Ann. Zool. Fennici 51: 43–48 ISSN 0003-455X (print), ISSN 1797-2450 (online)
Helsinki 7 April 2014

Towards computational techniques for identifying candidate chronofaunas

Ella Bingham[1] & Heikki Mannila[2]

[1)] *Helsinki Institute for Information Technology HIIT, Aalto University, P.O. Box 15600, FI-00076 Aalto, Finland (corresponding author's e-mail: ella.bingham@hiit.fi)*
[2)] *Department of Computer and Information Science, Aalto University, P.O. Box 15400, FI-00076 Aalto, Finland*

Received 2 Oct. 2013, final version received 4 Dec. 2013, accepted 9 Dec. 2013

Bingham, E. & Mannila, H. 2014: Towards computational techniques for identifying candidate chronofaunas. — *Ann. Zool. Fennici* 51: 43–48.

A chronofauna is a geographically restricted collection of interacting animal populations that maintains its base structure over a long period of time. We describe a simple computational method that can identify candidate chronofaunas on the basis of presence-absence matrices only: A candidate chronofauna is a collection of sites that share an exceptionally large number of taxa with the defining site of the chronofauna. We show examples of candidate chronofaunas in the NOW data (*see* http://www.helsinki.fi/science/now).

Introduction

The concept of chronofauna was defined by Olson (1952) as "*a geographically restricted, natural assemblage of interacting animal populations that has maintained its base structure over a geologically significant period of time*". A chronofauna can be viewed as a high-level structure in paleontological record. Eronen *et al*. (2009) showed how this concept can be used as an organizing principle in describing the interplay of environmental changes and faunal changes.

The notion of a chronofauna is, of course, a deeply paleontological concept: identifying interesting chronofaunas requires deep knowledge about the underlying characteristics of the taxa and the environment. Olson's definition is a paleontological one, as it has the qualifications of a natural assemblage and the interaction of the populations. These characteristics are not directly observable from the basis of presence-absence data only.

However, some aspects of the concept are computational and can be identified on the basis of presence-absence matrices. In this paper, we describe simple computational techniques that can be used for this.

Material and methods

Candidate chronofaunas: definition and basic approach

The definition of chronofauna in Olson (1952) implies that a chronofauna consists of both a collection of taxa and a set of sites. Informally, a candidate chronofauna (CCF) can be considered to consist of a set of taxa and a set of sites such

that the taxa occur sufficiently often at the sites so that some sort of larger structure can be identified. We thus viewed a chronofauna as a pair (T,S), where T is a set of taxa and S is a set of sites so that for each site in S sufficiently many taxa from T occur at the site, and each taxon in T occurs at sufficiently many of the sites of S. A formal definition of a CCF requires, of course, that the concept of "sufficiently many" is defined more formally.

A good definition of candidate chronofaunas should satisfy certain conditions dictated by the nature of paleontological presence-absence data. First, the definition should have tolerance for incomplete sampling. That is, we cannot require that all taxa in the CCF occur at all sites of the CCF. Second, the occurrence of additional taxa should not influence the CCF. That is, if we have identified a CCF (T,S), and then from some site in S we find some additional taxa that are outside T, we should still view (T,S) as a CCF.

Different similarity indices such as the Jaccard index or Simpson's index (Jaccard 1912, Simpson 1949) can be used to identify sites that might belong to a candidate chronofauna. Eronen *et al*. (2009) used the Raup-Crick similarity index in this way: the chronofauna consists of sites that have similarity with a preselected site higher than a threshold value. The use of the Jaccard index satisfies the first condition above: assuming the threshold is low enough, incomplete sampling does not cause problems. However, the second requirement is more problematic. Suppose site u has a very long list of taxa, and the list for another site v is much shorter. Then the (Jaccard) similarity between u and v will be quite low, even in the case when v consists of exactly those taxa T that would determine a candidate chronofauna (T,S). This phenomenon is caused by the symmetry of similarity indices: the similarity between two sites with long and short lists, respectively, will be small.

Our computational approach to finding candidate chronofaunas was based on looking at sites with a fairly long list of taxa. For each such site u, we searched for sites v such that u and v have more common taxa than could be expected. Then u together with such sites v formed a candidate chronofauna. We preferred to find CCFs that have many sites of different ages, following Olson's original definition cited in the beginning of the Introduction: a chronofauna spans over a geologically significant period of time.

Estimation for the number of common taxa for two sites

Evaluating whether two sites u and v might belong to the same CCF depends on the number of common taxa u and v have. This number, of course, depends very strongly on the number of taxa at two sites. Therefore we used a randomization technique to find out the expected value of common taxa for u and v.

Our approach was as follows. Given two sites u and v, let m and n be the numbers of taxa at the sites, respectively, and assume p taxa occur in both u and v. [Then the Jaccard coefficient between the sites would be defined as $p/(m + n)$]. We assumed that we in the data can identify the sets of sites U that have approximately the same age as u; in the NOW database we used the MN zonation system.

To find out whether the occurrence of p shared taxa in u and v differs from the expected, we randomly generated lists of length m using the occurrence frequencies of the taxa in U as the probability of each taxon. For each such list we calculated how many taxa it shares with v. Repeating the generation yielded a distribution of the number of common taxa under the hypothesis that a site of length m would be generated by assuming the frequencies of the taxa in the same age group as v. Denoting by e and d the average and deviation of this distribution, respectively, we classified the number p of common taxa between u and v as high, if $p > e + 2d$. (Here, of course, the choice of the factor 2 was arbitrary.) In this case we stated that the *intersection* of u and v is *large*.

Finding candidate chronofaunas

Given the above approach, we computed candidate chronofaunas as follows. For each site s with sufficiently many taxa, we computed the number of common taxa between s and all other sites, and identified the sites which had a large

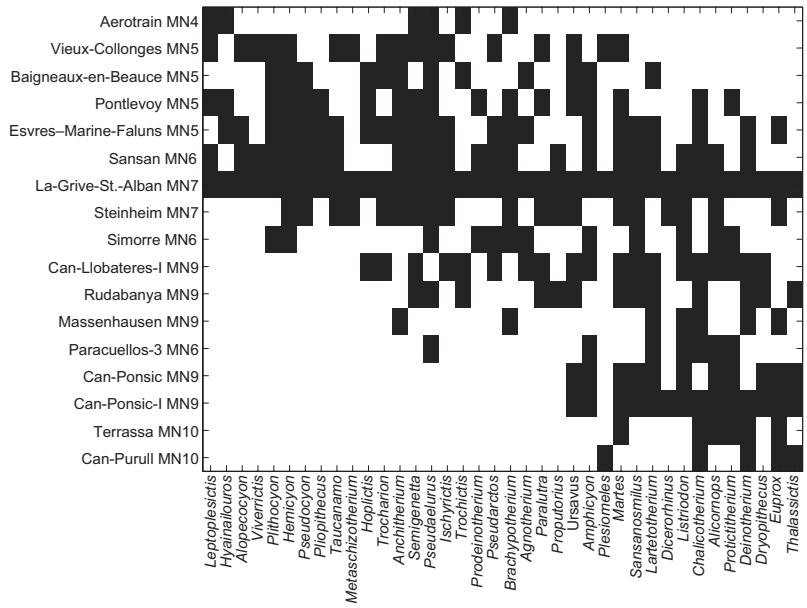

Fig. 1. The candidate chronofauna defined by the MN7 site La-Grive-St.-Alban (large mammals).

intersection with *s* according to the criterion described above. The chronofauna defined by *s* is the set *S* of such sites. The core taxa of the chronofauna were those that occurred in *s* and were present in at least 15% of the sites in *S*. (Again, the parameter was exchangeable.)

This approach produced one CCF for each defining site. We did not restrict overlap between CCF's of different defining sites: the sites belonging to the CCF of one site might overlap with the CCF of another site. One could define different score functions for the CCFs by using the number of sites, the number of common taxa, and the number of sites which have a large intersection with *s*. For simplicity we preferred to provide one candidate chronofauna for each defining site, as the interestingness of the CCFs depended also on other factors than such numerical ones.

Data

We used the NOW database (Fortelius 2011) as the data. We used all large mammal data in MN units MN1 to MN18. Our data contained taxonomic identification at least to the genus level, and we required that each locality had at least 7 taxa. After this we deleted all singletons, i.e. sites containing only 1 taxon occurrence. This left us with a dataset of 712 sites and 722 taxa. Our choice of data reflected that of Eronen *et al*. (2009), to facilitate comparisons, except that Eronen et al restricted their analysis to MN units MN7 to MN15 which are the most relevant to Pikermian chronofauna. We also conducted some tests on the data on small mammals from the NOW database, with the same selection procedure. This dataset had 741 sites and 573 taxa.

Results

To evaluate the method we used the NOW data to see whether it finds candidate chronofaunas that are in some way comparable with Pikermi. We ran the method on the NOW data by using as defining sites the sites with the highest number of taxa. As our goal was to identify other chronofaunas in addition to Pikermi, we omitted a handful of sites most similar to Pikermi when visualizing the results.

The candidate chronofaunas arising from three defining sites are visualized by showing the taxa and sites belonging to the chronofauna in Figs. 1–3. In these figures we ordered the rows and columns by using the barycentric algorithm (Sugiyama *et al*. 1981, Mäkinen and Siirtola

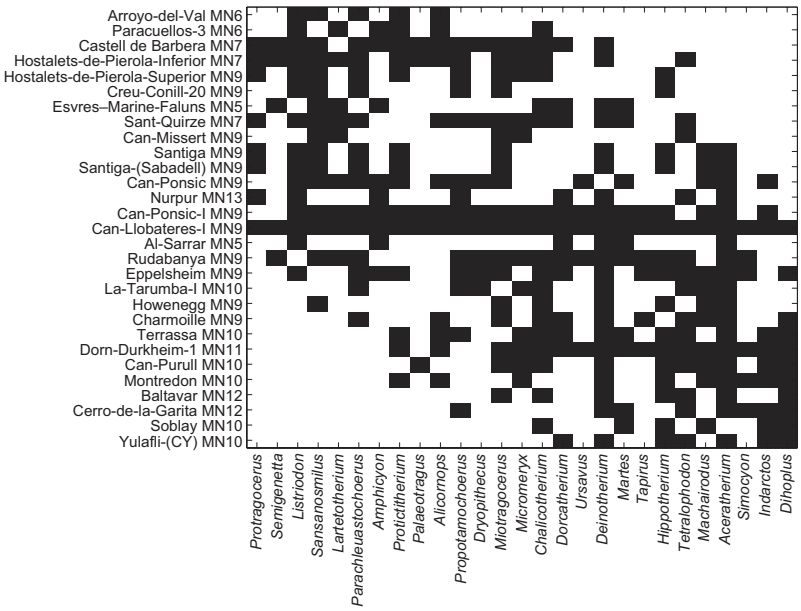

Fig. 2. The candidate chronofauna defined by the MN9 site Can Llobateres I (large mammals).

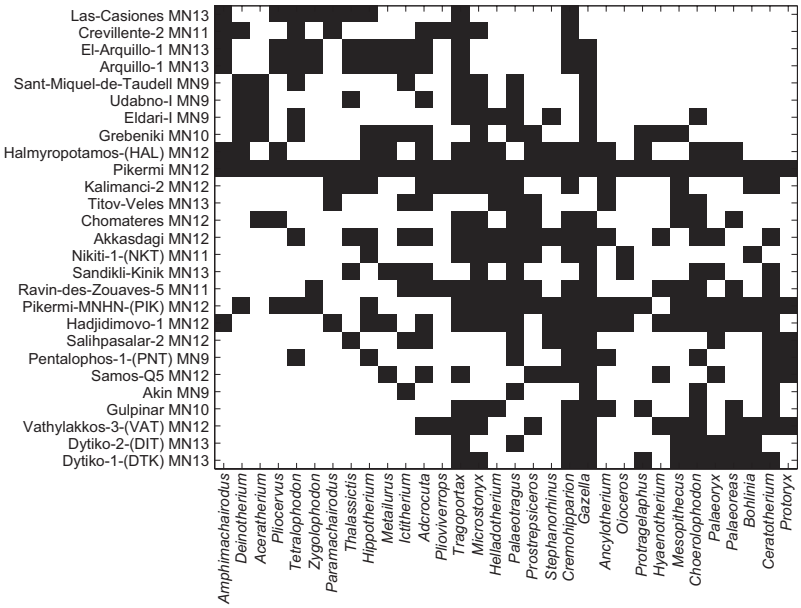

Fig. 3. The candidate chronofauna defined by MN12 site Pikermi (large mammals).

2005) that puts the 1s of the data matrix close to each other. Note that the ordering process had no information of the MN classes of the sites. The ordering still had a strong correlation with the MN units of the sites.

The number of sites, taxa, and the density of the candidate chronofaunas are quite similar to each other (Table 1). This suggests that there is some chronofauna-like structure in the NOW data in addition to the Pikermian chronofauna studied by Eronen *et al*. (2009).

When applied to the data on small land mammals from the NOW database, the results are similar (Table 2), with the exception that the temporal span of the candidate chronofaunas are longer than for large mammals. An example candidate chronofauna is shown in Fig. 4.

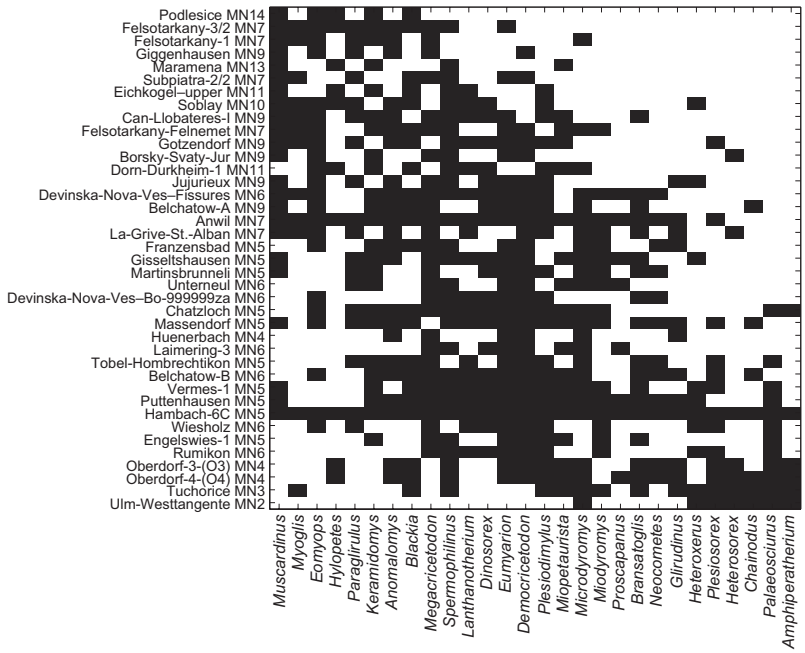

Fig. 4. The candidate chronofauna defined by MN5 site Hambach 6C (small mammals).

Discussion

Chronofaunas form an interesting high-level structure in paleontological data. The approach outlined in this paper can be used to find potential chronofaunas, defined as a set of sites that have an exceptionally large number of common taxa with the defining site of the chronofauna.

Table 1. Characteristics of candidate chronofaunas found from NOW data on large land mammals. Density is the frequency of occurrence of the taxa of the CCF in the sites of the CCF.

Defining site	Sites	Taxa	Density	First MN	Last MN
La-Grive-St.-Alban	17	38	0.39	4	10
Can-Llobateres-I	29	27	0.42	5	13
Pikermi	27	32	0.40	9	13
Esvres-Marine-Faluns	19	40	0.36	4	10
Dorn-Durkheim-1	23	22	0.43	9	13
Sansan	23	31	0.35	4	12
Yushe	28	16	0.41	9	16

Table 2. Characteristics of candidate chronofaunas found from NOW data on small land mammals. Density is the frequency of occurrence of the taxa of the CCF in the sites of the CCF.

Defining site	Sites	Taxa	Density	First MN	Last MN
Anwil	59	33	0.36	1	16
Podlesice	46	29	0.36	5	18
Soblay	46	30	0.37	4	16
Dorn-Durkheim-1	43	26	0.40	2	16
Weze-1	42	29	0.37	4	18
Ivanovce	35	30	0.39	9	18
Hambach-6C	39	28	0.46	2	14

Our method is simple and straightforward, and it does not require large amounts of computation. As mentioned earlier, we intend our technique to provide good starting point for the analysis of chronofaunas.

It would be interesting to see what the ecological characteristics of the candidate chronofaunas found by the method are. Computationally, the candidate chronofaunas resembled the Pikermian chronofauna. For small mammals the temporal span of the CCFs was longer than for larger ones; one can speculate on the relationship of this with the results on the longer lifespans of genera of small mammals (Liow *et al*. 2008).

An interesting experiment would be to repeat the computation for each site in turn, and see whether the resulting chronofaunas have some higher order characteristics that correlate with the MN units.

The method required setting a few parameters (which sites are considered as defining sites, how much deviation from the expected is required for a count of common taxa to be considered significant, and what fraction of occurrences is required for taxa). As the goal was to yield interesting viewpoints to the data, we recommend future users of the method to experiment on the suitable choices for these parameters.

The locations of the sites were not explicitly taken into account in the method. It would be easy to incorporate this into the process of selecting the candidates, but such decisions fit well into the post-processing stage of evaluating the interestingness of candidate chronofaunas.

Our approach views the taxa as unrelated. It would be interesting to see whether a gradual change in the taxonomic content of a chronofauna could be identified by using purely computational methods. A possible tool would be, e.g., singular value decomposition (SVD) that in information retrieval applications was shown to be able to determine similarities between words from 0-1 matrix data (Deerwester *et al*. 1988).

From a computational point of view we were able to observe the gradual change of the taxonomic composition in the sites of the candidate chronofaunas. The possible paleontological significance of the candidate chronofaunas remained, of course, completely open in this study.

Our method is on purpose blind to all other information except the presence-absence data (and the MN classes of the sites). The same type of approach was used in Fortelius *et al*. (2006) for seriation. A general interesting question in paleontological data analysis (and data analysis in general) is the interplay of the domain knowledge and general techniques.

References

Eronen, J. T., Ataabadi, M. M., Micheels, A., Karme, A., Bernor, R. L. & Fortelius, M. 2009: Distribution history and climatic controls of the Late Miocene Pikermian chronofauna. — *Proceedings of the National Academy of Sciences* 106: 11867–11871.

Deerwester, S., Dumais, S., Landauer, T., Furnas, G. & Beck, L. 1988: Improving Information Retrieval with Latent Semantic Indexing. — In: Borgman, C. L. & Pai, E. Y. H. (eds.), *Proceedings of the 51st Annual Meeting of the American Society for Information Science*: 36–40. Learned Information, Medford, New Jersey, U.S.A.

Fortelius, M. (coord.) 2011: *New and Old Worlds Database of Fossil Mammals (NOW)*. — University of Helsinki. [Available at http://www.helsinki.fi/science/now].

Fortelius, M., Gionis, A., Jernvall, J. & Mannila, H. 2006: Spectral ordering and biochronology of European fossil mammals. — *Paleobiology* 32: 206–214.

Jaccard, P. 1912: The distribution of the flora in the alpine zone. — *New Phytologist* 11: 37–50.

Liow, L. H., Fortelius, M., Bingham, E., Lintulaakso, K., Mannila, H., Flynn, L. & Stenseth, N. C. 2008: Higher origination and extinction rates in larger mammals. — *Proceedings of the National Academy of Sciences* 105: 6097–6102.

Mäkinen, E. & Siirtola, H. 2005: The barycenter heuristic and the reorderable matrix. — *Informatica* 29: 357–363.

Olson, E. C. 1952: The evolution of Permian vertebrate chronofauna. — *Evolution* 6: 181–196.

Simpson, E. H. 1949: Measurement of diversity. — *Nature* 163: 688.

Sugiyama, K., Tagawa, S. & Toda, M. 1981: Methods for visual understanding of hierarchical system structures. — *IEEE Transactions on Systems, Man and Cybernetics* 11: 109–125.