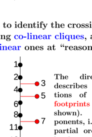
The Footprint Sorting Problem

C.Fried[1,2], W. Hordijk[3], S.J. Prohaska[1,2], C. R. Stadler[4], P.F. Stadler[1,2]

[1]Lehrstuhl für Bioinformatik, Institut für Informatik, Universität Leipzig, Germany
[2]Institut für Theoretische Chemie und Strukturbiologie, Universität Wien, Austria
[3]Department of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand
[4]Department of Computer Science, University of Leipzig, Germany

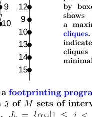
Tel: ++49 341 149 5120 Fax: ++49 341 149 5119 Email: {claudia, sonja, studla}@bioinf.uni-leipzig.de
WorldWideWeb: http://www.bioinf.uni-leipzig.de/

Non-coding DNA in the genome of eukaryotes contains a large amount of functional regions for the regulation of gene expression. These regulatory regions are subject to stabilizing selection and evolve much slower than adjected non -functional DNA. These so-called phylogenetic footprints can be detected with the technique of phylogentic footprinting. Automatic procedure for phylogenetic footprinting such as `tracker`[7] can produce large amounts of data. A seemingly simple problem is to sort footprints in their order along the genomes.



Sorting is complicated by the fact that not all footprints are co-linear: they may cross each other. Phylogenetic footprints (small balls) of different sequences (vertical lines) belonging to a common footprint clique are connected by lines.

The problem is to identify the crossing footprints, to sort the remaining co-linear cliques, and finally to insert the non-colinear ones at "reasonable" positions.



The directed graph $\vec{G}$ describes the relative locations of the phylogenetic footprints (not all arcs are shown). 2-connected components, i.e., obstructions to partial ordering are shown by boxes. The right side shows a well-ordering on a maximal set of co-linear cliques. The diagram also indicates the obstructing cliques at positions with a minimal number of conflicts.

The output of a footprinting program can be regarded as a collection $\mathfrak{J}$ of M sets of intervals, termed "footprint cliques", $J_k = \{\alpha_{kj} | 1 \le j < m_k\}$. A footprint clique contains at most one sub-interval of each sequence. For two sub-intervals $\alpha = [i; a, l]$ and $\alpha' = [i; a', \ell']$ on the same sequence i we have the order relation

$$\alpha < \alpha' \quad \Longleftrightarrow \quad \left\{ \begin{matrix} a < a' \\ \ell < \ell' \text{ and } a = a' \end{matrix} \right\} \tag{1}$$

Clearly, the ordering (1) implies an order-like $\prec^*$ on $\mathfrak{J}$: For $J, J' \in \mathfrak{J}$ we set $J \preceq^* J'$ iff for all $i \in \{1, \dots, N\}$ for which there is an $\alpha = [i, a, \ell] \in J$ and an $\alpha' = [i, a', \ell'] \in J$ we have $\alpha \le \alpha'$. Our task is hence to find a well-order on $\mathfrak{J}$ that is consistent with the order of the sub-intervals on each sequence, i.e., that is an extension of $\prec^*$ on $\mathfrak{J}$.

Minimium Feedback-Vertex Sets

The set $\mathfrak{J}$ can be regarded as the vertex set of a directed graph $\vec{G}$ with arcs $\alpha \to \beta$ iff $\alpha \preceq^* \beta$ and $\alpha \ne \beta$. The transitive closure of the relation $\preceq^*$ is a partial order if and only if the associated graph $\vec{G}$ is acyclic.

Given the vertex-weighted directed graph $\vec{G} = (V, A)$ with vertex set V and arc-set A, we have to find a maximal subset of vertices such that the induced subgraph $\vec{G}[U]$ is acyclic.

This problem is known as the *Minimum Feedback-Vertex Set Problem* [6].

We re-implemented the exact algorithm [1] which is not fast enough and a well-tested heuristic [3].

We compute the transitive closure of the acyclic graph by connecting two vertices by adding an arc from i to j iff there is a directed path from i to j. The graph $\vec{G}^\bullet$ is again acyclic and represents a partial order.

The computation of a well-ordering of the vertex set of a directed acyclic graph $\vec{G}$ is known as *topological sorting* and can be solved in $\mathcal{O}(|A|)$ time [4].

Sorting as Optimization Proplem

The standard sorting problem can be reformulated as a combinatorial optimization problem where $f(\pi)$ counts the number of mis-sorted pairs of objects with appropriate wights, if desired.

If $(X, <)$ is well-ordered then every adaptive walk that makes use of the canonical transpositions reaches the correct sorting. Therefore adaptive walk using arbitrary transpositions and/or reversals also guarantees to the find the correct sorting because these moveset contain the canonical transpositions.

The sorting problem can be extended to an arbitary relation $\prec$ on X which need not be complete or even antisymmetric. Clearly, there is a perfect solution π satisfying $f(\pi) = 0$ iff $(X, \prec)$ is a partially ordered set. A permutation thus satisfies $f(\pi) = 0$ iff π codes for a linear extension of $\prec$.

If $(X, <)$ is partial order then every adaptive walk that makes use of the (general) transpositions reaches a correct topological sorting.

To obtain a useful ordering π of the nodes of $\vec{G}$ by minimizing f we can use a local search algorithm such as simulated annealing [5].

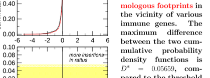
Given the permutation π that represents the approximation to the true ordering of the vertices, we would also like to identify a minimum feedback vertex set. A heuristic approximation proceeds by iteratively removing the vertex k with largest weight $f_k(\pi)$:

Input: $\vec{G} = (V, A)$, **U**, π
Output: $\bar{W}$ /* feedback vertex set */
1: $\bar{W} \leftarrow \emptyset$
2: **while** $f(\pi) > 0$ **do**
3: $\quad \ell := \arg\max f_k(\pi)$.
4: $\quad \bar{W} \leftarrow \bar{W} \cup \{\ell\}$.
5: $\quad u_{\ell i}, u_{i\ell} \leftarrow 0$ **for all** i.

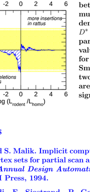
This produces a feedback vertex set; its minimality is not guaranteed even if π minimizes f.

An Application

Variations of the genome size can be caused by large-scale changes such as gene and chromosome duplications and by the accumulation of small-scale changes e.g. insertion and deletions. Recent analysis showed that rat chromosomes differ by several chromosome rearrangements from both the mouse and the human homologs [2].



Length variations of the non-conserved sequences between *adjacent* homologous footprints in different sequences can be used to estimate these local changes. Determination of adjacent footpints requires a correct sorting.



Comparison of length ratios of human/rat and human/mouse non-conserved sequence fragment located between homologous footprints in the vicinity of various immune genes. The maximum difference between the two cummulative probability density functions is $D^* = 0.05659$, compared to the threshold value $D_{95\%} = 0.05438$ for the Kolmogorov-Smirnov test. The two distributions are (just barely) significantly different.

References

[1] P. Ashar and S. Malik. Implicit computation of minimum-cost feedback vertex sets for partial scan and other applications. In *Proc. 31st Annual Design Automation Conf. (DAC)*, pages 77–80. ACM Press, 1994.

[2] A. Behboudi, E. Sjostrand, P. Gomez-Fabre, A. Sjoling, Z. Taib, K. Klinga-Levan, F. Stahl, and G. Levan. Evolutionary aspects of the genomic organization of rat chromosome 10. *Cytogenet. Genome Res.*, 96:52–59, 2002.

[3] P. Festa, P. Pardalos, and M. G. C. Resende. Algorithm 815: Fortran subroutines for computing approximate solutions of feedback set problems using GRASP. *ACM Transactions Math. Software*, 27:456–464, 2001.

[4] T. Hagerup and M. Maas. Generalized topological sorting in linear time. *Nord. J. Comput.*, 1:38–49, 1994.

[5] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220:671–680, 1983.

[6] A. Lempel and I. Cederbaum. Minimum feedback arc and vertex sets of a directed graph. *IEEE Trans. Circuit Theory*, 13:399–403, 1966.

[7] S. Prohaska, C. Fried, C. Flamm, G. Wagner, and P. F. Stadler. Surveying phylogenetic footprints in large gene clusters: Applications to Hox cluster duplications. *Mol. Phylog. Evol.*, 2003. submitted; SFI preprint #03-02-011.