A CLINICAL METABOLOMICS STRATEGY TO DISCOVER NEW BIOMARKERS IN COMPLEX DISEASE: AN OVERVIEW

Baumgartner C
Institute of Electrical and Biomedical Engineering, Research Group for Clinical Bioinformatics, UMIT, Hall in Tirol, Austria
christian.baumgartner@umit.at

Abstract: In clinical metabolomics biomarker discovery is conducted by the complementary power of clinical study design and execution, molecular profiling technologies and an efficient bioinformatics strategy for biomarker search, verification and interpretation. This survey article gives a review of useful bioinformatics methods for biomarker discovery, addressing the problem of data preprocessing, the data-driven search, prioritization and biological interpretation of metabolic biomarkers candidates in disease. Advanced data mining approaches and new strategies using network-based methods are discussed in more detail.

Keywords: Biomarker discovery, metabolomics, bioinformatics, complex disease

Introduction

Rapid progress in high-throughput technologies such as next generation sequencing or MS-based profiling techniques like GC or LC-MS/MS and in the development of related bioinformatics methods allow the systemic analysis and characterization of alterations in genes, proteins and metabolites. These technologies offer a broad spectrum of approaches to discover novel biomarkers and pathways activated in complex diseases. Since clinically relevant biomarkers are still lacking for aiding in diagnosis, disease screening or treatment, the complementary power of modern profiling techniques and powerful bioinformatics tools is applied for the discovery of new biomarker candidates. This large interest in biomarker discovery originates from their broad range of clinical applications and their fundamental impact on pharmaceutical industry [1].

In this contribution, emerging bioinformatics methods for biomarker discovery in metabolomics, i.e. the systematic search of low molecular weight biochemical compounds in complex biological mixtures, their selection and application to the problem of identifying, prioritizing, interpreting and validating metabolic biomarkers suitable for the clinical application are discussed.

Methods

In general, human biomarker discovery studies comprise a variety of experimental designs, including most frequently used retrospective case-control or more complex cohort study designs such as crossover or serial sampling designs. Latter, so-called longitudinal cohort studies allow patients to serve as their own biological control, which enable to study thoroughly the kinetics of circulating analytes by reducing the interindividual variability observed in multiple cohort studies. In metabolomic biomarker discovery bioinformatics plays a major role because this process is highly data-driven and, thus, constitutes the missing link between the initial discovery phases including experimental design, study execution and bioanalytics (i.e. sample preparation, separation and high-throughput profiling) and the search, verification and independent validation of biomarker candidates (see Figure 1).

Figure 1: Biomarker discovery process in human disease.

Once sample collection, preparation, separation and MS analysis have been carried out, technical reviewing of generated data is essential to insure a high degree of completeness, consistency and reproducibility. Data preprocessing is an additional necessary step to transform data into a format suitable for subsequent targeted analyses. This includes tasks such as data transformation and normalization, data sampling and outlier detection [2].

A pool of statistical bioinformatics methods is nowadays available for identifying, prioritizing and classifying robust and generalizable biomarker candidates, showing a high predictive value in terms of sensitivity and specificity. In general, data analysis tasks for the search of biomarker candidates in experimental data are “supervised” because study cohorts are typically well-phenotyped in carefully designed and controlled clinical trials. Commonly used supervised data mining methods for the search and prioritization of biomarker candidates in both, independent and dependent samples, include paired/unpaired null hypothesis testing, principal component analysis (PCA), uni- and multivariate feature selection methods
such as the information gain, reliefF, associative voting, the biomarker identifier, guilt-by-association feature selection, repeated measure analysis as well as more advanced methods like embedded or ensemble-based techniques (e.g., support vector machine recursive feature elimination, stacked feature ranking or the wrapper approach) [2]. Very recently, we proposed a prioritization model for classifying metabolic biomarker candidates according to their discriminatory ability by coupling a feature selection modality with a network-based approach to review and interpret major hubs (key metabolites) and their interactions and correlations in the network. In particular, the quantitative analysis of networks has become a novel technique for the biological and biochemical interpretation of alterations in disease-associated pathways. Therein, different types of topological graph descriptors, e.g., parametric or partition-based entropy measures are suggested to be used for analyzing such complex metabolic networks [3].

Generalizability and validation of biomarker candidates is a crucial step of the entire biomarker discovery process. Objective measures to assess the discriminatory or predictive value and the generalizable power of identified biomarker candidates are sensitivity and specificity or the area under the receiver operating curve (AUC). In longitudinal time series studies, alternative measures can be used to assess the predictive value of biomarkers in a similar manner, as described in [4]. In general, identified biomarker candidates need to be validated using larger sample sets, covering a more comprehensive cross-section of patients or populations. If such data is unavailable or impossible to collect, computational cross-validation strategies can be performed to assess generalizability on a single cohort. Usually, stratified n-fold cross-validation, bootstrapping or permutation analysis can be applied to overcome this problem. Nevertheless, prospective clinical studies are finally needed to verify and validate the clinical benefit of the selected panel of biomarker candidates before they can be applied to clinical applications.

A challenging discovery step is the biological and biochemical interpretation of putative metabolic biomarker candidates. In metabolomics, pathway mining tools are usually used to map, visualize and reconstruct a list of possible pathways by extracting metabolic information from network databases like KEGG. Such tools allow a direct functional annotation of metabolites, enzymes or reactions related to experimental findings. Hyperlinks to comprehensive databases such as OMIM or Swiss-Prot bring forth supplementary information about the underlying biochemical and biological mechanisms.

Applications

In a recent longitudinal biomarker cohort study we were able to identify, categorize, and profile kinetic patterns of early metabolic biomarkers of myocardial infarction using a bioinformatic-driven discovery strategy as described above. A panel of new metabolic signatures could be identified that appears as early as 10 minutes after the event. Some of them are promising candidates that may be useful in developing future diagnostic tests [4]. A new computational approach by coupling a generic search and prioritization strategy with a network-based approach was performed for profiling the human response to physical exercise. In this study a group of known, but also unexpected metabolic signatures were identified by studying the analyte kinetics at rest versus stress [3]. Figure 2 demonstrates figuratively this two step discovery strategy.

![Figure 2: Coupled network-based computational strategy for biomarker search, prioritization and interpretation in time series data. The pBI feature selection algorithm preselects major metabolites, which are subsequently used for inferring a kinetic network for studying metabolite interactions.](Image)

Conclusion

Major interest in metabolomics biomarker discovery studies originates from their broad range of medical applications - as clinically validated biomarkers can aid in diagnosis and disease prediction or serve as indicators of treatment efficiency, their impact on pharmaceutical industry and the public health system. Therein, bioinformatics is an essential tool for the biomarker search, bridging the gap between complex raw data generated by MS analysis and the data-driven search and verification of new biomarkers and pathways associated with disease.

Bibliography

SYSTEMS BIOLOGY ANALYSIS OF KINASE INHIBITOR ACTION IN LEUKEMIA TREATMENTS

Jacques Colinge
CeMM – Research Center for molecular Medicine of the Austrian Academy of Sciences, Vienna, Austria
jcolinge@cemm.oeaw.ac.at

Abstract: We propose a new computational method of relating kinase inhibitor protein targets to genetic defects associated to a disease through protein interaction network-based random walks. We illustrate the potential of this methods on two forms of leukemia chronic myeloid and Philadelphia positive acute lymphoblastic.

Keywords: Personalized medicine, systems biology, kinase inhibitor

Introduction
To be able to understand the mechanisms of action of drugs, predict their efficacy, and anticipate their potential side-effects is important during drug development. In diseases where the genetic background of patients modulates treatment response, it allows personalizing the therapy.

Methods
We have developed computational methods to analyze unbiased drug target profiles, measured by chemical proteomics, which we applied to kinase inhibitors. Chemical proteomics pulldowns are affinity purification methods that measure compound-protein interactions and they have revealed that kinase inhibitors can be promiscuous compounds [1,2], hence motivating a system-wide analysis approach. We were able to correlate chemical pulldown data with disease models constructed on the basis of known causative genes by means of diffusion processes, e.g. random walks (Fig. 1). In the case of chronic myeloid leukemia (CML), this allowed us to predict very likely immunosuppressive effects of bosutinib and new indications for CML drugs.

Recently, in still unpublished work, we measured the protein target profiles of four kinase inhibitors in two Philadelphia chromosome-positive acute lymphoblastic leukemia (Ph+ ALL) cell lines. We obtained predictions of treatment efficacy in reasonable agreement with experimentally determined IC_{50} values and therefore we able to recommend the most successful therapy from in silico models.

Discussion
Altogether, our work aims at developing techniques able to implement personalized medicine therapies by combining detailed knowledge on drug profiles and patient specific genetic background.

Figure 1: The human interactome – all known protein-protein interactions – constitutes a network which provides a convenient computational model of “cell biology”. Protein targets determined by chemical proteomics significantly influence part of this network. The same is true for gene mutations associated to a specific disease or patient. The computation of a correlation scores allows selecting the most likely efficient compound for a disease or patient or to screen diseases for repurposing compounds.

Bibliography


IDENTIFYING BIOLOGICALLY MEANINGFUL GROUPS OF GENES IN THE ANDROGEN RECEPTOR NETWORK

Marina Popovscaia¹, Christian Baumgartner¹, Michael Handler¹ and Helmut Klocker²

¹ Institute of Electrical and Biomedical Engineering, UMIT- The Health and Life Sciences University, Hall in Tirol, Austria
² Department of Urology, Division of Experimental Urology, Innsbruck Medical University, Innsbruck, Austria

marina.popovscaia@umit.at

Abstract: Androgen receptor is a transcription factor that plays a crucial role in the development of prostate cancer. Androgen receptor cofactors and androgen response genes are tightly connected to the androgen receptor transcriptional activity. In our study we built and analyzed a network which consist only of these groups of genes. In the constructed network we could identify a number of gene modules which potentially can be used as new biomarkers candidates for prostate cancer.

Keywords: prostate cancer, androgen receptor network, androgen receptor cofactors, androgen response genes.

Introduction

Prostate cancer (PCa) is one of the most frequent oncological diseases in men in Western industrialized countries. Up to date PCa is curable when detected at the early organ confined stages. The identification of PCa in early stages is therefore one of the top priorities of modern cancer research.

A fundamental role in the development of PCa, as well as in the growth, differentiation and maintenance of healthy prostate, play androgens - steroid hormones responsible for male reproductive function and behavior [1]. Effects of the androgens are meditated by the specific transcription factor - androgen receptor (AR). In its inactive form AR is located in the cytoplasm of the cell in complex with heat-shock proteins. Upon ligand (androgen) binding AR homodimerizes and translocates in the nucleus where it binds to the specific sequences on the DNA known as androgen response elements (AREs) [1]. It may also recruit cofactors – proteins that either enhance (i.e. coactivators) or reduce (i.e. corepressors) AR transactivation (see Fig. 1).

The aim of our work was to explore in detail the function of the complex "AR, AR cofactors and ARE containing genes" and to investigate the possibility of using one gene or groups of genes as biomarker(s) for PCa detection. For this purpose we constructed a network, further addressed as "AR-network", based specifically on these three groups of genes. Using various techniques we detected a number of modules in the AR-network, ranked them and investigated their ability to differentiate between cancer and benign tissue samples.

Methods

Construction of AR-network. For our study we initially chose 1913 genes: AR, 162 genes known to be AR cofactors and 1752 genes reported to be androgen controlled (containing AREs). The AR-network was inferred from microarray datasets GDS3111 and GDS2782 for LNCaP cell lines treated with dihydrotestosterone (DHT), downloaded from GEO (Gene Expression Omnibus) database. Both datasets had been generated using the platform GPL570 (Affymetrix Human Genome U133 Plus 2.0 Array). After normalization and data preprocessing a total number of 1692 genes were chosen for the AR-network construction. The C3NET method [3] which is based on mutual information (1) was used for inferring the network

\[
I(X, Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \frac{p(x, y)}{p(x) p(y)}. \tag{1}
\]

Identification of gene modules. Groups of genes were identified using hierarchical clustering.

Validation of results. In order to find out which of the gene groups might be biologically meaningful, we, first, ranked them based on the enrichment analysis of gene ontology terms. Second, we investigated the ability of highly ranked gene clusters to function as biomarker candidates and determine their discriminatory ability when comparing PCa versus non-PCa cases. For this step we used a SVM (support vector machine) classifier. Validation was performed on the GDS1439 and GSE32982 datasets containing expression data of benign, clinically localized and metastatic
Results

Upon construction of the AR-network using mutual information we were able to identify more than 60 gene modules of various dimensions. After ranking, we have chosen 18 modules for further investigation. One third of them could discriminate between PCa and non-PCa cases with various accuracy levels (Tab. 1).

As an example, we present here module No.6 (Fig. 2) which showed the highest accuracy (74.2%) of predicting PCa and non-PCa cases in the validation sets. AR cofactors are marked in green and ARE genes – in blue. One can immediately see that in this AR-network module AR cofactors tend to be the hubs centers.

Table 1: Validation of the gene modules.

<table>
<thead>
<tr>
<th>Modules</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>68.7%</td>
<td>72.6%</td>
<td>61.5%</td>
<td>66.1%</td>
<td>63.7%</td>
<td>74.2%</td>
</tr>
</tbody>
</table>

Enrichment analysis was performed to significantly (p-value < 0.05) associate Gene Ontology (GO) annotations to the genes in module No.6 using Genecodis web-based tool [4, 5]. The following GO terms were enriched in our group of genes: protein binding (Molecular function (MF)), pathways in cancer (KEGG pathways (KEGG)), focal adhesion (KEGG), small cell lung cancer (KEGG), p53 signaling pathway (KEGG), melanoma (KEGG), cell differentiation (biological (BP)), axon guidance (BP), intracellular protein transport (BP), cell migration (BP), protein export from nucleus (BP) and others (Fig. 3).

Discussion

Using our proposed discovery strategy we were able to construct and analyze the AR-network with respect to AR cofactors and ARE containing genes. Our aim was to identify single genes or biologically meaningful groups of genes which might be used as new biomarker candidates for PCa prediction. We could define a number of gene modules which showed significant accuracy of distinguishing between PCa and non-PCa cases in the validation sets. These promising results together with GO associations of the genes will help us to construct hypotheses for further validation of the results in the wet-laboratory.

Acknowledgement

This work was supported by the Oncotyrol Center for Personalized Cancer Medicine within the scope of the Austrian COMET program (project 1.3.1).

Bibliography


AUTOMATED DETECTION OF HUMAN BRAIN TUMORS USING IN VIVO $^1$H MR SPECTRA OF HEALTHY BRAIN TISSUE

Christian Hoffmann¹, Andreas Wille¹, Martin Busch² and Susanne Winter¹

¹Institut für Neuroinformatik, Ruhr-Universität Bochum, Germany
²Grönewege Institut für Mikrotherapie, Bochum, Germany

christian.hoffmann@ini.rub.de

Abstract: Presence of a brain tumor affects the metabolite concentrations of healthy brain tissue. This study inspects the qualification of MR spectroscopy recordings of this tissue for automatic tumor patient classification with linear discriminant analysis, artificial neural networks and support vector machines. With spectroscopy datasets reduced down to the concentrations of two different metabolites, a classification performance of approximately 80% could still be achieved.

Keywords: MR-spectroscopy, brain tumors, classification, multilayer perceptrons, support vector machines

Introduction

While the current criterion standard for brain tumor detection is biopsy with subsequent histological assessment, in vivo $^1$H magnetic resonance (MR) spectroscopy has been recognized as a possible non-invasive alternative over the past decades. This radiological method allows localized recordings of relative or absolute metabolite concentrations within the brain, which in turn differ significantly for healthy and pathological brain tissue [1]. Furthermore, it has recently been shown that tumors influence the spectrum of “healthy” tissue localized in the contralateral hemisphere as well [2]. In this study, the feasibility of several machine learning techniques for tumor patient classification based upon such spectra is demonstrated in detail. Especially the metabolite N-acetylaspartate can serve as a marker for neural integrity as its concentration depletes with most brain lesions. Busch et al. already emphasized its pivotal role for the oncological analysis of MR spectroscopy data [2]. This study additionally addresses the question whether other metabolite concentrations assessable by MR spectroscopy are also eligible for the classification task.

Materials

In total 264 spectra of 96 different persons were considered. 135 spectra were recorded from 52 patients with the region of interest placed above the ventricle in the hemisphere contralateral to the tumor. The remaining data originated from a healthy control group. Although the patients’ data was labeled according to the outcome of histological examinations (Tab. 1) and the tumor types are relevant from a medical point of view, they were not distinguished in this study due to low patient numbers.

Table 1: Number of spectra and patients with regard to the biopsy-based tumor classification.

<table>
<thead>
<tr>
<th>Tumor classification</th>
<th>Spectra</th>
<th>Patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Astrocytoma II</td>
<td>17</td>
<td>3</td>
</tr>
<tr>
<td>Astrocytoma III</td>
<td>20</td>
<td>6</td>
</tr>
<tr>
<td>Glioblastoma</td>
<td>88</td>
<td>28</td>
</tr>
<tr>
<td>Metastasis</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>Without histology</td>
<td>4</td>
<td>3</td>
</tr>
</tbody>
</table>

The data comprised relative concentrations of the metabolites N-acetylaspartate (Naa), choline containing compounds (Cho), creatine (Cre), myoinositol (Myo) and the sum of glutamate and glutamine (Glx). Each test dataset was build using 20% of the overall data, the corresponding training sets contained 60 - 80% depending on validation requirements of the algorithms. All test errors were averaged over 100 experiments to avoid bias which might be introduced by a specific choice of datasets.

Classification algorithms

Three different supervised learning methods were applied for the automatic distinction between healthy and diseased persons. The regularized linear discriminant analysis (LDA) [3] is a multivariate statistical method based on covariance estimates under the assumption of classwise normal distributed features. It strives to maximize the inter-class and minimize the intra-class variances by conducting a linear separation in the input space. Artificial neural networks (ANNs) consist of units – the so-called neurons – which are arranged in layers. The connections between neurons pertaining to subsequent layers are governed by adaptive weights. During the learning phase, these weights are adjusted by error backpropagation (IRprop+ [4]) employing a cross-entropy function. In this study, feed forward networks with one hidden layer consisting of at most seven neurons were used. The activation for these neurons was modeled by a Fermi function. Support vector machines (SVMs) [5] are maximum-margin classifiers. They implicitly perform a kernel-induced, not necessarily linear mapping of the input data into a so-called feature space in order to linearly separate the data afterwards. For the benefit of the overall classification performance, the soft-margin SVMs utilized in this study allow certain misclassifications under penalty. Gaussian kernels with diagonal covariance matrices were used as well as...
Table 2: Percentaged mean classification error for different metabolite combinations. Exclusively the best results are given for each algorithm. (N) indicates a normalization of all underlying concentrations with respect to Cre. For SVMs, the applied kernel is denoted by (G) for Gaussian, (L) for linear and (P) for polynomial.

<table>
<thead>
<tr>
<th>Metabolites</th>
<th>LDA (%)</th>
<th>ANN (%)</th>
<th>SVM (%)</th>
<th>LDA (N) (%)</th>
<th>ANN (N) (%)</th>
<th>SVM (N) (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naa, Cho, Cre, Myo, Glx</td>
<td>17.4±5.0</td>
<td>15.3±4.0</td>
<td>12.0±7.2(P)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Naa, Cho, Myo, Glx</td>
<td>18.0±5.6</td>
<td>15.7±4.4</td>
<td>14.2±6.0(P)</td>
<td>16.6±4.9</td>
<td>16.6±4.7</td>
<td>14.5±4.3(P)</td>
</tr>
<tr>
<td>Naa</td>
<td>37.5±6.0</td>
<td>40.2±5.5</td>
<td>43.0±6.5(L)</td>
<td>23.1±5.9</td>
<td>23.7±4.3</td>
<td>21.8±3.0(G)</td>
</tr>
<tr>
<td>Naa, Cre</td>
<td>23.7±5.4</td>
<td>24.2±3.9</td>
<td>25.7±4.5(P)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Naa (N), Cho, Cre, Myo, Glx</td>
<td>16.5±4.8</td>
<td>14.2±4.8</td>
<td>15.4±5.3(P)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cre</td>
<td>31.9±6.2</td>
<td>31.0±4.6</td>
<td>32.7±4.3(P)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cho, Cre, Myo, Glx</td>
<td>28.9±5.7</td>
<td>22.1±5.4</td>
<td>16.6±7.8(P)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cho, Cre, Myo</td>
<td>29.1±6.1</td>
<td>23.5±4.8</td>
<td>20.0±7.2(P)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cho, Cre, Glx</td>
<td>29.6±5.3</td>
<td>21.4±5.8</td>
<td>16.0±7.1(P)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cre, Myo, Glx</td>
<td>27.1±5.1</td>
<td>22.9±4.7</td>
<td>20.8±6.0(P)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cho, Myo, Glx</td>
<td>30.6±5.2</td>
<td>24.0±5.1</td>
<td>20.2±6.4(P)</td>
<td>44.0±6.5</td>
<td>38.4±7.3</td>
<td>36.4±4.9(P)</td>
</tr>
<tr>
<td>Cho, Myo</td>
<td>32.5±5.8</td>
<td>34.6±4.2</td>
<td>27.0±6.2(P)</td>
<td>42.2±5.9</td>
<td>45.8±4.8</td>
<td>38.8±4.3(L)</td>
</tr>
<tr>
<td>Cho, Glx</td>
<td>31.2±5.4</td>
<td>24.5±5.4</td>
<td>21.9±5.5(P)</td>
<td>45.3±6.5</td>
<td>39.5±9.8</td>
<td>35.9±5.1(P)</td>
</tr>
<tr>
<td>Myo, Glx</td>
<td>30.3±5.9</td>
<td>25.2±5.4</td>
<td>18.1±7.5(P)</td>
<td>44.7±6.1</td>
<td>41.0±6.4</td>
<td>32.1±9.0(G)</td>
</tr>
</tbody>
</table>

Discussion

Coincident with Busch et al. [2], a strong decrease in classification performance of LDAs was observed if Naa concentrations were discarded. ANNs and SVMs were also affected by the absence of Naa values in a similar way. However, combinations of the remaining metabolite concentrations contained almost an equal amount of information as normalized Naa values, leading to classification results of up to approximately 80%. Remarkably, it was not decisive which metabolites were included in these combinations. This finding is particularly useful for cases where some relative concentrations are missing.

In contrast to our expectations, normalizing the concentrations with respect to Cre generally led to increased error rates. This indicates the possibility of relations between the unnormalized metabolites, which are found implicitly by the algorithms. Overall, the results show that SVMs employing linear and polynomial kernels in particular are well suited for the tumor patient classification task. This may be especially relevant for early cancer diagnosis since the required data can be collected from apparently healthy tissue and therefore no indication of tumor presence is necessary in advance.

Bibliography


GENOME ASSEMBLY FRAMEWORK ON MASSIVELY PARALLEL, DISTRIBUTED MEMORY SUPERCOMPUTERS

Friedrich Menhorn¹,², Matthias Reumann²,³
¹Department of Computer Science, TU München, Germany
²Department of Computing and Information Systems, University of Melbourne, Carlton, VIC, Australia
³IBM Research - Australia Laboratory, Carlton, VIC, Australia
menhorn@in.tum.de

Abstract: Genome Assembly describes the process of assembling a long Deoxyribonucleic acid sequence out of next generation sequencing (NGS) data. Computational resources can become a bottleneck or large scale routine use. We propose a genome assembly framework for massively parallel, distributed memory supercomputers. Our framework builds on the simple idea to equally distribute the number of reads to each processor. Each processor holds the whole reference genome. Each processor aligns the short reads independently and sends the reads back to root processor together with the corresponding position and the whole genome can be aligned. We run our alignment framework on up to 8,196 processors of the IBM Blue Gene/Q "Avoca" at the Victorian Life Science Computation Initiative. The results show that more than 6 Million reads of over 324 Million nucleotides can be assembled in under 20 minutes versus previously requiring hours. Thus, our framework allows fast assembly of NGS data.

Keywords: Computation Biology, Genome Assembly, Supercomputing

Introduction

Next generation sequencing (NGS) has made feasible the use of sequencing in healthcare. However, routine use of NGS in a clinical setting on day to day basis has not yet been achieved due to lack of automatized workflows as well as the need for experienced bioinformaticians to carry out the analysis. Further, computational resources can become the bottleneck when scaling up NGS to large sample sizes. In NGS, the genome is divided in smaller partitions of about 50 to 250 base pairs with one base pair representing a Deoxyribonucleic acid (DNA) molecule. These so called reads are aligned to a reference genome. The alignment process is computationally expensive. However, the large number of short reads (millions for bacteria) can be distributed to a parallel system. Thus, the objective of this study was to build a framework for massively parallel, distributed memory supercomputers to carry out to speed up genome assembly to a reference genome, which we anticipate to be the most common form of assembly in high throughput mode.

Methods

Our frameworks builds on the simple idea to equally distribute the number of reads to each processor. Each processor then requires to hold the whole reference genome for the problem to become pleasantly parallel. Then, each processor can use the chosen alignment algorithm independently to find the proper position for its specific reads. After finishing, it sends the reads back to root processor together with the corresponding position and the whole genome can be aligned.

As an example system for massively parallel, distributed memory supercomputers we use the IBM Blue Gene/Q, a supercomputer “Avoca” at the Victorian Life Science Computation Initiative with peak performance of 838.86 teraFLOPS [7]. Avoca holds 4,096 compute nodes and a total of 65,536 PowerPC based 1.6GHz cores each with four hardware threads. Each compute node holds 16GB of RAM (1GB per core). The assembly framework is build using the Message Passing Interface (MPI) for inter-process communication at present. While we currently have not implemented a hybrid model, shared memory parallelism can easily be introduced in our framework using e.g. OpenMP for a single process per compute node that could yield up to 262,144 threads for the full Avoca system.

We use a slightly modified master-worker communication where process 0 and 1 are the root processors, thereby controlling the other processors. At the beginning, the root processors read the raw NGS output file. In parallel, the other processors read the file which contains the sequence of the reference genome. By using MPI_Scatter, we are able to distribute an equal number of reads m to each processor. The number of reads x for processor i of a given number of processors n are calculated through

$$x_i = \left\lfloor \frac{m}{n-2} \right\rfloor, i = 2 \ldots n$$

(1)

Note that this might not yield an integer value and care needs to be taken to adjust for rounding and distributing the right number of reads to each processor.

After having obtained its corresponding reads, each processor uses the chosen string pattern matching algorithm to find the alignment in the reference genome. A wide variety of pattern matching and alignment algorithms exist. We arbitrarily chose the Knoth-Morris-Pratt [?] and the Boyer-Moore [1] algorithms to compare performance of different alignment algorithms in our framework. However, any pattern matching algorithm could be implemented and inserted in the proposed framework. The algorithm returns an
They were able to partition more than 1.6 million fragments of over 1.25 billion nucleotides total size of maize genome into genomic islands in under 2 h. Moretti et al. introduced a different approach on the parallelisation of the assembly by using a campus grid. They used the Smith-Waterman algorithm to align 8 million reads of length 750 base pairs. They required about 18 hours on a the campus grid [?] for the whole genome assembly.

To conclude, our framework allows highly parallel genome assembly. While direct comparison with other methods is not straightforward due to use of different compute systems and data, the proposed framework suggests to yield fast genome assembly that could overcome the computational bottleneck for usage in routine NGS assembly workflows.

Acknowledgement
This research was supported by a Victorian Life Sciences Computation Initiative (VLSCI) grant number 1049 on its Peak Computing Facility at the University of Melbourne, an initiative of the Victorian Government, Australia. We thank Prof. Justin Zobel, Department of Computing and Information Systems, University of Melbourne, to host this student project as well as Dr. Kelly Wyres, IBM Research - Australia Laboratory for selecting the genome sequence used in this study.

Bibliography
Abstract: Chordoma is a malignant bone cancer that develops from notochordal remnants. The MUG-Chor1 chordoma cell line shows a phenotypic anomaly, as it comprises two subpopulations, which were separated by micromanipulation and analysed with Affymetrix microarrays. Such a low-template microarray analysis, however, poses a difficult statistical problem, as the variances between biological replicates are high, whereas the biological differences are modest. Hence statistical approaches to increase the detection power are required. Firstly, we have applied a filtering step, which could slightly improve the detection power. Secondly, we have conducted a gene set test shifting the analysis from single gene level to gene set level, as such techniques are more sensitive in detecting moderate changes when these changes affect a set of related genes.

Keywords: Chordoma, microarray analysis, low-template analysis, gene set test, detection power, filtering

Introduction

Chordomas are rare mesenchymal tumours occurring in the midline from clivus to sacrum. Therapeutic modalities are mainly restricted to surgery and irradiation. Additional treatment options are therefore urgently sought. Chordoma cell lines such as MUG-Chor1 are characterised by heterogeneous cell populations, ranging from small, non-vacuolated to large, vacuolated cells (Fig. 1). To characterise the differences in gene expression, we have separated the two subpopulations by micromanipulation and analysed them by means of Affymetrix microarrays. mRNA expression in single cells, however, can be very heterogeneous and thus can lead to high variances in biological replicates [1]. Furthermore, only modest changes in gene expression are expected, as the two cell types derive from the same cell line. These conditions impair the detection of differentially expressed genes using traditional analysis techniques. As a high number of hypotheses are tested in microarray experiments, numerous false positives will occur by chance. To avoid this problem, the p-values need to be adjusted for multiple testing, which in turn leads to a decrease in statistical power and often not a single gene meets the thresholds to be considered as significant. To improve the detection power, we have applied a filtering step to increase the statistical power of traditional techniques [2] as well as used a gene set analysis approach to detect differential expression of a particular gene annotation category. Gene set tests are knowledge-based analysis approaches and offer several advantages over traditional microarray analysis methods with the most important being the higher sensitivity in detecting moderate changes [3].

Methods

20 cells of each of the two subpopulations (type A - \( \varnothing \geq 40 \mu m \) and type B - \( \varnothing \leq 15 \mu m \)) of the MUG-Chor1 cell line were picked by micromanipulation in triplicates and processed with the NuGEN WT-Ovation™ One-Direct RNA Amplification System. The cDNA was fragmented and labelled by means of the NuGEN FL-Ovation™ cDNA Biotin Module V2 and finally hybridized to Affymetrix GeneChip® Human Gene 1.0 ST arrays.

Firstly, the data has been analysed using the Partek® Genomic Suite™ v6.6 and Panther 8.0. After preprocessing, differentially expressed genes have been computed using a one-way ANOVA. Secondly, the data has been preprocessed and analysed using a variance shrinkage approach in R 2.15.2. To increase the statistical power, genes that were not expressed in the two subpopulations and genes showing no expression changes across all samples have been removed. For computation of differentially expressed genes, the ‘Limma’ R package was used. p-values were adjusted for multiple testing with Benjamini and Hochberg’s method to control the false discovery rate and genes with a p-value <0.05 and a |log2-fold change| >1 have been selected as potentially significant. Thirdly, a self-contained gene set test,
GlobalTest [4] has been employed to detect differentially expressed sets of related genes, testing available gene sets associated with gene ontology (GO) terms and KEGG pathways as well as all curated datasets available in MSigDB. All p-values were adjusted for multiple testing with Benjamini and Hochberg’s method. GO associated and curated MSigDB gene sets with a p-value < 0.05 as well as KEGG pathways associated gene sets with a p-value < 0.1 have been selected as significantly differently expressed.

Results

Traditional analysis of the chordoma microarrays using a one-way ANOVA did not result in any differentially expressed genes after adjusting the p-value for multiple testing. Using a variance shrinkage approach improved the detection power slightly, but still did not result in any significant genes after adjusting the p-value. A filtering step, however, could further increase the statistical power and led to the detection of four differentially expressed genes.

The gene set analysis, GlobalTest, in contrast, resulted in 398 significant GO terms, 60 significant KEGG pathways and 118 significant curated gene sets and thus was powerful enough to overcome the statistical difficulties of low-template microarray analyses. An example plot for the gene set associated with the regulation of lipase activity is shown in Fig. 2, visualising the significant subtree of this gene set.

Discussion

The transcriptional state of a cell under different genetic, physiological, and pathologic conditions or at different stages of development is often characterised by moderate changes in a set of related genes rather than a strong change in individual genes. Such moderate changes may be missed by single gene analysis methods as they rely on strict thresholds. This is particularly pronounced in low-template analyses where the biological differences are modest and the variances between the biological replicates are high. In such scenarios, not a single gene may meet the requirements to be considered as statistically significant after correcting for multiple testing. Filtering techniques and gene set analyses can serve as a remedy and hence are able to improve the detection power, as we have shown here.

Variance shrinkage approaches have more statistical power than simple one-way ANOVA analyses, in particular with small sample sizes, as they borrow information across all genes. However, these alone were not powerful enough to identify valid differentially expressed genes for our experiment. A filtering step decreased the number of genes by eliminating the uninformative fraction and led to the detection of four significantly differentially expressed genes. Moreover, the self-contained gene set test, GlobalTest could successfully detect differentially expressed sets of related genes. Moderate change in individual genes may not be significant considering each gene separately. However, by defining a relationship between a number of genes, moderate changes can be detected as significant, as we have shown here. Apart from increased detection power, screening for the differential expression for a given set of genes can also associate the expression profile with a phenotype of interest and thus facilitate data interpretation.

In conclusion, our results show that gene set tests have more detection power than single gene level analyses and are important alternatives in analysing high-throughput experiments such as microarray and RNA-seq data, where multiple testing easily can lead to insufficient statistical power.

Acknowledgement

This work was funded by the DK Molecular Medicine and by the Austrian Centre of Industrial Biotechnology (ACIB) supported by the Austrian BMWFJ, BMVIT, SFG, Standortagentur Tirol and ZIT through the Austrian FFG-COMET Funding Program [FFG Grant 824186].

Bibliography


Abstract: Technological improvements have shifted the focus from data generation to data analysis. The availability of huge amounts of data like transcriptomics, proteomics and metabolomics raise new questions concerning suitable integrative analysis methods. We compare three integrative analysis techniques (co-inertia analysis, generalized singular value decomposition and integrative biclustering) by applying them to gene and protein abundance data from six life cycle stages of Plasmodium falciparum. We create a network view of the GO terms associated to cell cycle stages by all three methods.

Keywords: Integrative analysis, comparison, genes, proteins.

Introduction

Continuous technological improvements facilitate the availability of huge amounts of data resulting from the simultaneous characterization of the same organism or experimental condition. It is possible to measure the activity of thousands of genes, hundreds of proteins and hundreds of metabolites. Only the integrative analysis of all data types yields a deeper understanding of the system under study.

Methods

In this work we concentrate on gene and protein data. Most of the current analysis techniques are based on the assumption of a direct correlation between genes and proteins. This assumption does not hold due to post-transcriptional and post-translational expression regulation processes. Here we compare three alternatives to conservative analysis techniques. Co-inertia analysis (CIA) is an integrative analysis method used to visualize and explore gene and protein data [1]. The generalised singular value decomposition (GSVD) [2] has shown its potential in the analysis of two transcriptome data sets. Integrative Biclustering (IBC) applies Biclustering [3] to gene and protein data.

We compare CIA, GSVD and IBC by applying them to gene and protein abundance data of Plasmodium falciparum [4]. The data was gathered from samples in six life cycle stages of the parasite: merozoite, ring, trophozoite, schizont, gametocyte and sporozoite. For the comparison we add additional information in from of gene ontology terms related to biological processes.

Results

Using CIA we visualize in Figure 1 the six life cycle stages and GO terms in a 2D plane. Each cell cycle stage is represented by it’s projection in gene (circles) and in protein (squares) space connected through a line. The smaller the line between the two projections, the higher the concordance between the gene and protein data sets. Here we observe a very good agreement between the two data sets for all cell cycle stages. We notice the strict separation of the intraerythrocytic life cycle (ring, trophozoite, schizont and merozoite) from sporozoite and gametocyte stages.

Additionally, CIA offers the possibility of projecting GO terms (represented by numbers) onto the CIA plot and to associate them to a certain cell cycle stage (see Figure 1). The association of GO terms to the life cycle stages was done as follows. GO terms with positive x coordinates were associated to the stages of the intraerythrocytic life cycle. Due to the very close spatial position of these four stages no further discrimination between GO terms was possible. GO terms with negative x coordinates are associated to the gametocyte stage if they have positive y coordinates and to the sporozoite stage if they have negative y coordinates.

With GSVD we decompose the data sets in matrices with biologically meaningful interpretations (arraylets, generalized eigenvalues and genelets) and explore the processes captured by them. The genelets represent biological processes captured by the data sets and expressed in the corresponding arraylets with a relative significance measured by the generalized eigenvalues. GO terms are associated to cell cycle stages through gene/protein set enrichment analysis based on the cell cycle stage depended angular distances (see Figure 2). All cell cycle stages are associated with both gene and protein space resulting in a gene and protein set enrichment analysis of genes and proteins showing
the highest absolute values in the corresponding arraylets. Biclustering was applied to the gene, protein, life cycle stages and GO terms. The six life cycle stages are represented in Figure 3 by the left set containing the green (gametocyte), brown (trophozoite), red (ring), dark blue (schizont), light blue (merozoite) and pink (sporozoite) squares. Different biclusters are represented by distinct edge colors. The genes are colored in orange, the proteins in light blue and the GO terms in yellow. One can see that there are gene and proteins that strictly belong to one cluster (all edges of these nodes have the same color) as well as others that are associated to more than one bicluster (edges of these nodes have more than one color). GO terms are related to genes, proteins and cell cycle stages and they are connecting different biclusters.

We compare the results of the three integrative analysis methods showing in Figure 4 GO terms association to cell cycle stages common to all methods. Cell cycle stages not belonging to the intraerythrocytic cycle are either completely disconnected from the other stages (sporozoite) or connected by only one node (gametocyte linked through glycolysis) to the rest. The cell cycle stages of the intraerythrocytic cycle are densely interconnected.

**Discussion**

We have started this analysis with 4294 genes, 2903 proteins and 248 GO terms measured and annotated during six cell cycle stages of *P. falciparum*. The results of all methods were summarized in a GO term/cell cycle association network with 34 nodes (cell cycle stages and GO terms) and 42 edges. In concordance with the literature we observe a strong connectivity between the intraerythrocytic cell cycles and a low or non connectivity to the other stages. Each method produces a vast amount of results which are tedious to interpret. Inspection of the common associations is not only faster but it is more reliable and relevant because the results undergo a triple validation.

**Figure 4:** GO terms association to cell cycle stages common to all methods.

**Acknowledgment**

This work was funded by the Austrian Centre of Industrial Biotechnology (ACIB) supported by the Austrian BMWFJ, BMVIT, SFG, Standortagentur Tirol and ZIT through the Austrian FFG-COMET- Funding Program [FFG Grant 824186].

**Bibliography**

HIGH-THROUGHPUT CHARACTERIZATION AND COMPARISON OF MICROBIAL COMMUNITIES

Bettina Halwachs¹,², Johann Höftberger¹, Gernot Stocker¹, Rene Snajder¹
Gregor Gorkiewicz³ and Gerhard G. Thallinger¹,²

¹Institute for Genomics and Bioinformatics, Graz University of Technology, Austria
²Core Facility Bioinformatics, Austrian Centre of Industrial Biotechnology, Austria
³Institute of Pathology, Medical University of Graz, Austria

bettina.halwachs@acib.at

Abstract: The analysis of the huge amount of generated sequence data as well as pyrosequencing noise and chimeric sequences originating from PCR amplification pose a considerable challenge to the individual researcher in doing microbiome studies. The unbiased knowledge about microbial community composition and -structure as well as the interactions with the human host microbiome can give important insights into its role in human health and disease. Here we introduce SnoWMAn, the high-throughput microbiome analysis pipeline and additionally investigate the effects of sequencing noise on non denoised and data denoised using two different approaches.

Keywords: Next Generation Sequencing, Community Composition Analysis, Denoising, OTU inflation

Introduction

The overall goal of human microbiome studies is to represent complex community composition within a certain habitat of interest and compare it under different conditions, between time points or patients. To characterize and classify complex microbial communities gained directly from environmental samples, a certain variable region of the commonly shared 16S rRNA marker gene is directly amplified and sequenced. Before generated sequences can be classified into operational taxonomic units (OTUs) some preprocessing and filtering steps should be applied to guarantee unbiased community composition representation. Especially when working with pyrosequencing data, noise originating from longer homopolymer stretches (> 4 bps) can lead to an increase in OTUs called OTU inflation. Besides sequencing noise perceived diversity can be increased by chimeric 16S amplification products which were formed out of two or more sequence templates during polymerase chain reaction (PCR). During further analysis these hybrid products can be falsely interpreted as novel organisms, thus inflating apparent diversity and finally lead to false conclusions. The general microbiome data analysis workflow is illustrated in Fig. 1, where for each step a variety of tools and approaches are available. To simplify microbiome analysis from preprocessing over OTU picking to the final statistical analysis and visualization of the result, we developed the web-based analysis pipeline SnoWMAn. It addresses shortcomings of existing tools, such as number of sequences which can be analysed, reproducibility and usability. Additionally, SnoWMAn is unique in covering the complete analysis workflow, offering different analysis pipelines and reference databases as well as capabilities for statistical analysis and visualization.

Methods

To demonstrate the capabilities of SnoWMAn and to show the effects of pyrosequencing noise we reviewed a previous study on changes in the gut microbiome during diarrhea [1] by denoising the data with Acacia [2] and the mothur [3] implementation of AmpliconNoise respectively. Furthermore, contaminating sequences originating from the host genome as well as potential chimeric sequences had been removed from the amplified sequences by a BLAST approach and uchime respectively. OTUs were built using the Ribosomal Database Project (RDP)-Pyrosequencing approach and uchime respectively. OTUs were built using the Ribosomal Database Project (RDP)-Pyrosequencing approach using the Infernal alignment v1.1 [4] and a maximal cluster similarity of 6 and similarity steps of 1 %. Additionally, quality filtering based on given quality values per base, number of Ns (discard sequences containing Ns) and length (discard sequences < 150 bp) was applied. Final taxonomic classification was done by the RDP classifier 2.4 [5].

Results

In addition to the RDP pipeline used here, SnoWMAn users can chose according to the field of application and their study design between two reference based OTU picking pipelines (BLAT, JGast) and three de novo OTU picking pipelines (mothur, RDP, UCLUST). Depending on the selected pipeline various preprocessing- and pipeline parameters such as the applied reference database, the classification model and the clustering settings can be specified. To minimize analysis time as well as to improve result qual-
ity it is possible to filter sequences according their length, maximal mount of unidentified bases, or mean sequence quality thresholds. The newly introduced Acacia denoising tool was integrated into SnoWMAAn and can be used on demand for identification and removal of noisy sequences. Not only that it was shown to be about 2000x faster than existing tools, additionally, our comparison results in modest difference between the mothur re-implementation of AmpliconNoise and Acacia (see Tab. 1). In respect to chimeric sequences SnoWMAAn integrates mothur’s uchime for optional chimera detection and removal. After all necessary and optional parameters are specified the numerical intensive analysis task is automatically started. Once the calculation is finished SnoWMAAn offers various capabilities for statistical analysis and result visualization such as rarefaction curves for microbial diversity estimation and illustration of species richness (alpha diversity). Species turnover or beta diversity can be calculated and visualized using heatmaps. Comparison of individual microbiomes can be done by the integrated principal component analysis (PCA). Barcharts or piecharts can be used to represent the number of sequences for each sample and give an overview of sequence yields. Additionally, cumulative and endpoint depth of the taxonomic classification can be graphically illustrated. Line plots can be used to reveal sample composition at a specific taxonomic rank to point out compositional microbiome changes over time. Data can be presented in relative or absolute scale for all chart types. All the generated data can be easily exported either as Excel file or as figures in PNG or SVG format. The comparison of the effects of sequencing noise on community diversity results in enormous OTU-inflation when comparing the number of OTUs resulting from denoised vs. non denoised pyrosequencing data, see Tab. 1. Surprisingly, the number of potential chimeras varies depending on the denoising approach, especially for fecal samples. Moreover, the number of OTUs varies more than expected between AmpliconNoise and Acacia.

**Discussion**

Here we introduced SnoWMAAn as a comprehensive system for high-throughput analysis of microbial community sequencing data as well as the effects of two different denoising approaches. SnoWMAAn covers the whole microbiome analysis workflow and offers the two most common analysis approaches in one single pipeline. The user-friendly and intuitive web-interface makes it a convenient resource not only for classification and characterization but also for statistical analysis, visualization and reusing or sharing of the analysis result. Furthermore, the newly integrated denoising and chimera filtering tools satisfy latest findings towards sequencing noise. Although different denoising approaches showed modest variation of noisy sequences the effect on the number of chimeric sequences needs further investigations. The modular design of SnoWMAAn allows simplified extension of the classification tools to other genes than the 16S rRNA by providing appropriate reference databases and alignment models.

**Acknowledgement**

This work was funded in part by the Austrian Centre of Industrial Biotechnology (ACIB) supported by the Austrian BMWFW, BMVIT, SFG, Standortagentur Tirol and ZIT through the Austrian FFG-COMET- Funding Program [FFG Grant 824186].

**Bibliography**


<table>
<thead>
<tr>
<th>sample</th>
<th>contam.</th>
<th>no denoising</th>
<th>AmpliconNoise</th>
<th>Acacia</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>raw OTUs</td>
<td>noise chimera ok</td>
<td>OTUs</td>
</tr>
<tr>
<td>Feces</td>
<td>4</td>
<td>327439</td>
<td>3755</td>
<td>23810</td>
</tr>
<tr>
<td>Mucosa</td>
<td>3558</td>
<td>187773</td>
<td>3077</td>
<td>25570</td>
</tr>
</tbody>
</table>

Table 1: Number of sequences and OTUs at a cluster similarity of 0.03, without denoising as well as denoised by Acacia and the mothur re-implementation of AmpliconNoise. Contaminating sequences as well as potential chimeric sequences have been also removed prior to analysis.
DEVELOPMENT OF A COMMUNICATION SERVER FOR LONG-TERM TELEMONITORING OF PATIENTS WITH COPD

J. Schaaf1, H. Schneider1, M. Kaiser1, L. Mursina1

1Competence Centre for Information Technology, THM - University of Applied Sciences, Germany

jannik.schaaf-2@mni.thm.de

Abstract: The COPD is one of the most chronic diseases in the world. To reach an integrated accommodation of these patients it is necessary to deploy a telemedicine monitoring. Basically the vital signs of the patients need to be captured with medical device technology. Then the data is send to different actors in healthcare. Data security and interfaces to different systems must be noted in this case. Within the scope of the infrastructure there is a need of a communication server which can transfer the data to different systems. The goal is that the server is developed preferably open, expandable, maintenance-friendly and economic.

Keywords: COPD, telemonitoring, communication server, web service, service-oriented architecture

Introduction

The demographic change and the advance of the age are very important. These factors have a big impact of healthcare. That means that the costs will increase more and more [1]. With this change the rate of chronic diseases will increase [2]. Especially chronic airway diseases are playing an important role, especially the COPD which is the 4th cause of death in the world.

Meanwhile the telemedicine has a big meaning in the accommodation of chronic diseases. The goal of telemedicine accommodation is defined as comprehensive information processing which can optimize the patient accommodation and care. Through these factors the quality of living should be increase [1]. Some studies are proving that longer and more exact studies are needed for evaluating the quality of living from COPD patients [3, 4].

Currently we are developing in a project a platform that documents telemonitoring continuous device parameters and safety aspects [1].

Within this project, a platform is been created that adjusts the different communication systems and their interfaces to each other. This platform will enable long-term studies in the field of telemedicine monitoring of COPD patients. To perform a perfect monitoring, it requires a patient selection. At this point, patients with gold stage 3 or 4 should be selected.

First the vital signs are recorded, using medical-technical devices in the home environment of the patient. Then, the data is transmitted to relevant institutions to provide integrated care. That includes doctors, treating hospitals, emergency services and medical service providers. The data collection of the vital signs of the patients should be done with a Smartphone. For example, oxygen saturation, blood pressure and respiration values are documented.

Subsequently, the collected data is transferred out of the home environment of the patient. To ensure this, all newly collected medical data needs to be sent from the Smartphone to the communication server. This essential hub for transfer provides the server. The transmission to the communication is via a mobile internet network. These data has to provide to other systems.

As particularly important is a mechanism seen which may detect a life-threatening condition of a patient. This data must be send via SMS.

Methods

The development of the communication server as part of the project requires on the one hand software engineering methods but also the variety of software systems that are needed for developing and operating the server.

The development was based on the software life circle within the context of the requirements engineering, functional and non-functional requirements were defined, which should be placed on the created system. Furthermore, the design of the server, which was divided in course and fine design, has been defined. In the rough draft the global software architecture and subsystems were modeled. In the detailed design the classes and components for technical programming were created. This was followed by the implementation of the functionality of the server and adequate software testing.

Results

In the requirements analysis, the most important functional and not-functional requirements were defined. Here the most important requirement was an encrypted data transfer between phone and server to allow communication. The communication server checks if the transferred data is correct and gives feedback whether the data were correct or incorrect.

Subsequently, the successfully transmitted and checked data are stored persistently in the database. After this process a check to critical monitoring parameters is made.

If critical data is recognized, there must be a possibility of sending an SMS with this information. Furthermore, the possibility of communicating with other medical information systems, was defined.
The decision fell on the open-source software, Mirth Connect, which enables a unified and standardized information exchange with the communication standard HL7 in healthcare. Recently, the possibility of message exchange between doctors and patients can be guaranteed. The demand for platform independence led to the use of SOA-based web services. As seen in Figure 1, the communication between the individual components is only via the central SOA component that is implemented with web services possible.

![Figure 1: Service-oriented architecture of the Servers](image)

The implementation was done in the .NET Framework based on the Windows Communication Foundation and the C# programming language. The essential basis of the communication server is the server operating system Windows Server 2008 R2 on which the instances were installed, which were necessary for the use of the communication server. This includes a Microsoft SQL Server 2008 R2 that provides the database and Internet Information Service (IIS) that hosts the web server. Finally, the actual program code for the web service was developed in the programming language C#.

With help of white-box testing, the inner function of each software component has been tested.

**Discussion**

With the platform a telemedicine infrastructure is created which provides an integrated and improved care of patients with COPD. At this point we will see whether increasing the quality of life of the patients and reducing health costs with respect to the disease is possible.

Based on the determination of a certain group of patients, and other essential parameters, such as lung sounds, heart rate or oxygen saturation, the study can establish from the previous scientific studies to make more accurate and prolonged investigation.

The required communication infrastructure must prove stability, extensibility, maintainability and an ideal information exchange between all involved components.

The communication server provides with its service-oriented architecture as a web service and independent and bundled component in the communication infrastructure.

The decision to use a service-oriented architecture was especially important to carry out extensions and changes easily.

**Acknowledgement**

We thank the Hessian Ministry for Science and Art for financial assistance under the state program LOEWE.

**Bibliography**


TELEMONITORING FOR COPD PATIENTS: STANDARDIZED BLUE-TOOTH INTERFACE BETWEEN MEDICAL AND MOBILE DEVICES

S. Pfeiffer¹, H. Schneider¹, M. Kaiser¹, L. Mursina¹

¹Competence Centre for Information Technology, THM - University of Applied Sciences, Germany

stefanie.pfeiffer@mni.thm.de

Abstract: Increasing number of people owns a smartphone or tablet, so that these devices serve as a gateway in the field of telemedicine. Treatments with telemedicine gain importance to the health system in Germany. Especially, patients with lung disease benefit from this kind of care. The effort of this project is an innovative concept to receive personal health data with a mobile phone. The data transmission is realized by a standardized Bluetooth interface. This interface is used for a telemedical monitoring of patients with chronic pulmonary diseases.

Keywords: Telemedicine, IEEE 11073, interoperability, mobile devices

Introduction

Chronic obstructive pulmonary disease (COPD) is an irreversible illness of airways that goes along with exacerbations. Exacerbation is defined as a persistent worsening of symptoms [1]. The treatment of an exacerbated COPD is often performed in a hospital which causes high health care costs [2]. Telemedicine makes it possible to monitor the therapy and recognize complications early. Therefore costs as well as frequency of exacerbations are reduced effectively [3]. Previous studies show that especially COPD patients benefit from telemedical health care [4].

Telemonitoring can be carried out by different types of medical devices to record information about the patient’s condition. For the reason that the information is complete, the compatibility of the interface plays an important role. The challenge of this application is the interoperability between medical devices. The aim of this project is to create a concept for a wireless application which receives personal health data via a standardized Bluetooth interface.

Methods

Different platforms support the development for mobile application (Android, iOS, Windows Phone, Blackberry, etc.). The experts describe the Android operating system (OS) as one of the most flexible [5]. In addition, Android introduces a framework which supports the Bluetooth Health Device Profile (HDP). This framework makes it possible to create applications that use Bluetooth to communicate with medical devices in consideration of the ISO/IEEE 11073-20601 [6]. Therefore, Android is the favorite OS for this application. A pulse oximeter is used as medical device for implementation and test purpose. This device detects the arterial oxygen saturation non-invasive.

Android requires the software architecture pattern model-view-controller. The resources, which are characterized by format and usage, are used as model. The resources encapsulate the data and make them available to the controller. The controllers receive and react to the user input. For this purpose Android uses activities. Thus activities represent the logical part of the application. The responsibility of the view is to show the data. The design of the view is determined by layouts. The user inputs are entered with the touchscreen of the mobile device. Tablets offer more space to show information in contrast to smartphones. In order to support multi-screen, Google introduces fragments in Android. Fragments divide the Activities in encapsulated and reusable components that are displayed differently in tablets and smartphones. The display of the tablet can split into several areas.

Results

It is possible to connect a pulse oximeter with a first prototype of the application. The pulse oximeter sends the data via Bluetooth interface which uses the HDP. (Figure 2).

Figure 1: Different view on tablets and smartphones based on [7]

The action bar has access to the option menus in the resources and is therefore able to provide further functions of the application.

A database is used to store the received data. Google has integrated a compact library in Android, to enable the development of databases for applications. This library supports the SQLite database.

Results

It is possible to connect a pulse oximeter with a first prototype of the application. The pulse oximeter sends the data via Bluetooth interface which uses the HDP. (Figure 2).
Figure 3 shows a screenshot from the application that lists the medical devices. The action bar on top includes the logo, the title and the menu.

The menu provides several actions. By operating the arrows in the menu, the application establishes a connection with a selected device. The disk symbolizes the storage of a new medical device into the database. Furthermore, the patient can make the smartphone discoverable to other devices, by selecting this option within the application. To register the application as health sink, the user selects “App registration”. After the initial configuration, the connection to the device is established automatically.

**Discussion**

The result of this project implies that it is possible to develop a medical application which can be used for telemonitoring. The user interface between patients and software is the touchscreen. To engineer an application that is in line with the android core application, Google recommends the action bar. This enables intuitive handling. The data transmission is realized with the Bluetooth HDP. Therefore the application can communicate with all compatible pulse oximeters. However, currently only a few devices are available. This is unfortunately, because the use of data transmission via Bluetooth has three main advantages: First, the data transmission is wireless. As a consequence, no additional stumbling blocks in the home environment of the patient arise. Second and third, Bluetooth uses only a low power drain and is network-compatible [9]. The communication between medical devices that use HDP must not have less than Bluetooth version 3.0. That implies that for this application only new mobile devices can be used.

For the use in telemedicine, the application appears to be very promising. The main advantage is the interoperability which is achieved by the standards and the HDP. Because of this, every Bluetooth-compatible pulse oximeter, which uses the HDP, can communicate with the application.

**Acknowledgement**

We thank the Hessian Ministry for Science and Art for financial support within the national program LOEWE.

**References**

HARVESTING PATIENT OPINIONS ABOUT IMPLANTS FROM SOCIAL MEDIA SOURCES

Yves Noirjean, Markus Degen
Institute for Medical and Analytical Technologies, University of Applied Sciences and Arts Northwestern Switzerland, Switzerland
yves.noirjean@fhnw.ch

Abstract: The increasing use of social media (e.g. Facebook, Twitter, blogs, discussion boards) provides a new valuable information source for implant manufacturers about the public reception of a specific implant. On the other hand, patients and caregivers can get first hand experiences from other patients. However, because of the overwhelming amount of data available, tools are required to locate useful information. Often it is more important to get the "Wisdom of the crowd" instead of a single opinion. This paper describes a system which can be used to harvest data from social media and use different filters and analysis tools to aggregate search results.

Keywords: social media, sentiment analysis, cochlear implant, knee replacement, hip resurfacing

Introduction

Implant manufacturers are very interested in getting first hand and subjective information about the reception of their products at the patient side. The use of social media by users of medical products (e.g. patients who carry an implant) hold significant information for designers of the respective products [1]. There are common text analysis tools which can be used to get statements about a text without having to read the whole text. While most tools can be applied independent of the domain, three groups of medical implants were chosen as a showcase in this project: cochlear implants, knee replacements and hip resurfacing.

Methods

Figure 1 shows the architecture of the software system built.

Figure 1: System overview

Sources were manually chosen, and then continuously harvested, converted to a unified data model and written to a document store.

For the Twitter source, a Naive Bayes classifier [2] was trained with 1196 tweets. The tweets were manually tagged as wanted or unwanted where the wanted tweets contained personal experiences of a patient with an implant. After filtering with the classifier, the texts were grouped (e.g. by implant model or manufacturer) and analyzed. The remaining texts were used to form a control group.

Metrics

The following metrics were applied to each document:

**TF-IDF:**

Calculates the Term Frequency $tf$ (Eq. 1) multiplied by the Inverse Document Frequency $idf$ (Eq. 2) for a given set of terms. $tc$ represents the term occurrence count, $dc$ the count of documents containing the term and $tdc$ is the total document count.

$$tf(tc) = \begin{cases} \log_{10}(1 + tc), & \text{if } tc > 0 \\ 0, & \text{if } tc = 0 \end{cases} \quad (1)$$

$$idf(dc, tdc) = \log_{10}\left(\frac{tdc + 1}{dc + 1}\right) \quad (2)$$

$tf \times idf$ equals to the prominence of the analyzed terms [3].

**Sentiment Analysis:** SentiStrength [4] was used to estimate positive and negative sentiments expressed. A t-test resulting in a $p$-value was done for each pair of groups to determine the observed significance level of each comparison. Those results, along with the mean and the standard deviation for each group, are stored in the result document.

Results

The project is currently in its last stage (evaluation). Therefore the results shown here are preliminary.

Source evaluation

Figure 2 shows the quantity of hits for a selection of keywords in the domain of cochlear implants and hearing aids for different sources. The result demonstrates how unspecific queries lead to many hits which can be reduced by using more specific ones. For the alldeaf internet forum [5] this is true to a lesser extent because it is specific to the domain of hearing impairment.

Tweet classification quality

143 of the 1196 manually tagged Tweets were wanted. The classifier trained with half of the tagged set had a sensitivity
of 0.97 and a specificity of 0.72 when tested on the other half. During four months, 329666 Tweets were collected. Of those, 251022 (76%) held original content (no retweets) and 25149 (7.6%) were classified as wanted.

The mean and standard deviation values indicate a group of outliers within group B. Manual analysis of the texts for investigation confirmed the results and revealed reports of people having frequent visits to an audiologist for readjustments.

Table 2: mean and sd for TF*IDF of word "adjustment".

<table>
<thead>
<tr>
<th></th>
<th>control</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>0.0028</td>
<td>0.0074</td>
<td>0.0222</td>
<td>0.0049</td>
</tr>
<tr>
<td>sd</td>
<td>0.0469</td>
<td>0.0849</td>
<td>0.1368</td>
<td>0.0592</td>
</tr>
</tbody>
</table>

### Discussion

We have built a system which can be used to harvest patient opinions from several online sources. The system is fully configurable and easily extendable for further online sources and analysis methods. First results from analyzing the harvested data look promising.

Of the 445267 documents collected so far, 115601 are from the alldeaf forum and 329666 are from Twitter. The content of those documents totals to 106 MB of raw text data. With the planned addition of Facebook as a source and continuous harvesting of documents, these numbers will increase. The classifier based filtering of Twitter content improves the quality of the results at the cost of quantity. For the already analyzed search terms, the most valuable data is found in the domain specific, manually selected forum. The statistical results show that it is possible to find relevant aspects of implant product groups. Queries done while testing the system showed that it is important to carefully choose meaningful search terms - on one hand they have to be specific and on the other hand they have to be generic enough to allow a certain degree of fuzziness.

### Outlook

This project is now in its final stage and the main focus is now on the analysis of the harvested data. The final results will be available at the end of the project by mid 2013. Extensions to the system are conceivable in several areas:

- Improve classification methods, i.e. according to [7], which would probably lead to a better specificity
- Weight posts by analyzing the expertise of authors
- Add domain specific knowledge to the analyzer (i.e. extract "range of motion" after knee replacement)
- Add more data sources, i.e. use an aggregating search engine

### Bibliography


A FRAMEWORK FOR MULTI-MODEL SURGICAL WORKFLOW MANAGEMENT

Franke S¹, Neumuth T²

¹Innovation Center Computer Assisted Surgery, Universität Leipzig, Germany
²Innovation Center Computer Assisted Surgery, Universität Leipzig, Germany

Abstract: Surgical workflow management will make important contributions to future digital operating rooms. It allows the automation of tasks like pre-configuration of medical devices or resource logistics. The applications differ in the required process information aspects. Hence, a framework for flexible networks of process models is proposed based on the well-established pipeline concept. It allowed the integration of multiple different surgical process model types in one system. A prototypical framework was implemented and tested with a network for a neurosurgical use case.

Keywords: Surgical workflow, OR integration, workflow management, surgical assistance, digital operating room

Introduction

Surgical workflow management will make important contributions to future digital operating rooms. It allows the automation of tasks like pre-configuration of medical devices or resource logistics. Thus, it may unburden the surgeon and the OR staff peripheral responsibilities. Technical systems for workflow management in the surgical environment need to cope with high inter-process variance and uncertainty. Hence, new types of process representations beyond common business process models have been developed [1-3]. Most of them were adapted to specific aspects of workflow management. The establishment of a comprehensive intra-operative assistance requires a flexible integration of different methods for process description and management.

Methods

Various fields of application for surgical process information in intra-operative and peri-operative environments have been identified. The main use cases were information presentation [4], automatic configuration and orchestration of medical device functionalities [5], documentation, resource management and OR scheduling [3]. Each application required surgical process information. However, they differ in the required process information aspects. Hence, different types of process models were developed, tailored to specific perspectives and applications. A surgical workflow management system (sWFMS) should be designed for flexible integration of multiple surgical process models. Moreover, the process models may be interdependent as well. Consequently, a network of process models of different types was required. Hence, a network of components needed to be set up which reflects these dependencies within the architecture of the sWFMS. Also there is a need for functionalities in addition to the process model handling. Accordingly, two types of components could be distinguished: processors and modules.

Each process model was handled by a dedicated processor component. The processor combined input data with the process model and generated an output which was a certain aspect of surgical process information. The processor might also adjust the model during processing. Any component which did not own a process model was considered as a module. Typical tasks of a module were information fusion and external communication. This included for example information summary, estimation of characteristic quantities, information transfer and documentation.

Different intervention types might require different process model types and assistance modules. A generic conception of a surgical workflow management system should be applicable to a wide range of surgical intervention types. Hence, a flexible and extendable configuration of the component network was required. We developed a technical framework for this purpose. The framework used a pipeline concept to build up the component network and provided buffered and non-buffered connections with modification timestamp. The dynamic inter-connection required a machine-readable description of the I/O ports of each component. The ports were described by a unique label, data type information, a coding for semantic annotation and a human-readable short description. Furthermore each port could be defined as required or optional to the function of the component.

The core unit of the sWFMS managed the network of the components. This task included the pre-validation of the network configuration, the import of the required process models, the launch of the components and their interconnection as well as supervision of their functioning. A comprehensive pre-validation of the network configuration was established based on the component port descriptions. Hence, the sWFMS could automatically identify blocking circular dependencies, missing required inputs as well as unused outputs.

The process models were imported either locally or via web service interface. Therefor a process model database
was established. It included persistent storage as well as on-the-fly generation of process models. Additionally, the core provided a centralized timing port to synchronize any intervention time dependent model processing. The overall network of processors, modules and core formed a flexible surgical workflow management system that was able to combine multiple process models in one system.

Results

We implemented a surgical workflow management framework following the proposed concept. The framework included processors for generalized surgical process models (gSPMs) [1,3] and process models based on Hidden Markov Model theory. Additional modules for comprehensive logging, connection to an OR bus [5], intervention time prediction [3] and process information visualization [4] were incorporated. Furthermore, a prototypical network configuration for neurosurgical brain tumour removal was designed and implemented. An overview of the processing network is given in Figure 1.

Figure 1. Schematic representation of a prototypical processing network for surgical workflow management in neurosurgical use case.

The network was built upon a module for workflow recognition input. Its input was simulated during testing. A generalized surgical process model (gSPM) provided fine-granular activity information. It served as base for with multiple processors for abstract surgical process models (aSPMs) based on Hidden Markov Model theory. The aSPMs represented different perspectives on the surgical procedure. The additional modules provided the generated overall process information to external consumers.

By now, the framework was technically tested with forty recordings of performed brain tumour removals on standard hardware. We successfully combined two different types of surgical process models and various modules in one sWMFS. The first technical tests demonstrated stability and showed sufficient performance for intra-operative application.

Discussion

We proposed a concept for the integration of tailored process models into a common framework for surgical workflow management. The surgical process models formed a network based on the well-established concept of pipelines. The machine-readable description of the component input and output ports allowed automatic pre-validation of the network configuration. Additional modules used the generated process information to provide intra-operative and peri-operative assistance functionalities.

In future work studies have to be conducted to examine the enhancement of surgical workflow management impact by combination of different process models. The present work proposed a robust and flexible framework for process model networks. Thus, it will contribute to a successful integration of surgical workflow management into future digital operating rooms.

Acknowledgement

ICCAS is funded by the German Federal Ministry of Education and Research (BMBF) and the Saxon Ministry of Science and Fine Arts (SMWK) in the scope of the Unternehmen Region with grant numbers 03Z1LN12 and by funds of the European Regional Development Fund (ERDF) and the state of Saxony within the frame of measures to support the technology sector.

Bibliography

This paper presents a project where a platform for telemonitoring and treatment protocols [4]. Afterwards these data can be used to optimize the monitoring and quality assurance of medical software to work out risk factors. This work is to result in a prototype which has the requirement to obtain a certificate.

**Keywords:** Medical information system, telemedicine, remote patient monitoring, web application, .NET

**Introduction**

Due to the steadily declining population in Germany, the number of inhabitants will sharply decrease within the next 50 years. The ratio between young and old people will change in favor of a continuous aging of the society [1]. Taking these demographic changes into consideration, one can assume a rising incidence of chronic diseases such as COPD (chronic obstructive pulmonary disease), as well as increasing costs of health care [2]. In this regard, telemedicine can be seen as a meaningful addition to traditional medical care because it opens up the chance to reduce treatment costs while increasing quality [3].

The endeavour of telemedicine is primarily to achieve qualitative and economic reforms as well as to enhance the transparency of medical care. One aspect of telemedicine is the telemonitoring used for remote monitoring of a patient's condition. To ensure such monitoring, the necessary data is collected by means of appropriate equipment. Afterwards these data can be used to optimize the monitoring and treatment protocols [4]. This paper presents a project where a platform for telemedical monitoring is developed and is intended for supporting long-term studies of COPD patients. Within the project, a platform component responsible for the control and evaluation of the collected data is designed and implemented in a prototype.

This work aims at highlighting the efficient and effective benefits of telemonitoring by a prototypical implementation. Simultaneously it is objective to identify and document security-critical aspects.

**Abstract:** The background of this work is the need for visualization of medical patient data, which can be used for optimizations of the monitoring- and treatment protocols. Advantages of the remote patient monitoring should be highlighted and safety-critical aspects should be documented by using a prototype implementation. The conception is to involve privacy concerns as well as standards for the development and quality assurance of medical software to work out risk factors. This work is to result in a prototype which has the requirement to obtain a certificate.

**Methods**

**Project management:** For planning and implementation a web-based open source software called "Redmine" is used. It allows managing project requirements using a ticket system and connecting them with descriptive text in the form of a wiki.

**Software Engineering:** As software development process an iterative and incremental process model is applied. In this approach the basic idea is to develop a system by repeating all activities of development process to detect any errors or unclear defined requirements as early as possible. For the test aspect to be considered at an early stage and to reinforce the previously mentioned effect, test-driven development methodology is used. In this way, the design activities are carried out much more consciously and a testable software design can be guaranteed from the beginning. Additionally, implementing test cases produces documentation close to source code.

**Environment:** The information system is implemented as a web-based application using the C# programming language in conjunction with the ASP.NET MVC 4 Framework. As an integrated development environment Visual Studio 2012 Ultimate is used. The Ultimate version has the most powerful features of the Visual Studio series and provides support for modelling, testing, and implementation.

Hosting is provided by a Microsoft Server 2008 R2, which is equipped with the Microsoft Internet Information Services 7 web server. The backend is supported by a Microsoft SQL Server 2008 R2 database.

**Class Libraries:** To develop the information system we use different existing class libraries in order to raise the focus especially on the implementation of the essential requirements. For an object relational mapping ORM (object relational mapping) the open source framework “ADO.NET Entity Framework” is applied. For the implementation of the tests the MS Test Framework takes place. Some of the tests will be so-called unit tests, which shall test small components isolatedly. Often simulations or the so-called “Mocking” for individual components eased drastically by using the class library “Moq” is needed.

The Design of the software has been created by the design pattern “Inversion of Control” (IoC) in form of the “Dependency Injection” pattern. The IoC container “Ninject”, responsible for the whole configuration for creating the
objects is used to keep the effort for its implementation as low as possible.  

**Version Management**: For the version management of the source code the distributed version control system “Git” is applied.

**Results**

At the beginning of the project the scope of the functional and non-functional requirements were determined in sufficient depth. The information system will be able to process collected measurement data of medical devices and display this data in an appropriate and user-friendly way. The frontend will be besides the connector and the communication server the third important part of the architecture of the project and will represent the interface between user and device. The most important functional requirement is the connection between the communication server and the visualization of the devices data. This data will be assigned to a patient for a certain time and will be displayed in form of a line chart. If further data from other devices exist for this patient at any moment, all data will be displayed in the same diagram. Fig. 1 shows a prototype of such a dashboard.

![Figure 1: UI Mockup - Dashboard](image)

The system shall bring medical data in a personal context and will therefore be only accessible for responsible and authorized persons. The security facet is very important in the area of non-functional requirements. We keep an eye on only following the highest and latest security standards to encrypt data and prevent unauthorized persons from getting access and compromising the system. During the whole development process a complete documentation of all constructive activity will be logged. A certification made out by an external institution is aspired.

**Discussion**

Displaying a huge amount of data in a clear and understandable way will be possible with the prototypical information system. By means of this graphic illustration the analysis results required for the optimization of the diagnostic shelf can be improved.

Risk factors occur especially in correlation with the security of the system. Medical data is brought in a personal context and allow discussions about privacy and data protection law. It is necessary to minimize the risk by identifying security critical aspects and initiate counter-steps. An additional risk factor is the authenticity of data. Thus, a mechanism needs to be developed to prove the authenticity and ensure reliability of the data.

**Acknowledgement**

We thank the Hessian Ministry for Science and Art for financial assistance under the state program LOEWE.

**Bibliography**