

## Integrating Scientific Information: a Case Study from Neuroscience

Partha P Mitra  
Cold Spring Harbor Laboratory

The need for information integration is keenly felt in neuroscience research. There are 344 neuroscience journals indexed by Thomson ISI, 40,000 papers each year on Pubmed, and 177 disparate databases listed currently on the Society for Neuroscience website. To make matters worse, there is no agreed upon central conceptual framework in the field of neuroscience (as compared to say cellular or molecular biology), almost no standardization of data types, and terminology varies from laboratory to laboratory. It is a truism that understanding brain function is a “final frontier” for modern science, so the related information integration problems are well worth solving.

This white paper presents a case study that illustrates the issues involved, strategies that have worked and problems that await resolution. The example chosen involves the study of song development in the zebra finch<sup>1</sup>. The zebra finch is of interest to neuroscientists as a model system for studying the neural basis of vocal learning. A male zebra finch chick, if exposed to song from an adult tutor during a sensitive period, learns to make a copy of the tutor song. The song development process takes about three months. A principle reason for interest in this model system is that the areas of the brain involved in the song development are well characterized and have been studied using a variety of methods. The study in question initially involved continuous acoustic recordings of song throughout the three month period from ~100 birds, generating large data sets (~10 terabytes of raw data). More recent work has included other data types (electrical recordings from the brain, pressure and flow measurements from the vocal apparatus, and video recordings). While it is an ongoing project, the study has given rise to a number of findings, and in each case information integration has played an important role. We will discuss in further detail two issues: integration across multiple time scales, and integration across heterogeneous data types.

The data in question span about ten orders of magnitudes in time scale: there is meaningful structure in both the acoustic and neural recordings at a millisecond scale, while the recordings span three months. There is hierarchical structure at several intermediate scales: songs are composed of notes which are formed into bouts; bouts show diurnal patterns and the song shows a developmental time course over days to months. This large hierarchy of timescales is not unusual, and may also be found in technological settings, such as in internet traffic. Similar problems arise if one were to make longitudinal recordings of human behavior (audio or video) over extended periods of time. The strategy that we found effective, was to identify the natural scales of the temporal hierarchy (sub-second, seconds, minutes, hours and days), and to condense the information at each scale into a small set of features determined using a combination of domain knowledge and exploratory analysis. Feature time series obtained at each scale were treated using standard signal processing methods including spectral analysis, as well

---

<sup>1</sup> Tchernichovski, O., Mitra, P.P., Lints, T., Nottebohm, F. (2001) Dynamics of the vocal imitation process: how a zebra finch learns its song. *Science*, 291 (5513) 2564-2569.

as segmentation and clustering. Treating the full data set at a single scale would have been both intractable and non-informative. Note also that different phenomena were observed at different scales, so that scale-invariant or methods would not have fared well either. Finally, separating data into scales also aided integration of the different data types. That large data sets need to be broken up into a hierarchy of scales is not in itself surprising; however, two things are of note. A naïve application of multi-scale methods (such as wavelets) would not in this case yield good results, since the data in question do not have scale invariance. Secondly, we were able to use more or less the same methods at each scale. This leads us to an interesting conclusion: in breaking large data sets up into hierarchies, techniques may be 'scale-invariant' even if the data themselves are not.

While we have had success in integrating across time scales, integrating across heterogeneous data types has been harder. Our current database of raw data files is ~10 terabytes, stored as a Unix file system. We are currently evaluating relational databases versus semantic web technologies for building an integrative analysis platform for our data sets, but neither seem optimal. The ultimate goal is to achieve as much automation as possible in the data analysis: ideally, we would have a software agent perform the majority of the analysis tasks. We currently have meta data about the experiments stored in ad hoc formats (this is typical of neuroscience laboratories). One option is to build an entity-relation model of the existing meta data and store the information in a relational database. The advantage of this approach is that we can access well worked out, robust software tools. The disadvantage is lack of flexibility: if the experiments, and therefore meta data, change in the future (as is common in a neuroscience laboratory), then the relational database may have to be redone. More importantly, as we perform some analysis on the data sets, we generate new types of meta data on an ongoing basis, again causing the existing relational model to break. Semantic web techniques are attractive precisely because of their flexibility and extensibility. However, we are finding it difficult to locate correspondingly robust and efficient software tools for the usual database operations. Semantic web methods have been proposed to allow for better searching and extraction of information from web pages; we feel they will also be useful for integrating heterogeneous data types within a laboratory or across laboratories in neuroscience, but this will require the development of a body of software that performs database management operations robustly and efficiently. Work will also be required to build up the domain specific structured vocabularies and data models; however, this in and of itself will not solve the problems we are facing until better management tools are available.

While we have focused on a specific case study, the problems we have encountered are quite similar to problems facing other laboratories. Neuroscience provides a rich application area for computer scientists working in the area of information integration. It is to be noted that such applications can only be expected to be successful if carried out in close collaboration with an active laboratory or research group. On the other hand, more mature general purpose tools for dealing with heterogeneous data sets, perhaps using semantic web concepts, will also be useful.