RHM

RHM is an alternative method for calculating phylogenetic distances in order to construct phylogenetic tree based on compression algorithms. The method was proposed in 2006 by Teemu Roos, Tuomas Heikkilä and Petri Myllymäki (hence its name) (Roos et al. 2006). Given a set of textual documents, the method produces a bifurcating stemma. RHM operates in a manner similar to the maximum parsimony method with certain important differences. Roos and Heikkilä (2009) have argued that RHM and maximum parsimony actually yield the best results when constructing cladistics based stemmata, but they also point out that the computational cost is high – i.e. computing a stemma for a tradition with anywhere between 10 and 50 manuscripts may take considerable time, (hours rather minutes).

The RHM method uses an approximation of Kolmogorov complexity which – theoretically – is defined as the smallest possible but complete description of an object (e.g. compressing "aaaaa" into "5a"). Theoretically smallest because for formal languages (like computer languages) it is mathematically impossible to prove that such a description is actually the smallest possible. In practice, therefore, such smallest possible descriptions are always approximated. RHM uses such an approximation to evaluate the distance (i.e. the amount of dissimilarity) between witnesses while constructing a phylogenetic tree by using GZIP compression as the approximation. The use of GZIP automatically gives greater weight to longer variants, e.g. the weight assigned to the variation *"beatus"* vs. *"sanctus"* is six units while the weight assigned to the variation "ex" vs "in" is only three units. Similarly, variation in word order is usually assigned a smaller weight than variation in the actual words. All of this is based only on the actual information content and not on scholarly evaluation.

Because RHM uses compression-based comparison without user-intervention, all weighting of variations is based on information immanent in the text without scholarly evaluation intervening. This means that the application of RHM requires less effort than an analysis based on carefully constructed encodings where, for example, variation that is considered insignificant (capitalisation, punctuation, etc.) is removed by normalisation, variation in word order is encoded using special characters, and so on. This also results in RHM using as its input aligned text files which contain the actual words, instead of encoded variant readings using arbitrary characters such as *A,B,C,…* as is done in, for instance, the Nexus data matrix format.

A difference between RHM and typical maximum parsimony implementations, such as that in PAUP or PHYLIP, is the search procedure used to find highly scoring tree structures.The search technique used in RHM takes a user-defined parameter, the number of search steps or iterations. The more iterations, the longer the search takes but also the better a solution can be expected. The maximum parsimony implementation in PAUP and PHYLIP on the other hand, is faster.

References

– Roos, Teemu, Tuomas Heikkilä, and Petri Myllymäki. 2006. "A Compression-Based Method for Stemmatic Analysis." In *ECAI 2006: Proceedings of the 17th European Conference on Artificial Intelligence: August 29 – September 1, 2006*, edited by Gerhard Brewka et al., 805–806. Amsterdam: IOS Press.
– Roos, Teemu, and Tuomas Heikkilä. 2009. "Evaluating Methods for Computer-Assisted Stemmatology Using Artificial Benchmark Data Sets." *Literary and Linguistic Computing* 24 (4): 417–433.

JZ, TR