Bioinformatics and Neuroinformatics

Chapter 5

Bioinformatics

Samuel Kaski, Janne Nikkilä, Merja Oja, Jaakko Peltonen, Jarkko Venna, Antti Ajanki, Andrey Ermolov, Ilkka Huopaniemi, Arto Klami, Leo Lahti, Jarkko Salojärvi, Abhishek Tripathi

5.1 Introduction

New so-called high-throughput measurement techniques have made possible genome-wide studies of gene function. Gene expression, gene regulation, protein content, protein interaction, and metabolic profiles can be measured and combined with the genetic sequence. The methods are used routinely in modern biology and medicine, and now the current challenge is to extract meaningful findings from the noisy and incomplete data masses, collected into both community resource and private data banks. The data needs to be analyzed, mined, understood, and taken into account in further experiments, which makes data analysis an integral part of biomedical research. Successful genome-wide analyses would allow a completely novel systems-level view into a biological organism.

Combining the different kinds of data produces new systems-level hypotheses about gene function and regulation, and ultimately functioning of biological organisms. We develop probabilistic modeling and statistical data analysis methods to advance this field. Our main novel contributions stem from the cross-breeding of the methodological basic research, in particular on Modeling of Relevance, and collaboration with top groups in Biology and Medicine. We have had long-standing collaboration with Laboratory of Cytomolecular Genetics (Prof. S. Knuutila) and Neuroscience Center (Prof. E. Castrén), University of Helsinki, University of Uppsala (Prof. J. Blomberg), Turku Centre for Biology (Doc. T. Aittokallio), VTT (Prof. M. Oresic), and smaller-scale collaboration with several other groups. During 2007 we started new projects with EBI, UK (A. Brazma) and Finnish CoE in Plant Signal Research, University of Helsinki (Prof. J. Kangasjärvi) with promising results that will be reported in the next biennial report.

In 2006 we started a new conference series in collaboration with Prof. E. Ukkonen and J. Rousu of University of Helsinki. The conference "Probabilistic Modeling and Machine Learning in Structural and Systems Biology" inspired a special issue in a main journal, and yearly conferences in Evry, France, in 2007, and in Belgium in 2008.

References

[1] Juho Rousu, Samuel Kaski, and Esko Ukkonen, editors. *Probabilistic Modeling and Machine Learning in Structural and Systems Biology. Workshop Proceedings; Tuusula, Finland, June 17-18*. Helsinki, Finland, 2006.

[2] Samuel Kaski, Juho Rousu, and Esko Ukkonen. Probabilistic modeling and machine learning in structural and systems biology. *BMC Bioinformatics*, 8(Suppl 2):S1, 2007.

5.2 Translational medicine on metabolical level

Translational medicine is a research field which attempts to more directly bring basic research findings to clinical practice. One of the necessary steps of this process is to translate inferences made on the molecular level, for example about metabolites, in model organisms into inferences about humans. Such translation is extremely challenging and the existing knowledge, if there is any, is currently largely tacit and only known to experts of the specific disease and model organism.

Metabolomics is the study of the set of all metabolites found in a sample tissue. Metabolite concentrations are affected strongly by diseases and drugs, and hence they complement the genomic, proteomic, and transcriptomic measurements in an excellent way, in studies of the biological state of an organism.

We are in the process of developing new computational methods for translational medicine, for mapping between the observed metabolomics data from model organisms and humans. In project TRANSCENDO we apply the methods to studies of the emergence of Type I diabetes, by computing mappings between non-obese diabetic (NOD) mice and children, and between the effects of a disease in several tissues. The project is collaboration within a consortium involving computational systems biology (Matej Oresic, VTT), semantic modelling (Antti Pesonen, VTT), probabilistic modelling (us), and pharmacology and animal models of metabolic disease (Eriika Savontaus, University of Turku).

Metabolomic development in humans

Metabolic development of children and its differences between the genders is not yet well understood. These dynamic changes may, however, affect strongly the susceptibility to diseases and the responses to drugs.

We are studying a metabolomic data set derived from a collection of blood samples collected during the first years of life from boys and girls. We assume that the metabolic profiles are generated by a set of unobserved metabolic states, and we model those states and the data with a Hidden Markov Model (HMM). HMM fits the assumption of latent states very well and is easy to compute and interpret. Moreover, HMM provides a way for probabilistic re-alignment of the time series, which takes into account the individual variation in the dynamics. Simulations have indicated that HMMs can separate the boys' and girls' metabolic states more efficiently apart than traditional linear method; classification accuracy is 73% for HMM, and under 60% for linear methods. Figure 5.1 presents the model structures for girls and boys.

Disease-related dependencies between multiple tissues

A common setting in medical research is that a disease may be mainly located in a specific organ, for example in lungs, but it indirectly affects multiple tissues. Giving drugs to patients induces an analogous setup: the drugs may affect multiple other tissues in addition to the target tissue (and hence disease). We are developing new methods for discovering the disease-related metabolic dependencies between the multiple tissues, with the goal of revealing potential side effects of the diseases and drugs.

In practice, we have metabolomics data from mice belonging to 4 classes: healthy and untreated, sick and untreated, healthy and treated, sick and treated. A fast and straightforward way of digging out disease-related dependencies is to first find disease-related aspects with partial least-squares classifiers, and then dependencies with canonical correlation analyses and more straightforward correlations between contributing metabolites.

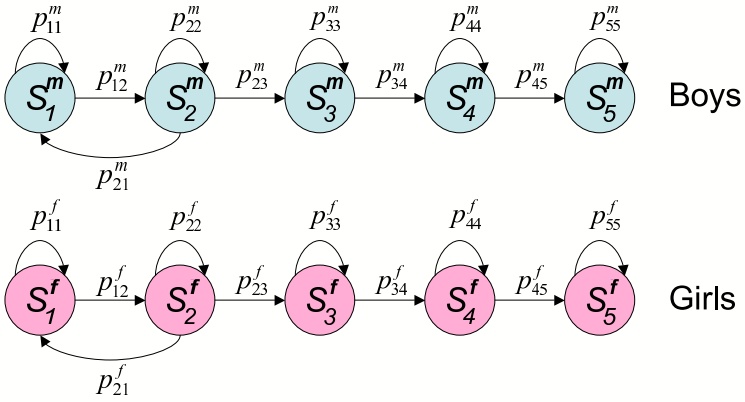


Figure 5.1: HMM models for metabolic states in boys and girls. The nodes represent hidden metabolic states, and the arrows possible transitions. Note that the states form a chain in order to force the models to focus on progressive changes in metabolite concentrations.

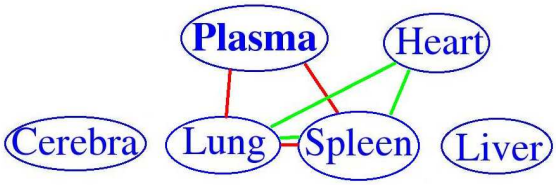


Figure 5.2: Disease-related dependencies between tissues before treatment (red), and after treatment (green). The disease is located in the lungs so the dependencies between lungs and plasma and spleen are logical, but note that after the treatment the dependency with plasma disappears and a dependency to heart emerges. This might be a sign of a side effect of the treatment.

This multivariate approach complements the traditional metabolite-wise linear models. Figure 5.2 shows the dependencies found between tissues before and after drug treatment.

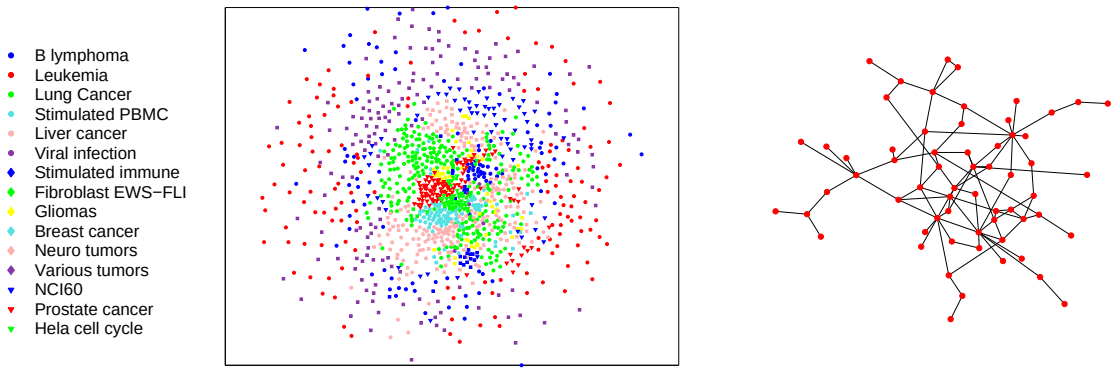


Figure 5.3: *Left:* Sample visualization of a gene expression atlas of cancer samples by curvilinear component analysis. Each dot denotes one microarray; the colors show the cancer class of the sample. *Right:* Part of yeast gene regulatory interaction network visualized by local multidimensional scaling.

5.3 Visualizing gene expression and interaction data

A large community-resource or private gene expression databank consists of numerous data sets submitted by several parties. They may have been measured for different purposes, with different treatments and methods in different laboratories. Several such databanks have been established and they continue to grow. A key challenge is how to best use the databanks to support further research. Currently information in these databanks is accessed using queries on the imperfect meta-data, that is, textual annotations and descriptions. In the future more sophisticated search methods, that take the actual data into account, are needed. Our study [2] aimed at comparing the different methods applicable as a visual interface that reveals similarities of data sets.

We compared several different visualization methods in the task of visualizing a large collection of gene expression arrays. Several new methods have been recently proposed for the estimation of data mani- folds or embeddings, but they have so far not been compared in the task of visualization. In visualizations the dimensionality is constrained, in addition to the data itself, by the presentation medium. It turned out that an older method, curvilinear components analysis, outperforms the new ones in terms of trustworthiness of the projections. Even though the standard preprocessing methods still need to be improved to make measurements of different labs and platforms more commensurable, the good news is that the visualized overview, expression atlas, reveals many of the cancer subsets (Fig. 5.3). Hence, we conclude that dimensionality reduction even from 1339 to 2 can produce a useful interface to gene expression databanks.

Biological high-throughput data sets can also be visualized as graphs that represent the relations between the biological entities. We applied our visualization methods for visualizing gene interaction graphs, and showed that Local Multidimensional Scaling performs very well in this task (Fig. 5.3; [1]).

References

[1] Jarkko Venna and Samuel Kaski. Visualizing Gene Interaction Graphs with Local Multidimensional Scaling In *Proceedings of ESANN'06, 14th European Symposium on Artificial Neural Networks*, pages 557–562, d-side, Evere, Belgium, 2006.

[2] Jarkko Venna and Samuel Kaski. Comparison of visualization methods for an atlas of gene expression data sets *Information Visualization*, 6:139–154, 2007.

5.4 Fusion of gene expression and other biological data sets

While analysis of gene expression data is a corner stone in modern bioinformatics, it is not a sufficient description of cellular state. The cell is an extremely complex system, and gene expression is only a partial view, among all the other omics. Only integration of information from multiple sources can reveal the true potential of the modern high-throughput measurement methods, such as gene expression data.

Integration is not trivial since the data types and scales can vary dramatically. Moreover, what is a proper way of doing the integration depends on the analysis task. Our main novel contribution has been to develop and apply new methods for searching for relevant features by combining data sources (described in Section Modeling of Relevance). We have additionally developed more specific methods for taking into account the known regulatory and context variables in modeling gene expression.

Relevant features through data fusion

We consider a data fusion problem of combining two or more data sources where each source consists of vector-valued measurements from the same object or entities but on different variables. The task is to include only those aspects which are *mutually* informative of each other. This task of including only shared aspects of data sources is motivated through two interrelated lines of thought. The first is noise reduction. If the data sources are measurements of the same entity corrupted by independent noise, discarding source-specific aspects will discard the noise and leave the shared properties that describe the shared entity. The second motivation is to analyze what is interesting in the data. One example is the study of activation profiles of yeast genes in several stressful treatments in the task of defining yeast stress response. In this example what is in common in the sources is what we are really interested in. The "noise" may be very structured; its definition is simply that it is source-specific.

A recent application is search for asbestos-related effects in gene expression by combining several cell lines [3].

We recently showed that there is a simple and computationally fast way of doing data fusion such that shared, relevant features are retained and source-specific noise is discarded [2]. The method is based on the classical canonical correlation analysis; it is surprising that there are still new practically important uses for so old methods! The method has been applied to several gene expression studies: classification of cell cycle regulated genes in yeast, identification of differentially expressed genes in leukemia, and defining stress response in yeast. The software package is available at `http://www.cis.hut.fi/projects/mi/software/drCCA/`.

Modeling context specific gene expression regulation

The biological state of the cell is to a large part defined by which genes are expressed at a certain moment or under certain environmental conditions. Regulation of gene expression is thus the key to understanding, for example, the reasons why some cells become transformed to cancer cells. Regulation of expression has been under intensive study during the past years, but analysis with statistical models has proved to be extremely difficult because the sample sizes are always small due to high measurement costs. The effective sample sizes become even smaller when analyzing context specific regulation, where the data becomes divided according to the context or experimental setup. We have introduced ways of context-specific modeling with one of the most often used model families, the Bays networks.

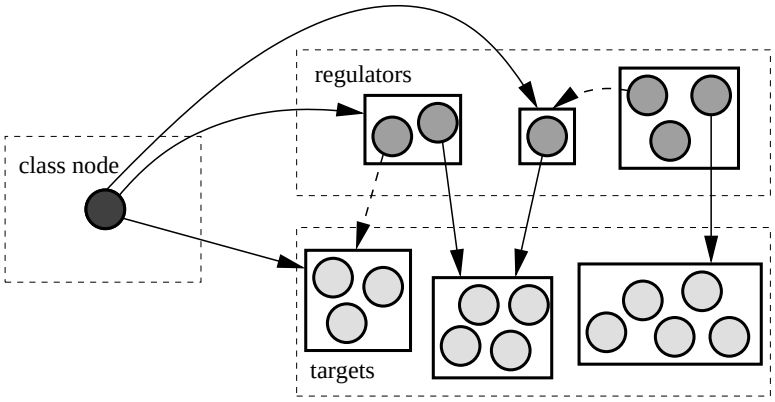


Figure 5.4: The structure of condition-dependent Bayesian network. Similarly behaving genes are grouped into modules. The edges depict the regulatory interactions. The dashed edges indicate interactions that are active in one of the conditions only. Genes linked from the class node may behave differently in different conditions.

The gene regulatory relationships form a complex network in which genes can be regulated by multiple regulators or through long chains of regulatory interactions. The regulatory network adapts to the conditions outside the cell by activating or stopping regulatory interactions in response to changes in the environment. We have studied regulatory networks in yeast with new *condition-dependent Bayesian network* [1]. The data has been divided into several conditions or contexts indicated by a context or class variable that is treated as a covariate. The model has the novel capability of identifying interactions that are active only in subset of conditions. The output of the method is a graphical representation of the network where the possible condition-dependent interactions are highlighted. Figure 5.4 depicts a conceptual example of a condition-dependent network.

We analyzed the regulation in yeast cultures which had been subjected to normal and stressful growth conditions. The method identified 25 regulators which are active only in stressful conditions. The majority of them (20 out of 25) have been annotated stress-related in the literature. The rest are new potential stress regulators.

References

[1] Antti Ajanki, Janne Nikkilä, and Samuel Kaski. Discovering condition-dependent Bayesian networks for gene regulation. In *Proceedings of Fifth IEEE International Workshop on Genomic Signal Processing and Statistics*, 2007.

[2] Abhishek Tripathi, Arto Klami, and Samuel Kaski. Simple integrative preprocessing preserves what is shared in data sources. *BMC Bioinformatics*, 9:111, 2008.

[3] Penny Nymark, Pamela M Lindholm, Mikko V Korpela, Leo Lahti, Salla Ruosaari, Samuel Kaski, Jaakko Hollmen, Sisko Anttila, Vuokko L Kinnula, and Sakari Knuutila. Gene expression profiles in asbestos-exposed epithelial and mesothelial lung cell lines. *BMC Genomics*, 8:62, 2007.

5.5 Human endogenous retroviruses

The human genome includes surviving traces of ancient infections by retroviruses that have become fixed to human DNA. These surviving traces are called *human endogenous retroviruses* (HERVs). HERVs are interesting because they can express viral genes in human tissues, and because their presence in the genome may affect the functioning of nearby human genes. If ancient highly mutated elements are included, HERV sequences form 8% of the human genome [1].

In earlier research we had used Self-Organizing Maps to analyze the classification of HERVs into families [2]. In recent research we have moved to estimating the relative activities (expression levels) of the HERVs across several human tissues. We analyze activity for individual HERV sequences (rather than groups of sequences); this is vital for analyzing their individual control mechanisms and their possible roles in diseased and normal cell functions.

To find evidence of HERV activity, we use probabilistic modeling methods for expressed sequence tags (ESTs) gathered from public databases. We introduced a generative mixture model for EST sequences where each component of the mixture was associated with a particular HERV (see the top subfigure of Fig. 5.5). In our experiments we compared this rigorous model with a fast heuristic method; it turned out that the fast method performed reasonably accurately on simulated data, which made it possible to analyze very large HERV collections.

We first used the models to analyze overall activities across different tissues and conditions. In addition to comparisons on simulated data, we performed several experiments on real HERV data; the probabilistic method for a smaller and the fast method for a larger set having 2450 HERVs [3]. Lastly the probabilistic model was used to estimate tissue-specific expression of HERVs from the HML2 family [4].

Overall, 7% of the HERVs were estimated to be active; the majority of the HERV activities were previously unknown. HERVs with the retroviral *env* gene were found to be more often active than HERVs without *env*. We were also able to analyze which parts of the HERV sequences the EST data match to; see [4] and its supplementary material for figures. For the HERV family HML2, activity profiles of HERVs over tissues are shown in the bottom subfigure of Fig. 5.5; some of the HML2 HERVs display tissue-specific expression (e.g. activity in male reproductive tissues or in the brain).

References

[1] Eric S. Lander *et al.* Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

[2] Merja Oja, Göran O. Sperber, Jonas Blomberg, and Samuel Kaski. Self-organizing map-based discovery and visualization of human endogenous retroviral sequence groups. *International Journal of Neural Systems*, 15(3):163–179, 2005.

[3] Merja Oja, Jaakko Peltonen, Jonas Blomberg, and Samuel Kaski. Methods for estimating human endogenous retrovirus activities from EST databases. *BMC Bioinformatics*, 8(Suppl 2):S11, 2007.

[4] Merja Oja. In silico expression profiles of human endogeneous retroviruses. In *Proceedings of the Workshop on Pattern Recognition in Bioinformatics (PRIB 2007)*, 2007.

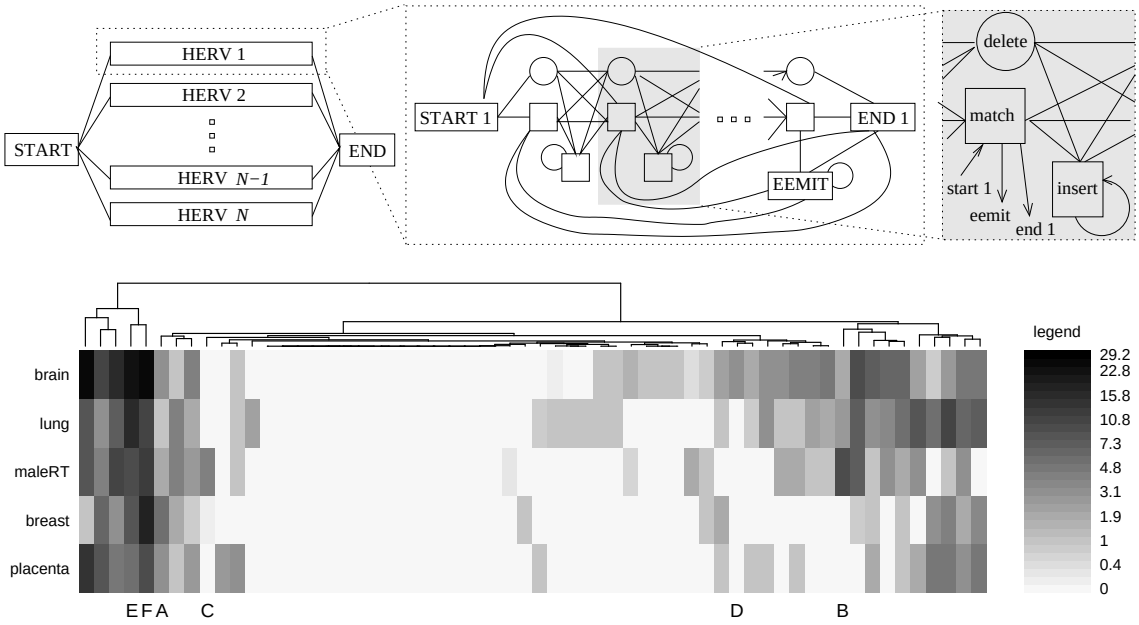


Figure 5.5: Top: the new probabilistic mixture model introduced for estimating activity of human endogeneous retroviruses (HERVs) from expressed sequence tags (ESTs). Bottom: activities of HERVs from the HML2 family in different tissues. Each column depicts the expression profile of an individual HERV sequence; the columns have been ordered by hierarchical clustering based on the profiles. Numbers next to the legend are probabilistic EST counts. Letters A-F at the bottom identify individual HERVs that have been analyzed in [4].

Chapter 6

Neuroinformatics

Ricardo Vigário, Jaakko Särelä, Sergey Borisov, Astrid Pietilä, Jan-Hendrik Schleimer, Jarkko Ylipaavalniemi, Alexander Ilin, Samuel Kaski, Eerika Savia, Erkki Oja

6.1 Introduction

Neuroinformatics has been defined as *the combination of neuroscience and information sciences to develop and apply advanced tools and approaches essential for a major advancement in understanding the structure and function of the brain.* Aside from the development of new tools, often the fields of application include the analysis and modelling of neuronal behaviour, as well as the efficient handling and mining of scientific databases. With the current configuration, the group aims at proposing algorithmic and methodological solutions for the analysis of elements and networks of functional brain activity, studying several kinds of communication mechanisms. These are to be applied in the understanding of ongoing brain activity, as well as responses to natural stimulation.

From a methodological viewpoint, the neuroinformatics group has been involved in studying certain properties of ICA, such as its reliability and applicability to the analysis of electrophysiological brain data (namely electroencephalograms, EEGs and magnetoencephalograms, MEG), as well as to functional magnetic resonance images (fMRI). Within the study of ICA reliability, subspace effects have been made evident, and their potential in functional network interpretability suggested (see Sec. 6.2).

Two explorative studies into functional brain networks have been started. One made explicit use of complex stimulation in fMRI, resulting in the detection of several networks of functional activity with clear interpretability (see Sec. 6.3). Another targeted phase synchrony, which is expected to play a central role in the communication within the central nervous system, as well as between this and the peripheral nervous system (see Sec. 6.4).

Several other topics have been researched in the field of biomedical signal processing, which are not thoroughly reported here. In particular, the denoising source separation framework (DSS) introduced earlier in the laboratory of computer and information science, has been used in the study of phonocardiographic signals, as well as in the investigation of different possible origins for high- and low-amplitude alpha-activity in EEG. We have as well studied measurement fMRI artefacts using a reliable ICA approach with a standard spherical phantom. All of these topics are collected in Sec. 6.5. The application of our methods to tissue segmentation in magnetic resonance imaging (MRI), to the detection of brain lesions will appear in the report of next biennial period.

Research reported in this section has been carried out in collaboration with experts in neuroscience and cardiology.

References

[1] Schleimer, J.-H., and R. Vigário. Clustering limit cycle oscillators by spectral analysis of the synchronisation matrix with an additional phase sensitive rotation. In *Proc. 17th Int. Conf. on Artificial Neural Networks (ICANN'2007)*, Porto, Portugal, pp. 944–953, 2007.

[2] Schleimer, J.-H., and R. Vigário. Order in Complex Systems of Nonlinear Oscillators: Phase Locked Subspaces. In *Proc. of 15th European Symposium on Artificial Neural Networks (ESANN'07)*, Bruges, Belgium, pp. 13–18, 2007.

[3] Ylipaavalniemi, J., E. Savia, R. Vigário, and S. Kaski. Functional elements and networks in fMRI. In *Proc. of 15th European Symposium on Artificial Neural Networks (ESANN'07)*, Bruges, Belgium, pp. 561–566, 2007.

[4] Ylipaavalniemi, J., and R. Vigário. Subspaces of Spatially Varying Independent Components in fMRI. In *Proc. 7th Int. Conf. on Independent Component Analysis and Blind Signal Separation (ICA'2007)*, London, England, pp. 665–672, 2007.

[5] Schleimer, J.-H., and R. Vigário. Reference-based extraction of phase synchronous components. In *Proc. 16th Int. Conf. on Artificial Neural Networks (ICANN'2006)*, Athens, Greece, pp. 230–238, 2006.

[6] Borisov, S., A. Ilin, R. Vigário, and E. Oja. Comparison of BSS methods for the detection of α-activity components in EEG. In *Proc. 6th Int. Conf. on Independent Component Analysis and Blind Signal Separation (ICA'2006)*, Charleston, South Carolina, USA, pp. 430–437, 2006.

[7] Pesonen, M., M. Laine, R. Vigário, and C. Krause. Brain oscillatory EEG responses reflect auditory memory functions. *abstr. 13th World Congress of Psychophysiology, International Organization of Psychophysiology (IOP'2006)*, Istanbul, Turkey, 2006.

[8] Pietilä, A., M. El-Segaier, R. Vigário, and E. Pesonen. Blind Source Separation of Cardiac Murmurs from Heart Recordings. In *Proc. 6th Int. Conf. on Independent Component Analysis and Blind Signal Separation (ICA'2006)*, Charleston, South Carolina, USA, pp. 470–477, 2006.

[9] Ylipaavalniemi, J., S. Mattila, A. Tarkiainen, and R. Vigário. Brains and Phantoms: An ICA Study of fMRI. In *Proc. 6th Int. Conf. on Independent Component Analysis and Blind Signal Separation (ICA'2006)*, Charleston, South Carolina, USA, pp. 503–510, 2006.

6.2 Reliable ICA and subspaces

In contrast to traditional hypothesis-driven methods, independent component analysis (ICA) is commonly used in functional magnetic resonance imaging (fMRI) studies to identify, in a blind manner, spatially independent elements of functional brain activity. Particularly, in studies using multi-modal stimuli or natural environments, where the brain responses are poorly predictable, and their individual elements may not be directly relatable to the given stimuli.

In earlier reported work, we have analysed the consistency of ICA estimates, by focusing on the spatial variability of the components. The optimization landscape of ICA is defined by structure of the data, noise, as well as the objective function used. The landscape can form elongated or branched valleys, containing many strong points, instead of singular local optima. Multiple runs of the ICA algorithm with varying random initial conditions and re-sampling allows to characterise the optimisation landscape and the robustness of the estimates.

Previous studies have analyzed the consistency of independent components, and suggested that some components can have a characteristic variability. The goal was to provide additional insight into the components, that is not possible to attain with single run approaches. Complex valleys can also be considered as separate subspaces, where statistical independence is not necessarily the best objective for decomposition.

We have now proposed a novel method for reliably identifying subspaces of functionally related independent components. We also proposed two approaches to further refine the decomposition into functionally meaningful components. One refinement method uses clustering, to distinguish the internal structure of the subspace. Another method is based on finding the coordinate system inside the subspace that maximally correlates with the temporal dynamics of the stimulation. The directions are found with canonical correlation analysis (CCA).

A study of subspaces was conducted on multi-modal fMRI recordings, including several forms of auditory stimulation. In the following figure, we can see a set of components, strongly related to auditory stimulation. Each component is consistent, appearing in all or most of the 100 runs. The mixing variability is also minimal. However, the spatial variance reveals a coincident location of variability, shared by all components. The variability links the components into a three dimensional subspace, even though ICA has consistently identified directions within the subspace.

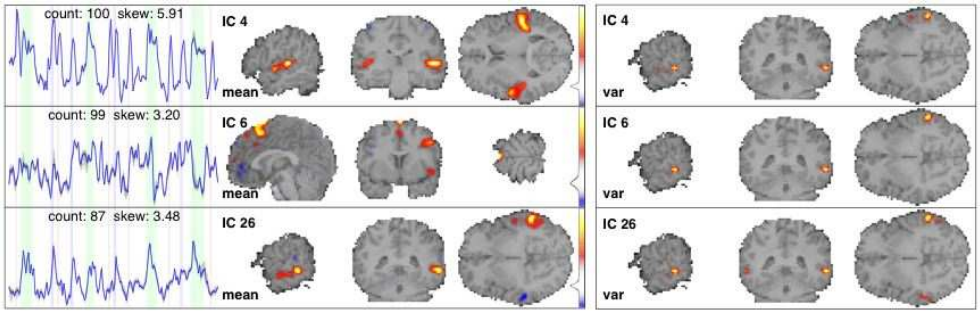


Figure 6.1: Tested approaches for alpha extraction from simulated data.

Within the study, we postulate that, based on spatial variance, components can be roughly divided into 3 classes: individual and consistent components, with distributed

variance due to noise; consistent members of a subspace, with focal variance coincident with the variance of the other members; and inconsistent subspaces, with variances coincident with their own mean. Such subspaces can provide information on networks of related activity in a purely data-driven manner. Criteria to disambiguate each subspace will then be of crucial relevance.

6.3 Towards brain correlates of natural stimuli

Natural stimuli are increasingly used in functional magnetic resonance imaging (fMRI) studies to imitate real-life situations. Consequently, challenges are created for novel analysis methods, including new machine learning tools. With natural stimuli it is no longer feasible to assume single features of the experimental design alone to account for the brain activity. Instead, relevant combinations of rich-enough stimulus features could explain the more complex activation patterns.

We proposed a novel two-step approach, where independent component analysis is first used to identify spatially independent brain processes, which we refer to as functional patterns. As the second step, temporal dependencies between stimuli and functional patterns are detected using dependency exploration methods. Our proposed framework looks for combinations of stimulus features and the corresponding combinations of functional patterns.

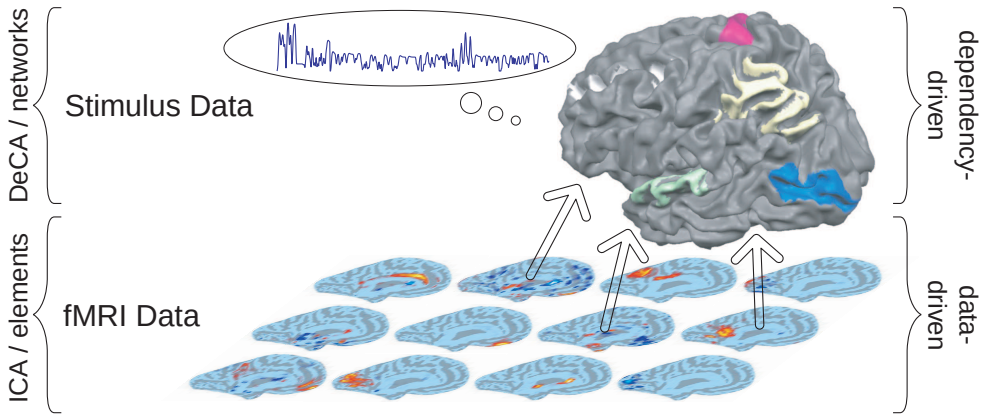
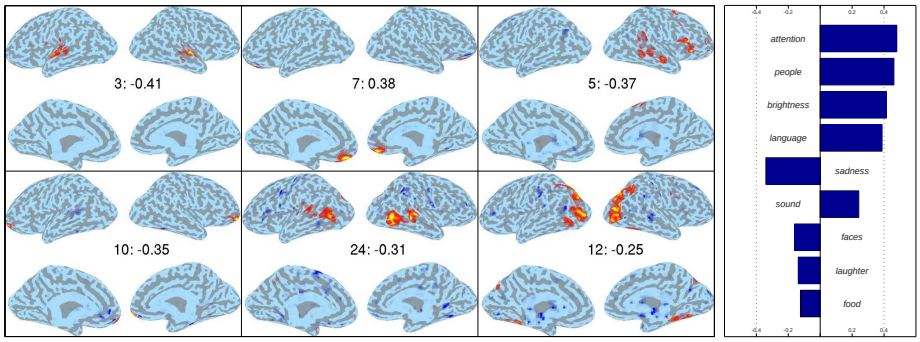


Figure 6.2: The proposed framework: elements of functional brain activity emerge from the data via ICA. Functional networks are revealed by DeCA, based on covariation between the elements and task goals, encoded as features.

This two-step approach was tested on fMRI recordings of brain responses to natural stimuli, consisting of a movie with 20 minutes duration. Rather subjective features were extracted from the movie, including labels such as *"attention"*, *"sadness"*, *"people"* or *"laughter"*.

As an illustrative example, we can look into a network comprising brain areas that individually correspond to, *e.g.*, auditory (IC3), visual (IC12), and multi-modal integration (IC24). This suggests that the functional role of the whole network is related to combining information from many sensory inputs. Indeed, the four highest scoring features of the dependent component are *attention*, *people*, *brightness* and *language*.

The found networks seem plausible, considering the limited and very subjective nature of the available stimulus features. Some elements were a part of several networks, with different functional contribution to each networks common task. A more controlled study has been carried out since, to verify the results and to further develop the approach. These will be reported in the next biennial report.



(a) (b)

Figure 6.3: (a) The 6 ICs corresponding to the highest loadings in the first dependent component. The loading value is shown in the middle of each square. (b) Respective loadings of the stimulus features.

6.4 Synchrony exploration

Interest in phase synchronisation phenomena has a long history, when studying the interaction of complex, natural or artificial, dynamic systems. Although not completely adopted, synchronisation was attributed a role in the interplay between different parts of the central nervous system as well as across central and peripheral nervous systems. Such phenomena can be quantified by the phase locking factor (PLF), which requires knowledge of the instantaneous phase of an observed signal.

Linear sources separation methods treat scenarios in which measurements do not represent direct observations of the dynamics, but rather superpositions of underlying latent processes. Such a mixing process can cause spuriously high PLF's between the measurements, and camouflage the phase locking to a provided reference signal. Essentially, synchronisation is either caused by a common input or by interactions between neurons.

Reference-based approach

The PLF between a linear projection of the data and a reference can be maximised as an optimisation criterion, revealing the most synchronous source component present in the data, with its corresponding amplitude. This is possible despite the amplitude distributions being Gaussian, or the signals being statistically dependent, common assumptions in blind sources separation techniques without a-priori knowledge, e.g. in form of a reference signal.

We first addressed this reference-based problem, and proposed a new algorithm capable of extracting sources phase-locked with a reference. In the following illustration one can see the efficiency of such a method. The sources, depicted on the right frame, were chosen so that neither high-order statistics methods, e.g., FastICA, nor methods based on temporal decorrelation, e.g., SOBI would perform the desired source estimation.

We tested this approach on MEG recordings, with a 306-sensor Vectorview neuromagnetometer, together with left and right hand EMG's. The subject was instructed to simultaneously keep isometric contraction in left and right hand muscles, using a special squeezing device. We then used the right hand EMG as a reference for the phase exploration into the MEG recordings. The results achieved agreed with early studies performed in the same recordings.

We also addressed the "internal neuronal synchronisation" problem, where no clear reference is available, proposing to cluster a population of oscillators into segregated sub-

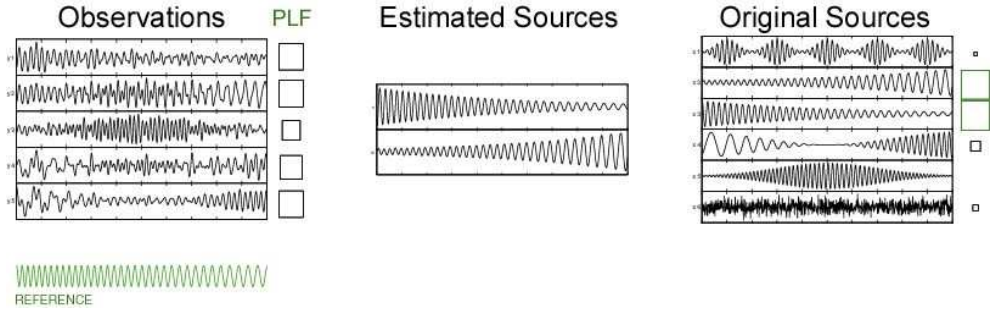
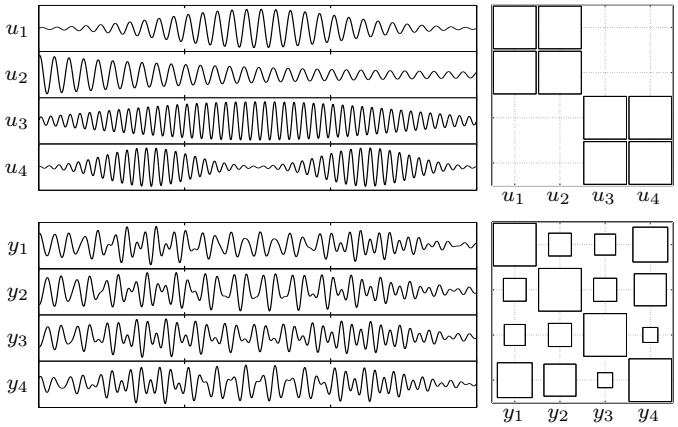


Figure 6.4: Six signals, of which only two are locked in phase.



(a) Oscillators with distinct amplitudes, grouped in synchronous subspaces. PLFs are depicted as a Hinton diagram. ($P = 0.08$)

(b) Mixtures with an overall spuriously phase locking. ($P = 3.66$)

populations, exhibiting high internal interactions. Approaches to solve this problem have often assumed different frequencies for the various sub-populations, usually neglecting phase information. These assumptions pose a restriction to the analysis of the dynamic world of natural systems, where communication can be unrelated to the natural frequency of the constituent oscillators.

Our solution makes explicit use of phase information, extracted from known models of physical interactions. The approach relies on a post-rotation of the eigenvectors of the synchronisation matrix.

With simulations, we show the effect of the post-rotation, in the estimation of underlying sources for which their frequency has been drawn from a global distribution. In neurobiological terms, this means that the neuron's system parameters, which determine its natural frequency, do not depend on the synaptic connections it has formed. In such formulation, frequency can not be used to identify the sources anymore. Phase clustering is then crucial for the task.

We have also proposed a method to reveal phase-locked subspaces, based on a concept of order in complex systems of nonlinear oscillators. Any order parameter quantifying the degree of organisation in a physical system can be studied in connection to source extraction algorithms. Independent component analysis, by minimising the mutual information of the sources, falls into that line of thought, since it can be interpreted as searching components with low complexity. Complexity pursuit, a modification minimising Kolmogorov complexity, is a further example.

Using such concept of order, we designed an algorithm capable of revealing subspaces of oscillators such that: oscillators of the same subspace are completely phase locked; whereas between subspaces there is no Phase locking. The following illustrations exemplify the algorithm's performance. Estimated sources coincide with the true ones.

6.5 Overview of other topics

Within the duration of this biennial report, several research topics have been addressed with a more prospective view. Some will be the subject of more thorough development in further reports, whereas other will stay as simple case studies. This section reviews four of those.

Extraction of alpha activity

Following earlier work on the characterisation of low- and high-amplitude alpha brain activity, we tested a two-step blind source separation approach for the extraction of such rhythms from ongoing EEG. The method comprised a denoising stage, performed by DSS, followed by either a high-order statistical independent component analysis source estimation, FastICA, or a temporal decorrelation one, TDSEP.

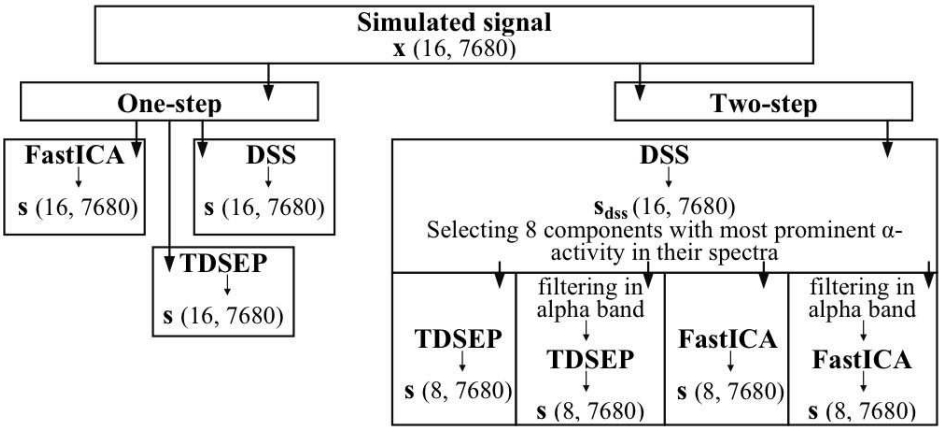


Figure 6.5: Tested approaches for alpha extraction from simulated data.

The main findings are that denoising has, as expected, a positive effect in rendering the subsequent source separation algorithms more efficient. In addition, we observed that high-order statistics ICA was more adequate in such separation than TDSEP, in spite of the latter being particularly suited for dealing with temporally structured sources. A targeting α-filter, placed between the denoising and the TDSEP modules, resulted in good estimates, rendering the combination rather efficient. Such filtering seems to not affect significantly FastICA.

Artefact removal in ERD/ERS study

Still within the rhythmic activity of the brain, we participated in a study of brain oscillatory EEG responses to auditory memory functions. The analysis concentrated on event related de-synchronisation and synchronisation (ERD and ERS, respectively), in the theta and alpha frequency ranges for ERS and also in beta for ERD.

The outcomes of that study suggested that theta frequency ERS responses may be associated with working memory functions, whereas alpha ERD/ERS responses robustly dissociate between auditory memory encoding and recognition.

ICA showed to be crucial in denoising the raw ongoing EEG, prior to wavelet processing. Several subjects displayed considerable artefacts that rendered most of the event-related responses virtually unusable.

BSS of cardiac murmurs

A significant percentage of young children present cardiac murmurs. However, only one percent of them are caused by a congenital heart defect; others are physiological. An automated system for an initial recording and analysis of the cardiac sounds could enable the primary care physicians to make the initial diagnosis and thus decrease the workload of the specialised health care system. independent component analysis source estimation, FastICA, or a temporal decorrelation one, TDSEP.

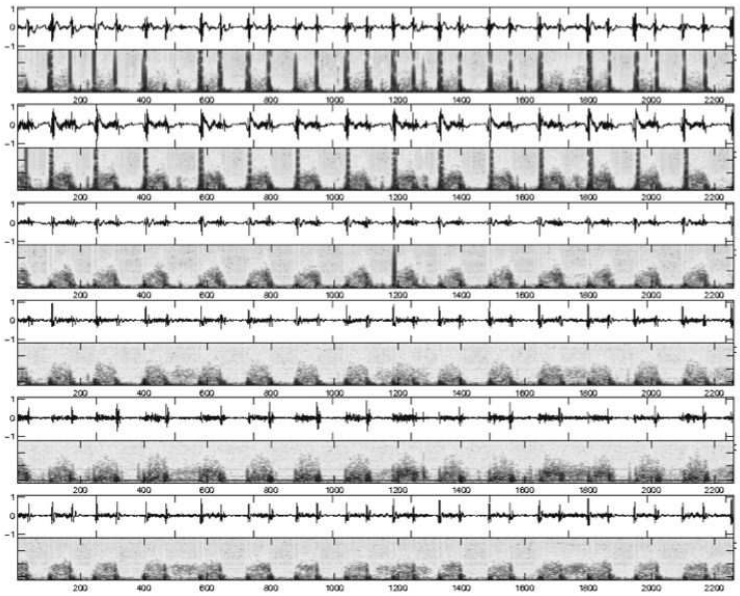


Figure 6.6: Six-channel PCG recordings from a patient, together with their spectrograms. The S1 and S2 are clearly visible in the first spectrogram as periodic pairs of vertical bars covering all the frequencies. Murmurs are visible in the systole, between the S1 and S2, present on all six recordings.

The first step to such analysis is the identification of the different components of the cardiac cycle, with particular emphasis to the separation of the murmurs. We have proposed a new methodological framework to address this issue, combining ICA and DSS. independent component analysis source estimation, FastICA, or a temporal decorrelation one, TDSEP.

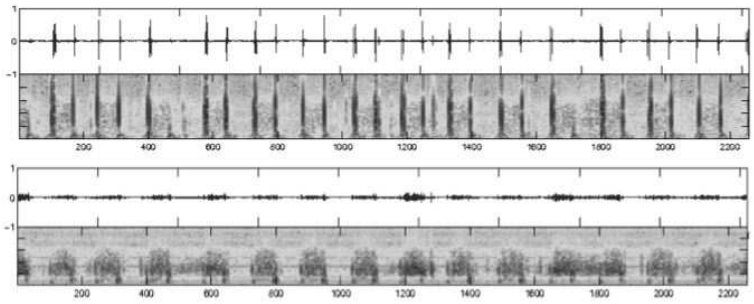


Figure 6.7: Heart sounds S1 and S2, clearly isolated from all other signals. In the second frame are uncontaminated murmurs.

Using such approach, we have been able to isolate rather efficiently the murmurs, as well as heart sounds S1 and S2 and artefacts such as voices recorded during the measurements.

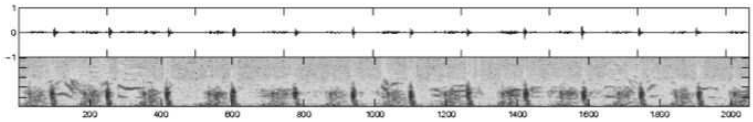


Figure 6.8: Speech artefacts present in PCG recordings. Formant structures are clearly visible.

With the aforementioned results, the collaboration with the Lund University Hospital, Sweden, has been strengthened, and further research outcomes are expected in the next reports.

Phantom study in fMRI

Phantom measurements are routinely used for verifying and calibrating the quality of MRI machinery. However, data-driven analysis of phantom fMRI data has been largely overlooked, possibly due to the lack of a method for assessing the reliability of the solutions. We have now used a reliable ICA approach to such analysis, and revealed evidence for possible misinterpretations in ICA studies with real subjects.

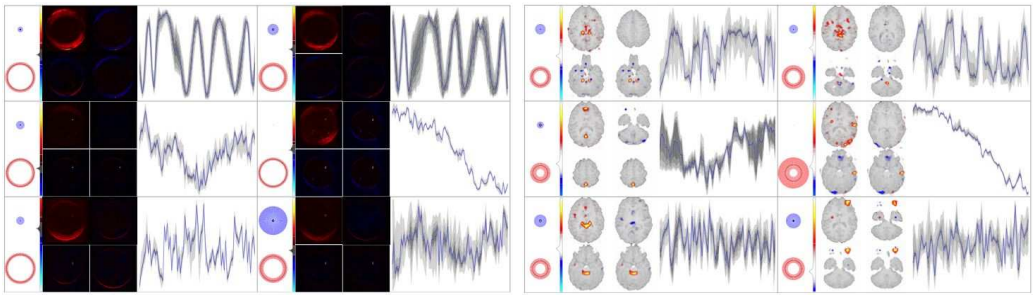


Figure 6.9: Reliable independent components, extracted from fMRI of a spherical phantom (a), and a real subject (b). Corresponding temporal 'activation' patterns are shown on the right of each estimate.

Several independent components found on a real subject presented a temporal structure that follows clearly that of phantoms. We speculate that methods other than ICA can also suffer from a similar kind of misinterpretation. We therefore suggest the need for a better understanding of the artificial, scanner- or environmentally-induced artefacts, prior to the automatic analysis of any fMRI recording. A comparison between real brain ICA and phantom-based decompositions may help in the validation of the estimated components.