Proceedings of the UK e-Science All Hands Meeting 2007, Nottingham, UK, 10th-13th September 2007, ISBN 978-0-9553988-3-4

The CARMEN e-Science pilot project: Neuroinformatics work packages.

L.S. Smith[1], J. Austin[2], S. Baker[3], R. Borisyuk[4], S. Eglen[5], J. Feng[6], K. Gurney[7], T. Jackson[2], M. Kaiser[3], P. Overton[7], S. Panzeri[8], R. Quian Quiroga[9], S.R. Schultz[10], E. Sernagor[3], V.A. Smith[11], T.V. Smulders[3], L. Stuart[2], M. Whittington[3], C. Ingram[3].

[1]University of Stirling, [2]University of York, [3]Newcastle University, [4]University of Plymouth, [5]University of Cambridge, [6]University of Warwick, [7]University of Sheffield, [8]University of Manchester, [9]University of Leicester, [10]Imperial College, [11]St. Andrews University.

Abstract

The CARMEN (Code Analysis, Repository and Modelling for e-Neuroscience) project aims to apply e-Science technology to Neurophysiology to enable the sharing of data resources and the services for processing this data. This paper describes the application oriented services: the companion paper [1] describes the technology underpinning this application. Overall, the CARMEN project is developing a Neuroinformatics resource which will change the way data is processed in experimental Neuroscience by providing archiving and a set of powerful services, supporting collaborative neurophysiology research.

1. The CARMEN project

The CARMEN project is an e-Science pilot project which started in October 2006, and is funded for four years by the UK EPSRC. The primary aim of the project is the establishment of (i) data and metadata repositories for Neuroscience data, and (ii) a set of expandable services which can be composed together that operate on that data.

A prototypical Neuroscience experiment consists of some neural tissue (in vivo or in vitro), instrumented in some way. The most common measurement technique is electrophysiology, where voltage or current is measured by a (very small) electrode. The electrode may be intracellular, but is more commonly extracellular. Frequently many electrodes are used simultaneously (a multi-electrode array or MEA is used). Signals are small, and often arise from electrical activity in a number of neighbouring neurons. Many neurons signal using action potentials (otherwise called spikes), and there is considerable interest in how these spikes reflect the processing in the neural tissue. In addition to electrophysiological measurements, there are also optical measurements that can be made. These use dyes which are sensitive to ionic concentrations (particularly Ca^{++}), or to voltage. Imaging from these dyes can produce measurements which reflect activity over an area of the tissue.

The project has a "hub and spoke" architecture, with seven work packages. One work package (the "hub", Work Package 0 (WP0) described in [1]) aims to provide the hardware and software underpinning the repositories and the services. The other work packages, WP1 - WP6 are Neuroinformatics based and aim to provide and utilise services, and to enable the use of these services in a straightforward way by both computational and experimental neuroscientists. The "spoke" work packages are concerned with (WP1) spike detection and sorting, (WP2) information theoretic analysis of signals, (WP3) automated parameter estimation in conductance based neural models, (WP4) intelligent database querying, (WP5) measurement and visualisation of spike synchronisation, and (WP6) multilevel analysis and modelling in networks.

After a brief discussion of the structure of the CARMEN project, the paper is organised by work package.

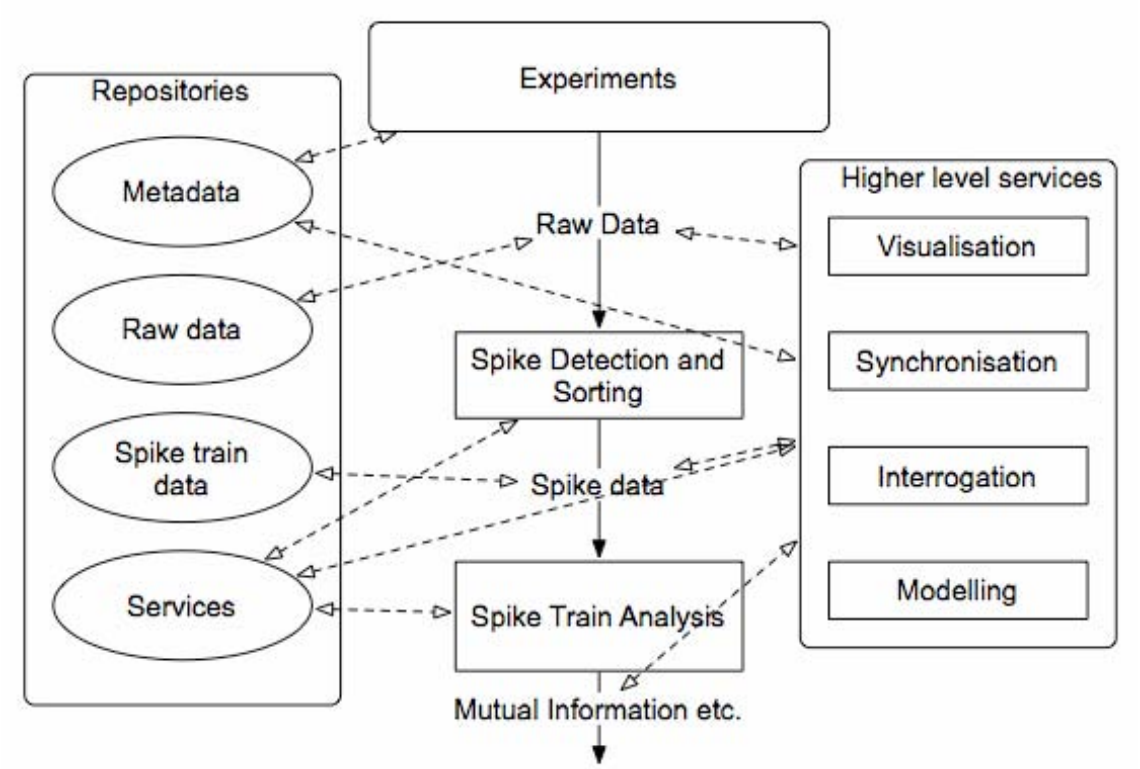

Figure 1: Overall organisation of the CARMEN project. The infrastructure for the repositories is WP0. The actual data comes from different work packages. Metadata and raw data come directly from the experiments (or from data supplied by experimenters either within or outside of the project). The basic services of spike detection and sorting create spike train data sets for the repository, and are being developed by WP1. Spike train analysis provides another set of services for application to spike trains, and is being developed by WP2. Higher level services operate on the data produced by these services (perhaps using workflows created from these services) to provide interpretation and interrogation facilities. WP3 is concerned with developing constraints for neural models, and will use the different modalities of data within the repository to achieve this. WP4 aims to provide interrogation services (capabilities), including content-based searching. WP5 is concerned with providing visualisation services (which will use all the modalities of data within the repository). WP6 is more integrative in scope, and aims to provide high quality raw data and metadata for the repository, as well as creating workflows for data interpretation and multi-site collaboration, both in on-line and off-line experiments. In addition, it aims to supply high-level statistical techniques (such as Bayesian inference) as services, for discovering structures within the data

2. The CARMEN Workpackages

Figure 1 provides an overall view of the project. The six "spoke" work packages support the overall goals of the project in different ways. WP1 aims to provide the basic services which are required for further interpretation of electrophysiological signals. WP2 aims to develop information theoretical analyses of the spiking data provided by the services in WP1. WP3 is about constraining models of neurons using the data from the repository. WP4 aims to use previously developed technology to permit the interrogation of both the raw data and the spiking data. WP5 aims to permit both the measurement of the degree of synchronisation of spikes across neurons, and the visualisation of the data. WP6 is more integrative in scope: it aims both to provide high quality data to the repository, and to combine both electrophysiological recording and optical (dye-based) recording techniques. In addition, it aims to use the tools developed in WP 1, 2, 4, and 5 to investigate in vivo and in vitro neural systems, and to develop new dynamic Bayesian network algorithms to trace paths of neural information flow. The overall architecture is service-oriented: that is, the work packages will

provide a set of services which will be deployed at the hub. These services will be able to be composed by users (using a workflow service

2.1 Work Package 1: Spike detection and sorting (LS Smith, Quian Quiroga)

One major class of users of the CARMEN system are experimental neurophysiologists who collect neurophysiological datasets using single and multiple electrode array systems: most of the datasets are extracellular recordings. The first critical steps in extracting accurate spike (action potential) sequences from these signals is the ability to resolve spikes generated by one neuron from the background noisy electrical signal generated by distant neurons or other sources (spike detection), and the ability to classify each spike as arising from one particular close-by neuron (spike sorting). The aim of this work packages is (i) to enable existing techniques to be deployed on the hub system as services, and (ii) to develop new and better techniques, and to enable these to be deployed on the hub system.

This research builds on methods for assessing different spike detection and sorting techniques, and partly on developing novel spike detection techniques which are more robust against the types of noise most likely to be found in real experiments. The assessment technique has been based on the generation of data using a biophysically justified model developed in MATLAB [2], as well as on the use of real datasets from elsewhere in the CARMEN project. Work on novel spike detection systems based on the techniques described in [3] and on mathematical morphology techniques [4] (more often applied to image processing, but useful here as a method of noise reduction) is proceeding, and these, and other existing techniques, are being assessed on both synthetic and real datasets, as well as being recoded suitably for provision as services.

The other thread of this research is focused on the optimization of a recently proposed spike-sorting algorithm developed by Quian Quiroga [5]. The method combines the wavelet transform, which localizes distinctive spike features, with superparamagnetic clustering [6], which allows automatic classification of the data without assumptions such as low variance or Gaussian distributions. The algorithm is unsupervised and fast, which makes it suitable for the analysis of large datasets from multiple electrode recordings. Current work is on the optimization of implementation details so that the algorithm will be useful for the analysis of other datasets from researchers of the deployed at the hub), and facilities for users to develop and arrange the deployment of further services will be provided. CARMEN project. Moreover the algorithm will be extended to the analysis of data from tetrode arrays in which the electrodes are very close to each other so that signals from the same neuron may be recorded on more than one electrode.

2.2 Work Package 2: Information Theoretic Analysis of Electrically- and Optically-Derived Signals (Schultz, Panzeri).

Multi-electrode recording techniques are increasingly allowing neuroscientists to record action potentials from large populations of neurons. At the same time, recently developed optical recording approaches (such as two-photon imaging) allow action potential induced calcium signals to be imaged in essentially all of the neurons in a region of tissue, *in vivo*. The combination of these technologies offers particular promise. There are, however, many challenges inherent to the analysis of large datasets, particularly when the data is acquired from recent technologies such as two-photon imaging: processing and analysing such datasets, to the point of testing useful scientific hypotheses, is a non-trivial task. This is the challenge WP2 sets out to address, with the assistance of CARMEN's e-Science technology.

There are two principal challenges faced in the work package, as illustrated in Figure 2 below. The first is the processing of disparate electrical and optical signal types, in order to convert them into a common representation to which existing and novel data analysis algorithms can be applied. This will require the development and implementation of procedures for automatic selection of pixels corresponding to single cell activity (e.g. by techniques based on independent component analysis), and then where appropriate detecting events using spike detection algorithms. The second challenge is to analyse the resulting high dimensional datasets using Information Theory. The Investigators in this Work Package have been at the forefront of the development and application of methods for alleviating the sampling problems inherent to measuring information, and were the first to address in a systematic way the role of spike timing and spike correlations in neuronal population coding [7, 8, 9]. This previous work will aid the design and implementation of procedures (and hence services) allowing the information content and statistical structure of the different signals to be quantified.

Panzeri's group will work on developing novel methods for the estimation of information

carried by neural populations; in parallel, the Schultz group will work on applying these information-theoretic methods to two-photon imaging data. This will involve the acquisition of new experimental data using a two-photon microscope at Imperial College, as well as solving the particular computational problems involved in applying the algorithms to this new type of neuroscience data.

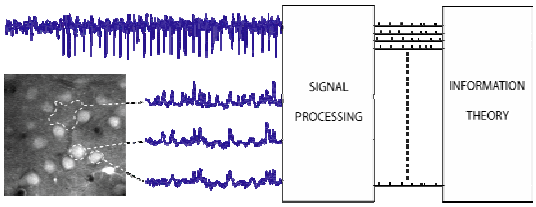

Figure 2: The research carried out in this WP2. Electrical (upper) and optical (lower) signals are recorded from single neurons and from neuropil in the intact brain, while sensory stimulation is occurring. In the first analysis block, the different signals are processed in order to convert them into a common representation. In the second block, information theoretic analysis procedures are applied in order to decompose the sensory representation in terms of the statistical structure of the signals.

2.3 Work Package 3: Automated search techniques for parameter estimation in conductance based neural models (Gurney, Overton)

The WP aims to: (i) refine and extend a technique (initially developed at the University of Sheffield), which automates the search for key parameters used in building biologically realistic neuronal models [10]; (ii) deliver a general purpose, useable service (rather than simply 'research code') with all necessary documentation; (iii) promote the donation of current clamp data to the CARMEN repository; (iv) start to build a canonical set of conductance based models using the tools developed, and the data garnered.

Neuronal modelling takes place at many levels of description – from simple leaky integrate and fire neurons, through so-called 2-dimensional 'reduced models', to large scale conductance-based models. While it is likely that all levels of description will continue to play a role in computational neuroscience, this work package addresses the most detailed level – that of multi-compartment, conductance-based models. Such models are necessary for understanding synaptic processing and dendritic signal integration - key elements in a complete description of neural computation.

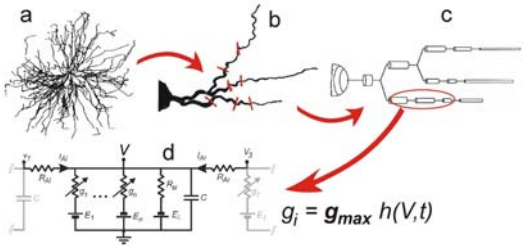

Figure 3: From neuron to parameterised multi-compartment model. See text for details.

One of the principal difficulties in developing conductance-based models is establishing suitable values for the extensive set of parameters required. A typical model build might proceed as shown in figure 3. Morphology (a) is idealised into a series of compartments (b) and (c) which are then modelled using a set of equivalent electrical circuits (d). The key here is establishing the ionic channel conductances g_i. These are expressed in terms of maximal *conductances* g_{max}, and dynamic or *kinetic* terms, $h(V,t,\zeta)$, which are functions of membrane voltage, V, time, t, and parameter sets ζ. These kinetic parameters are determined by the molecular structure of the ion channel and may be found either in the literature, or in online databases. However, the maximal conductances will depend critically on the morphology of the particular cell for which the modeller has physiological data (in the form of membrane potential deflections under current clamp). Therefore, these parameters must be *fitted* within the model to the available data set, rather than grafted on from other studies. Hand fitting here is very labour intensive and does not, therefore, encourage the exploration of the model space. General purpose search techniques are often stochastic and extremely compute intensive. However, based on the observation that the neural membrane equation is linear in the g_{max} it is possible to develop deterministic methods using comparatively simple optimisation techniques. Pilot work using this approach will be extended to deliver a powerful, general purpose modelling tool that will be part of the CARMEN repository.

2.4 Work Package 4: Intelligent Database Querying (Jackson, Austin).

The raw data to be stored by CARMEN consists of spatiotemporal signals expressed, either as time-series recordings collected by single electrodes or MEAs, or as image files, collected at regular intervals using various optical recording techniques. The data model to be implemented by CARMEN will allow these data structures to be integrated across space and time, so that experiments that make use of combined techniques can intuitively co-register data, and will ensure that derived data (for example, the products of analysis) are bound to both the raw data source and the experimental metadata.

CARMEN will challenge and extend e-Science and neuroscience capabilities by allowing users to query these raw and derived data for abstracted (e.g. cut from existing data) or simulated (e.g. modelled or drawn) activity patterns that are of interest. This might, for example, allow the modeller to programmatically perform sweeping, serial iterations to improve the biological accuracy of their simulations, through interaction with real data, or allow neuroscientists to accurately correlate neuronal firing patterns with specific events. As well as providing this access to data CARMEN will play a significant role in allowing theoreticians to accelerate the identification of flaws and limitations in hypotheses and algorithms.

Querying will be provided at two levels: at a high level by storing the derived data in a database to utilise spatiotemporal pattern matching functions in the database native query language and at the "signal" level by applying and extending the utility of Signal Data Explorer (SDE) [11]. The SDE was originally developed by the DAME [12] project to search for patterns in temporal signals of vibration data across distributed repositories. The main view of the SDE is shown in figure 4 whilst being used to explore and search retinal data (data supplied by Evelyne Sernagor). A user can search using example patterns (highlighted in blue below), using previously stored patterns, by drawing patterns and by the construction of complex patterns (built from a combination of the previously described patterns).

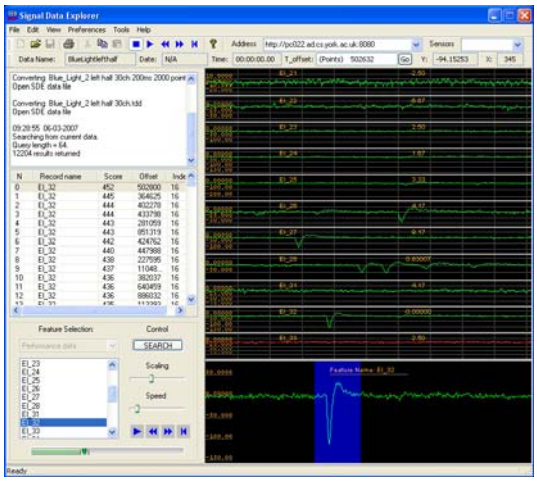

Figure 4: Using Signal Data Explorer to explore and search retinal data.

The SDE currently can search local and distributed databases, apply a range of data filters and provide various viewing modes to complement the main view shown above. A view that allows the user to see "wide" spatial and or temporal views is also available – this can be viewed simultaneously with the main view and is useful in providing macro and micro views of the data as well as allowing the user to "play and zoom" large data sets.

The facilities provided by the SDE will be significantly enhanced during the CARMEN project to increase its utility to the neuroscience domain, for example, to provide spike detection, etc. The WP4 research programme covers four main strands:

- Improvement of the intelligence of the SDE search process and its two distributed software search components: Pattern Match Control (PMC) and Pattern Match Engine (PME).
- Extension of the PME service to search for patterns in both raw signal and derived data (e.g. spike trains) discretely and simultaneously.
- Evaluation of techniques to allow the tool to interact with the spatiotemporal query functionality that is provided by modern RDBMS systems.
- Further development of the user interface based on the SDE's ability to allow users to draw the pattern that they wish to search for. This will involve high performance desktop visualisation in order to allow the user to harness the SDE by simulating or extrapolating from results that are displayed in the visualisation interface.

2.5 Work Package 5: Measurement and Visualisation of Spike Synchronisation (Baker, Aertsen (University of Freiburg), Borisyuk, Feng, Gerstein (University of Pennsylvania), Stuart).

Whereas traditional single unit neurophysiology has measured neural firing rate, the use of multiple electrode techniques permits the examination of firing relationships between cells. This opens up new possibilities: higher-order coding schemes where information is represented by the coordinated firing of small populations of neurons are a good candidate for exploration. A particular type of such coding which has received much interest is synchronisation – when neurons fire nearly simultaneously. However, this also poses a technical challenge, as the assessment of spike synchrony requires novel analysis techniques. WP5 will develop suitable methods and integrate them with the e-Science infrastructure, thereby making them available to all neuroscientists.

WP5 will build on existing methods already developed by the investigators:

- Time-resolved cross-correlation (see figure 5)
- Spatiotemporal repeating pattern detection
- Gravitational clustering [13]
- MANOVA for detection of synchronous neural assemblies [14]
- Non-linear Granger causality analysis
- Correlation distance

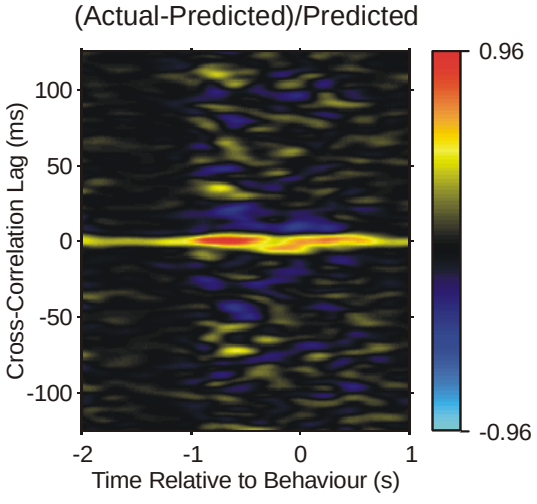

Figure 5: Time resolved cross-correlation. Normalisation to remove correlation expected by chance, and assessment of the statistical significance of any features, is a challenging problem in the face of non-stationary neural data.

In all cases, advanced statistical methods are required to exclude synchronisation which occurs by chance. Often, Monte Carlo methods must be used; the consequent heavy computational requirements would prevent routine adoption by the neuroscience community. However, once the methods developed by WP5 are implemented on the CARMEN infrastructure, users will benefit from the substantial computing resource of the core processors.

Sophisticated analysis techniques often present challenges in the visualisation of results; a key component of WP5 is the development of new methods to display and interact with data. By running these over the e-Science network, users will be able to explore results using complex graphical structures in real time.

WP5 seeks to compress the experiment-to-analysis cycle by developing technology to stream multiple electrode data to the e-Science network in real time. Experimenters will then be able to benefit from near-instantaneous analysis of their data, allowing adjustment of experimental parameters based on preliminary results. This will also facilitate long distance collaboration: colleagues at a distant site will be able to view data as it is captured, and provide advice on experimental optimisation. Finally, this WP aims to 'close the loop', and to develop algorithms which will automatically move electrodes to find and maintain clean recordings based on the results of real-time analysis.

2.6 WP6: Multilevel Analysis and Modelling in Networks (Sernagor, Whittington, Kaiser, Eglen, VA Smith, Smulders).

Understanding activity dynamics within neuronal networks is a major challenge in neuroscience that requires simultaneous recording from large numbers of neurons. In addition, two major types of activity can contribute to understanding how neurons communicate within these networks. Cutting-edge experimental techniques, such as imaging and multi-electrode arrays (MEAs), can generate a comprehensive picture of the spatiotemporal dynamics of network activity. This part of the project is developing analytical techniques to resolve coordinate activity within large networks in cortex, cerebellum, hippocampus and retina. These techniques will allow integration of data from different recording paradigms. The three primary goals are: (i) integration of existing and novel network analysis techniques into CARMEN in order to build comprehensive models of

network dynamics; (ii) populating the CARMEN repository with data of exceptional quality and detailed provenance for analysis of network properties; and (iii) development of new dynamic Bayesian network algorithms to trace paths of neural information flow in networks.

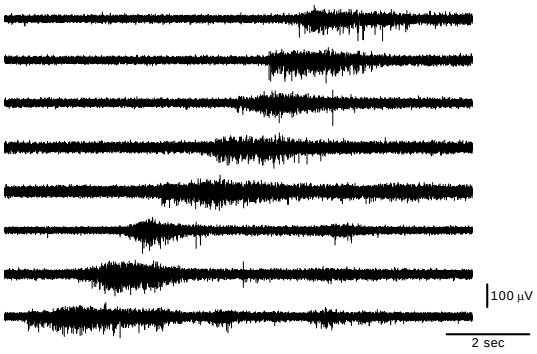

Figure 6: Recording using a multi-electrode array (MEA) (allowing recording from many adjacent sites: electrode separation is 200µm, and electrode diameter is 30µm) from a turtle retina (age 1 month post hatching). The retina has been mounted flat on a 60 electrode MEA, with the ganglion cell layer facing down on the electrodes. What is shown is a spontaneous wave sweeping across one column of electrodes.

This integrative work package requires a broad range of inter-related and inter-disciplinary skills. Sernagor is studying the cellular mechanisms underlying oscillatory propagating activity patterns in the developing retina [15, 16]. Figure 6 shows an example of the data for the CARMEN repository from this work. Whittington is designing *in vitro* models of EEG rhythms and understanding network mechanisms for neuronal population behaviour associated with sensory processing. Kaiser is working on modelling and network analysis; analysis of the spatial and topological organisation of correlation networks and of network changes over time [17]. Eglen is developing analysis and visualization tools for studying spontaneous neural activity; investigating network properties and how they change over time [18]. VA Smith is developing dynamic Bayesian network inference algorithms for recovering neural information flow using data from arrays of single-unit as well as multi-unit recordings [19]. Smulders will contribute multi-electrode array recordings from rat hippocampus and from the avian brain.

3. Discussion

CARMEN's aim of applying e-Science based Neuroinformatics to neuroscience datasets is timely because of (i) the low (and decreasing) cost of machines and storage, particularly when compared with the costs of experimental neuroscience, (ii) the effective physical underpinning of UK e-Science (both network infrastructure, and infrastructure supported by the e-Science centres), and (iii) the ability to build on to successful earlier e-Science projects both for the CARMEN infrastructure [1] and for searching [11]. In addition, the recently established International Neuroinformatics Co-ordinating Forum (which the UK has now joined) should make it possible to extend this UK funded project internationally.

It will be necessary for the project to become financially independent after the four years of the project. Even at this relatively early stage, discussions are ongoing with industrial partners, journals and the research councils on this topic.

Acknowledgements

This work was supported by the EPSRC, grant number EP/E002331/1. Thanks are also due to the other members of the CARMEN consortium.

References:

[1] P. Watson, T. Jackson, G. Pitsilis, F. Gibson, J. Austin, M. Fletcher, B. Liang, P. Lord, The CARMEN Neuroinformatics server, this meeting.
[2] L.S. Smith, N. Mtetwa (2007), A tool for synthesizing spike trains with realistic interference, J. Neurosci. Methods, 159, 170-180.
[3] N. Mtetwa, L.S. Smith (2006), Smoothing and thresholding in neuronal spike detection, Neurocomputing, 69 , 10-12, 1366-1370.
[4] J. Serra (1983), Image Analysis and Mathematical Morphology, Academic Press.
[5] R. Quian Quiroga, Z. Nadasdy, Y. Ben-Shaul (2004) Unsupervised spike detection and sorting with wavelets and superparamagnetic clustering. Neural Computation 16:1661-87.
[6] M. Blatt, S. Wiseman, E. Domany (1996), Super-paramagnetic clustering of data, Physical Review Letters, 76, 3251.
[7] S. Panzeri, S.R. Schultz, A. Treves and E.T. Rolls (1999). Correlations and the encoding of information in the nervous system. Proceedings of the Royal Society of London Series B: Biological Sciences, 266: 1001-1012.

[8] S. R. Schultz and S. Panzeri (2001). Temporal correlations and neural spike train entropy. Physical Review Letters, 86(25): 5823-5826.
[9] S. Panzeri and S. R. Schultz (2001). A unified approach to the study of temporal, correlational and rate coding. Neural Computation, **13**(6), 1311-1349.
[10] M.D. Humphries, K. Gurney (2006), A means to an end: validating models by fitting experimental data, Neurocomputing, in press.
[11] Signal Data Explorer: http://www.cybula.com/flyers/SignalData.pdf.
[12] Distributed Aircraft Maintenance Environment Project (DAME): http://www.cs.york.ac.uk/dame/.
[13] S..N. Baker, G.L. Gerstein (2000): Improvements to the Sensitivity of Gravitational Clustering for Multiple Neuron Recordings. Neural Computation 12(11): 2597-2620
[14] P.M. Horton, L. Bonny, A.U. Nicol, K.M. Kendrick, J.F. Feng (2005) Applications of multi-variate analysis of variances (MANOVA) to multi-electrode array data, *Journal Of Neuroscience Methods*, 146 22 – 41
[15] E. Sernagor, S. Eglen, B. Harris, R. Wong (editors) (2006) Retinal Development, Cambridge University Press.
[16] E. Leitch, J. Coaker, C. Young, V. Mehta, E. Sernagor (2005) GABA type-A activity controls its own developmental polarity switch in the maturing retina. J Neurosci 25: 4801-4805.
[17] O. Sporns, D. Chialvo, M. Kaiser, C.C. Hildetag (2004) Organization, Development and Function of Complex Brain Networks, Trends in Cognitive Sciences, 8: 418-425
[18] J. Demas, S.J. Eglen, R.O.L. Wong (2003) Developmental loss of synchronous spontaneous activity in the mouse retina is independent of visual experience. J Neurosci. 23:2851-2860.
[19] V.A. Smith, J. Yu, T.V. Smulders, A.J. Hartemink, E.D. Jarvis (2006) Computational inference of neural information flow networks PLoS Computational Biology 2:e161