-J. Bandelt and A. Dress. Split decomposition: a new and useful approach to phylogenetic analysis of distance data, *Molecular Phylogenetics and Evolution* **1**(3) (1992b) 242-252.

[7] .-J. Bandelt, V. Chepoi, A. Dress, and J. Koolen, Combinatorics of lopsided sets. *Eur. J. Comb.*, 27 (2006) 669–689.

[8] H.-J. Bandelt and M. Steel, Symmetric matrices representable by weighted trees over a cancellative abelian monoid. *SIAM. J. Disc. Math.* **8** (1995) 517-525.

[9] J.-P. Barthélemy and A. Guénoche, Trees and proximity representations (Wiley, 1991).

[10] S. Böcker and A. Dress, Recovering symbolically dated, rooted trees from symbolic ultrametrics. *AiAM.* **138** (1998) 105-125.

[11] B. Bowditch, Notes on Gromov's hyperbolicity criterion for path metric spaces. in: *Group theory from a geometric viewpoint*, E. Ghys, A. Haefliger, A. Verjovsky eds., World Scientific, 1991, 64-167.

[12] P. Buneman, The recovery of trees from measures of dissimilarity. In *Mathematics in the Archeological and Historical Sciences*, 387-395, F. Hodson et al. eds, Edinburgh University Press, 1971.

[13] D. Bryant and V. Berry, A family of clustering and tree construction methods. *AiAM* **27** (2001) 705-732.

[14] D. Bryant and A. Dress, Linearly independent split systems. *Europ. J. Comb.*, to appear.

[15] A. Dress, Trees, tight extensions of metric spaces, and the cohomological dimension of certain groups: A note on combinatorial properties of metric spaces. *AiM* **53** (1984) 321-402.

[16] A. Dress, Towards a classification of transitive group actions on finite metric spaces. *AiM* **74** (1989) 163 - 189.

[17] A. Dress, Towards a theory of holistic clustering. In *Mathematical hierarchies and biology*, (Piscataway, NJ, 1996), 271-289, DIMACS Ser. Discrete Math. Theoret. Comput. Sci., 37, Amer. Math. Soc., Providence, RI, 1997.

[18] A. Dress, Proper Gromov transforms of metrics are metrics. *AML* **15** (2002) 995-999.

[19] A. Dress. Graphs and Metrics. In *Encyclopaedia of Genetics, Genomics, Proteomics and Bioinformatics*, Shankar Subramaniam, Adam Kuspa et al. eds., John Wiley and Sons, 2005.

[20] A. Dress, The Category of X-nets. In *Networks: From Biology to Theory*, Jianfeng Feng, Jürgen Jost, and Minping Qian eds, Springer, 2006.

[21] A. Dress, Phylogenetic analysis, split systems, and boolean functions. In: *Aspects of Mathematical Modelling*, Roger Hosking and Wolfgang Sprössig eds, Birkhäuser, 2007, to appear.

[22] A. Dress, A note on group-valued split and set systems. *Contributions to Discrete Mathematics*, to appear.

[23] A. Dress, Split decomposition over an abelian group, Part I: Generalities. *AoC*, to appear.

[24] A. Dress, Split decomposition over an abelian group, PartII: Group-valued split systems with weakly compatible support. *Discrete Applied Mathematics, Special Issue on Networks in Computational Biology*, to appear.

[25] A. Dress. Split decomposition over an abelian group, Part III: Group-valued split systems with compatible support. submitted.

[26] A. Dress, B. Holland, K. T. Huber, J. H. Koolen, V. Moulton, and J. Weyer-Menkhoff, Δ additive and Δ ultra-additive maps, Gromov's trees, and the Farris transform. *Discrete Applied Mathematics*, Discrete Applied Mathematics, 146, 2005, 51-73.

[27] A. Dress, K. Huber, A. Lesser, and V. Moulton, Hereditarily optimal realizations of consistent metrics. *AoC, Special Volume on Biomathematics*, 10:63–76, 2006.

[28] A. Dress, K. Huber, J. Koolen, and V. Moulton, Compatible decompositions and block realizations of finite metric spaces, submitted.

[29] A. Dress, K. Huber, J. Koolen, and V. Moulton, An algorithm for computing virtual cut points in finite metric spaces, Proceedings of COCOA 2007, Lecture Notes in Computing Science.

[30] A. Dress, K. Huber, J. Koolen, and V. Moulton, Cut points in metric spaces, *AML*, in press.

[31] A. Dress, K. Huber, J. Koolen, V. Moulton, and A. Spillner, A note on the metric cut point and the metric bridge partition problems. submitted.

[32] A. Dress, K. Huber, and V. Moulton, A Comparison between two distinct continuous models in projective cluster theory: The median and the tight-span construction, *AoC*, **2**, 299-311, 1998.

[33] A. Dress, K. T Huber, and V. Moulton, Some variations on a theme by Buneman. *AoC*, **1** (1997) 339-352.

[34] A. Dress, K. T. Huber, and V. Moulton, An exceptional split geometry. *AoC* **4** (2000) 1-11.

[35] A. Dress, K T. Huber, and V. Moulton, An explicit computation of the injective hull of certain finite metric spaces in terms of their associated Buneman complex. *AiM* **168** (2002) 1-28.

[36] A. Dress, K. T. Huber, and V. Moulton, Some uses of the Farris Transform in Mathematics and Phylogenetics — A Review. *AoC, Special Volume on Biomathematics*, to appear.

[37] A. Dress, J. Koolen, and V. Moulton: "$4n-10$": A note on 2-compatible split systems. *AoC*, **8** (2004) 463 – 471.

[38] A. Dress, V. Moulton, and W. Terhalle, T-theory: An overview. *Europ. J. Comb.* **17** (1996) 161-175.

[39] A. Dress and M. Steel, Mapping edge sets to splits in trees: the path index and parsimony. *AoC, Special Volume on Biomathematics*, **10** (2006) 77-96.

[40] A. Dress and M. Steel, Phylogenetic Diversity over an Abelian Group. *AoC, Special Volume on Biomathematics*, to appear.

[41] J. S. Farris, On the phylogenetic approach to vertebrate classification. In: Major patterns in vertebrate evolution, Eds: M. K. Hecht, P. C. Goody, and B. M. Hecht. Plenum Press, 1976.

[42] J. S. Farris, On the naturalness of phylogenetic classification. *Sys. Zool.* **28** (1979) 200-214.

[43] J. S. Farris, The information content of the phylogenetic system. *Sys. Zool.* **28** (1979) 483-519.

[44] J. S. Farris, A. G. Kluge, and M. J. Eckardt, A numerical approach to phylogenetic systematics. *Sys. Zool.* **19** (1970) 172-189.

[45] M. Gromov, Hyperbolic Groups. Essays in Group Theory, MSRI series vol. 8, S.Gersten, ed., Springer-Verlag, 1988.

[46] . Grünewald, K. Forslund, A. W. M. Dress, and V. Moulton, QNet: An agglomerative method for the construction of phylogenetic networks from weighted quartets. *Mol. Biol. Evol.*, 24:532–538, 2007.

[47] D. Huson, SplitsTree: a program for analyzing and visualizing evolutionary data. *Bioinformatics* **14** (1998) 68-73; see also *http://www-ab.informatik.uni-tuebingen.de/software/splits/welcome_en.h tml.*

[48] D. Huson and A. Dress.
Constructing Splits Graphs. *ACM Transactions in Computational Biology and Bioinformatics*, **1** (2004), 109 - 115.

[49] J. Isbell, Six theorems about metric spaces. *Comment. Math. Helv.* **39** (1964) 65-74.

[50] N. Jardine, Towards a general theory of clustering. *Biometrics* **25** (1969) 609-610.

[51] P. J. Lockhart, A. E. Meyer, and D. Penny, Testing the phylogeny of Swordtail fishes using Split Decomposition and Spectral Analysis. J. Mol. Evol. **41** (1995) 666-674.

[52] A. Parker-Rhodes and R. Needham, A reduction method for non-arithmetic data, and its application to thesauric translation. *in* "Information processing, Proceedings of the International Conference on Information Processing, Paris, 1960", UNESCO, pp. 321-325.

[53] C. Semple and M. Steel, Tree representations of non-symmetric. group-valued proximities. *AiAM.* **23** (1999) 300-321.

[54] C. Semple and M. Steel, Phylogenetics. Oxford University Press, 2003.

[55] M. Steel, Recovering a tree from the leaf colourations it generates under a Markov model. *AML* **7** (1994) 19-24.