Tagsnp selection using Weighted CSP and Russian Doll Search with Tree Decomposition

D. Allouche, S. de Givry, M. Sanchez, T. Schiex

UBIA, UR 875, INRA, F-31320 Castanet Tolosan, France.
{allouche,degivry,msanchez,tschiex}@toulouse.inra.fr

Abstract. In recent a paper [5], we have proposed a new method called "Russian Doll Search with Tree Decomposition" to improve optimization in weighted CSP frameworks. This approach is based on identifying conditionally independent subproblems, which are solved independently and whose optimum is cached, in the Branch and Bound algorithm.

The central idea was to (inductively) solve a relaxation of each subproblem to strengthen bounds in order to overcome weakness associated to the locality of bounds induced by decomposition, which often hampers the practical effects of this result, because subproblems are often uselessly solved to optimality.

The algorithm obtained, BTD-RDS, generalizes both RDS and tree-decomposition based algorithms such as BTD or AND-OR Branch and Bound. Experimentation have been done on the one hand on CELAR benchmark [1], closing a very hard frequency assignment instance which has been open for more than 10 years and on the other hand on a new benchmark from bioinformatic field: the tagsnp selection.

This problem occurs in genetics and polymorphism analysis. Single nucleotide polymorphisms, or SNPs, are DNA sequence variations that occur when a single nucleotide (`A`,`T`,`C`,or `G`) in the genome sequence of an individual is altered. For example a SNP might change the DNA sequence `AAGGCTAA` to `ATGGCTAA`. For a variation to be considered a SNP, it must occur in at least 1% of the population. There are several millions SNPs in the 3 billions nucleotides long human genome, explaining up to 90% of all human genetic variation. SNPs may explain a portion of the heritable risk of common diseases and can affect response to pathogens, chemicals, drugs, vaccines, and other agents. The TagSNP problem is a sort of lossy compression problem which consists in selecting a small subset of SNPs such that the selected SNPs, called tag SNPs, will capture most of the genetic information. The goal is to capture a maximally informative subset of SNPs to make screening of large populations feasible [2]. From the combinatorial point of view the TagSNP problem is equivalent to a set covering problem (NP-hard) with additional quadratic criteria.

Modelization and practical experimentation will be present on HAPMAP chromosome 1 [4] data.

The practical exploitation of tree decompositions to solve combinatorial optimization problems that have structure is not always straightforward. Our experiments show that, even on problems that have a nice visible structure, it is often very profitable and sometimes crucial to restrict the maximum size of the separators of the decomposition exploited. Theory

says that separator size influences the space complexity of the structural algorithm like BTD and RDS-BTD. In practice, the improvement in efficiency is mostly explainable by the added freedom in variable ordering allowed by cluster merging, an observation consistent with [3] conclusions.

In this presentation, we will first focus on our recently published results about tagsnp experimentation using wcsp model and a comparison will be done with PLNE approach. Furthermore, we will specifically address an ongoing work on the problem decomposition and stress a putative heuristic approach allowing to adapt dynamically the maximum size of separators of the decomposition based on the specific structure of a given instance.

References

1. B. Cabon, S. de Givry, L. Lobjois, T. Schiex, and J.P. Warners. Radio Link Frequency Assignment. *Constraints*, 4(1):79–89, 1999.
2. J.N. Hirschhorn and M.J. Daly. Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, 6(2):95–108, 2005.
3. P. Jégou, S. N. Ndiaye, and C. Terrioux. Dynamic management of heuristics for solving structured CSPs. In *Proc. of CP-07*, pages 364–378, Providence, USA, 2007.
4. Z. S. Qin, S. Gopalakrishnan, and G. R. Abecasis. An efficient comprehensive search algorithm for tagsnp selection using linkage disequilibrium criteria. *Bioinformatics*, 22(2):220–225, 2006.
5. M. Sanchez, D. Allouche, S. de Givry, and T. Schiex. Russian doll search with tree decomposition. In *In Proc. of IJCAI'09*, Pasadena, USA, July 11-17 2009.