Translational Medicine today seems within the reach of any team of professionals with the access to biotechnology and internet, thanks to sequencing and publishing of human genome, and advances in commercially available biotechnology. It is a collaborative effort to bring the latest scientific results from basic research rapidly into applications in medicine. This can be exemplified by a provision of diagnostic tools, a development, and improvement of, procedures. Education, the spread of knowledge, also falls under this category. Polish authors publishing within this trend include Guzik (2010); Bartnik et al. (2012); Derwińska et al. (2012); Wiśniowiecka-Kowalnik et al. (2013), the latter three publications were prepared with the cooperation of the author while working on this thesis. More than 1000 patients were diagnosed by the Institute of Mother and Child (IMID), Warszawa, with the cooperation of Baylor College of Medicine, TX, USA, and with the help of the software developed for that occasion described briefly in Section 6. Many of these patients carry inborn genetic diseases: autism, epilepsy, mental retardation, certain heart defects. This group of genetic diseases is mediated by aberrations in the genomic material of the carrier. These diseases may be inherited, or they may be introduced de novo into patients genome during pre-embryonic and embryonic phases.

Pinpointing in a patients genome the exact genetic aberration responsible for the disease requires two things: i) a close look into patient’s genome sequence which allows to see mutations, e.g. segment deletions, segment duplications, ii) the understanding of each aberration’s consequences for the phenotype, i.e. patient’s health. The former problem is approached by several biotechnological methods, such as karyotyping, Fluorescent in-situ
Hybridization (FISH), DNA microarrays, and recently the full genome Next Generation Sequencing (NGS). This thesis is concerned with the analysis of data from Array Comparative Hybridization (aCGH) – a technique based on DNA microarrays, which is described in Chapter 2. The latter problem is constantly being solved through gathering of reports on patients, and their experiments results, in databases around the world. In Chapter 5 presented is the analysis connecting data from our IMID2py database of aCGH results from IMID with aforesaid external databases, in this case those are International Standards for Cytogenomic Arrays database (ISCA) (Faucett, 2010), Genetic Association Database (GAD) (Zhang et al., 2010) and Database of Genomic Variants (DGV) (Zhang et al., 2006).

The importance of information infrastructure in translational medicine cannot be overestimated. The publicly available reference sequence of human genome is essential in annotating, relating, and referring data from labs around the world. The arriving technology of Semantic Web, with marvelous tools for storing and referencing graph databases of knowledge, and the Linked Open Data (LOD) publishing method, are slowly transforming the landscape. In this regard in Chapter 6 summarized is an Early Adoption project of Apache Internet Knowledge Stack, where the IMID2py database is extended with LOD semantic data from UniProt and other databases.

The role of computational methods is undeniable either, with the expectation of sophistication constantly growing. The aCGH technology for genomic aberrations detection is based on statistical analysis and transformations of the outputted data. Plethora of software is implemented for this task, a survey of which is provided by Karimpour-Fard et al. (2010). In Chapter 3 of this thesis an integrated solution to the problem of segmentation and, at the same instance, noise separation in aCGH data is proposed; a solution rooted in the Bayesian framework of Graphical models and Markov fields. Furthermore, in Chapter 4 we implement a statistical measure, and a modification of a popular software, the circular binary segmentation (CBS) algorithm, which is used to compare the quality of different aCGH microarray designs.

Detection of DNA copy number changes in patient’s genome is crucial in precise diagnosis of genetic diseases, in understanding of thereof, and the aCGH technology was, and still is, pivotal in medicine, one of the reasons being some of the pathogenic changes are mosaic and not detectable in conventional karyotyping, reports Stankiewicz and Beaudet (2007).
DNA copy number changes, or Copy Number Variations (CNVs), are gains or losses of chromosomal material. They are associated with many types of genomic disorders like mental retardation, congenital malformations, or autism, according to Lupski (2009); Shaw et al. (2004). Genetic aberrations are characteristic of many cancer types and are thought to drive some cancer pathogenesis process, by Lai et al. (2007); O’Hagan et al. (2003); Snijders et al. (2005); Wang et al. (2006).

The aCGH technology is widely used for identification of segmental copy–number alterations in disease genomes, which is corroborated by many publications including Boone et al. (2010); Miller et al. (2010); Perry et al. (2008). In a typical experiment, DNA is extracted from two genomic samples (test vs reference) and labeled differently. Samples are mixed together and then hybridized to a microarray spotted with DNA probes. Signal fluorescent intensities of each spot from both samples are considered to be proportional to the amount of respective genomic sequence present. A more detailed description is found in Chapter 2.

The aCGH microarrays can be classified into two types. Targeted arrays aim in detection of known, clinically relevant copy number changes and thus provide a better coverage of selected regions, see e.g. Caserta et al. (2008); Thomas et al. (2005). On the other hand, the whole-genome arrays, provide a coverage of the entire genome Barrett et al. (2004). Each design is constrained by the number of DNA probes on the microarray – hundreds of thousands to more than a million probes on a microarray in 2014. Nevertheless, in many applications, especially clinical, the design of the array should combine these two goals resulting with the exploration of the whole genome, with the special focus on certain specific regions (e.g. containing genes related to the disease under study). An exon array CGH approach proposed recently accurately measures copy-number changes of individual exons in the human genome. Chapter 2 contains description and a diagram of an iterative process of aCGH microarray design, and in Chapter 4 a method to compare designs is outlined and tested.

Assigning significance to signals found in aCGH data is a challenging task, a combination of statistical analysis and of human verification by geneticists. Methods proposed in Chapter 5 aim at improving this task. It’s an ongoing effort: improving, automating, and verifying protocols for detection of rare CNVs which underlie diverse spectrum of diseases in human, a perfect example from translational medicine.
Collaboration and cooperation catalyzes translational medicine to happen. This thesis stems from the collaboration between the groups of Ania Gambin from the University of Warsaw, who lead bioinformaticians Tomek Gambin, and Maciek Sykulski (the author), and the group of Paweł Stankiewicz, who lead the Cytogenetic Lab at the Institute of Mother and Child (IMID), Warsaw, where the team was Barbara Wiśniowiecka-Kowalnik, Katarzyna Derwińska, Magdalena Bartnik, and others, under supervision of Ewa Bocian. Moreover, Paweł Stankiewicz connected Polish efforts with the efforts of Baylor College of Medicine (BCM), Houston, USA where leading research on aCGH microarrays and their clinical application takes place. The Polish teams at UW and IMID took part in the design and testing of V8.x OLIGO aCGH chip – a custom-designed array with approximately 180,000 selected "best-performing" DNA oligonucleotide probes on it – which is used as a research and diagnostic tool at IMID and BCM, of which reports in print were made by Bartnik et al. (2012, 2014); Boone et al. (2010); Derwińska et al. (2012); Wiśniowiecka-Kowalnik et al. (2013).

Main Results

We present the process of aCGH technology from particular to general and back. By that understood is the cycle of i) signal vs noise considerations of incoming aCGH microarray data, on which the design of microarrays is dependent, ii) processing of data stream in order to derive its segmentation: a structured signal, aligned along human genome used as reference coordinate system, this step is often termed Copy Number Variants (CNVs) calling, iii) collecting, and referencing collections of, CNVs, either benign or pathogenic, in a suitable data infrastructure, iv) assigning either medical, or evolutionary significance to CNVs, a step involving aggregation of knowledge from various sources on human genetics, where the reference system is either genome sequence coordinates (e.g. genome annotation databases such as UCSC), or names of knowledge network nodes such as genes, proteins, RNA transcripts, transcription factors and their binding sites, v) mapping significant knowledge bits back to the genome sequence, to influence the design of aCGH microarrays, with DNA probes printed on which are sensitive to selected genomic aberrations, and which have good signal-to-noise characteristics. The outline of this dissertation aligned with the aforementioned cycle is sketched on Figure 1.

In Chapter 2 we acquaint the reader with the technology of DNA
Figure 1: Plotted is the outline of this dissertation. We aspire to follow the main steps in the process of aCGH design, research, and clinical validation. A more detailed diagram of aCGH design process can be found in Chapter 2.

We analyze characteristics of log₂ ratio data from aCGH microarrays, quantify its heteroscedasticity and signal-to-noise ratio on real data, to later introduce log₂ ratio segmentation problem, and the most popular Circular Binary Segmentation approach. The section is closed with description and a diagram of the iterative process of targeted microarray design.
Figure 2: Factor graph for $\mathcal{L}_{\text{BSMF}}(\theta; Z, x, \Omega)$ model. Log2 ratio data $\vec{x}$ marked gray. Gaussian Markov random fields $\vec{a}, \vec{b}$, defined on graphs $G_{\text{spatial}}, G_{\text{genome}}$ accordingly, in white. Precision latent variables $\tau, \rho, \nu$ marked blue, their prior Gamma distributions not charted on the graph, Normal prior for $\vec{b}$ not charted. Breaks in the fields indicator $[0, 1]$ variables $Z = \{\vec{y}, \vec{z}, \vec{s}\}$ marked green. Background, and segment fields break probabilities $p, q$ marked in yellow. The $\vec{s}$ is a segment category state indicator, which controls Gaussian mixtures which are priors to segment field $\vec{a}$. The multivariate $N_{G_{\text{genome}}}, N_{G_{\text{spatial}}}$ correspond to Gaussian Markov Fields on edges of the respective graphs.

The Background and Segments Markov random fields model (BSMF) for segmentation and spatial denoising is declared in Chapter 3. There, we introduce Bayesian Graphical model with conjugate priors, which is a Markov Random Field defined on two graphs: spatial grid $G_{\text{spatial}}$, and genomic line $G_{\text{genome}}$. 

The total posterior likelihood for this Background-Segment-Markov-Field problem (BSMF), where \( \theta = \{a, b, \tau, \nu, q, p\} \) is a set of latent to be optimized parameters: \( a, b \) – Markov fields, \( \tau, \nu, q, p \) – Gaussian precision parameters, \( q, p \) – probabilities of breaks in the fields; where \( Z = \{y, z\} \) is a set of \([0, 1]\) field break indicator variables, and given a set of constant parameters \( \Omega = \{(\Omega_{r0}, \Omega_{r1}), (\Omega_{q0}, \Omega_{q1}), (\Omega_{p0}, \Omega_{p1}), (\Omega_{q0}, \Omega_{q1}), (\Omega_{p0}, \Omega_{p1}), \tau_b, \Omega_a\} \), is

\[
\mathcal{L}_{\text{BSMF}}(\theta; y, z; x, \Omega) \propto \prod_{\{i,j\}\in E_{\text{spatial}}} \sqrt{\frac{\tau}{2\pi}} y_{ij} \exp \left( -\frac{1}{2} \sum_{\{i,j\}\in E_{\text{spatial}}} y_{ij} \tau (b_i - b_j)^2 \right)
\prod_{\{i,j\}\in E_{\text{spatial}}} (1 - q)^{y_{ij}} q^{1 - y_{ij}}
\prod_{\{i,j\}\in E_{\text{genome}}} \sqrt{\frac{\rho}{2\pi}} z_{ij} \exp \left( -\frac{1}{2} \sum_{\{i,j\}\in E_{\text{genome}}} z_{ij} \rho (a_i - a_j)^2 \right)
\prod_{\{i,j\}\in E_{\text{genome}}} (1 - p)^{z_{ij}} p^{1 - z_{ij}}
\prod_i \sqrt{\frac{\nu}{2\pi}} \exp \left( -\frac{1}{2} \sum_i \nu (x_i - a_i - b_i)^2 \right)
\tau^{\Omega_{r0}-1} \exp (-\tau \Omega_{r1}) q^{\Omega_{q0}-1} (1 - q)^{\Omega_{q1}-1}
\rho^{\Omega_{p0}-1} \exp (-\rho \Omega_{p1}) p^{\Omega_{p0}-1} (1 - p)^{\Omega_{p1}-1}
\nu^{\Omega_{q0}-1} \exp (-\nu \Omega_{q1})
P_{\text{prior a}} \cdot P_{\text{prior b}}
\]  

(1)

The Factor graph for the above likelihood is presented on Figure 2. The problem is framed as a partially Quadratic Programming problem, its posterior conditional distributions are given, Expectation Maximization scheme for its optimization is proposed and implemented. The model is then extended with Hidden Markov Model (HMM) state-like prior mixture for segment field, the double linkage modification of genomic neighborhood graph is analyzed. The BSMF Markov Chain possibility is briefly remarked, then the Expectation Maximization scheme for BSMF implementation is explained. Results of the algorithm on real data from IMID2py database are described and plotted, its performance is compared with CBS results, sensitivity to variability in setting of priors is analyzed.
In Chapter 4 the problem of array design and comparison thereof is taken on. Synthetic data, and modification of real data, with imposed noise is generated. The measure of robustness to noise is proposed for a single DNA probe, and later extended to a whole microarray design resulting with the measure of relative noise-induced discrepancy. Method is parametrized by the segmentation algorithm used to identify aberrations. We implemented the efficient Monte Carlo method for testing noise robustness within CBS procedure. Results on synthetic data and in the optimization of a concrete aCGH design are presented.

In Chapter 5 we propose a novel multiple sample aCGH analysis methodology aiming in rare CNVs detection. The majority of previous approaches dealt with cancer data sets, while we focus on inborn genomic abnormalities identified in a diverse spectrum of diseases in human. We consider a large log₂-ratio matrix $L$, with dimensions $|S| \times |Q|$, and one of its $k$-windows $L^S_Q$ containing data coming from a set of patients $S = \{1, \ldots, n\}$, and from consecutive probes from the set $Q = \{p, \ldots, p + k - 1\}$ (here probe ordering respects probes positions on the reference genome). The transformation of each of $k$ columns into ranks and division of resulting ranks by $|S| + 1$ yields pseudo–ranks matrix $R^S_Q$.

$$ R^s_q = \frac{\text{rank of } L^s_q \text{ in } L^S_q}{|S| + 1}, \quad s \in S, q \in Q $$

Our method for discriminating outliers is based on a statistics computed for each of $n$ patients: mean $L_q$ distance to other rank vectors.

$$ \mu^q(s) = \frac{1}{|S|} \sum_{j \in S} \left( \sum_{l=p}^{p+k-1} \left| R^s_q - R^j_l \right|^q \right)^{\frac{1}{q}}, \quad s \in S, q \in (0, \infty] $$

For the purpose of this work we selected $L_1$ distance measure, both for simplicity and greater robustness than $L_2$. In the dissertation the null distribution for $\mu^1(s)$ is analyzed further and a statistical test for outliers is constructed. Our method is tested on exon targeted V8.1 OLIGO aCGH microarray by analyzing 366 patients affected with developmental delay/intellectual disability, epilepsy, or autism. The proposed algorithm can be applied as a post-processing filtering to any given segmentation method. With the additional information obtained from multiple samples we efficiently detect significant segments corresponding to
rare CNVs responsible for pathogenic changes. The robust statistical framework based on rank statistics applied in our method eliminates the influence of a technical artifact termed in literature as ‘waving’.

In Chapter 6 described are the design and features of IMID2py database used at IMID to gather and analyze aCGH results. Later, we present a semantic extension to our database, namely the results of our Early Adoption project of Apache Internet Knowledge Stack, which involves using the Apache Stanbol software (Auer et al., 2012) to and annotate records in the database with Linked Open Data, and search within it. Uniprot RDF release with Gene Ontology terms, PubMed abstracts, GeneID references is indexed using Stanbol. A concept of a tree of enhancements is introduced, with a set of modules: enhancers, which facilitate certain specific searches within the semantic graph. At the end of the chapter we summarize results obtained by IMID researches with the use of IMID2py database.

**Scientific publications and other published resources**

Results in Chapters 4, 5 stem from the joint work with Tomasz Gambin who published some of these results in his dissertation Gambin (2012). Results in Chapter 3 were obtained in cooperation with Bogusław Kluge, manuscript in preparation, to whom a more detailed thanks are given at the end of the chapter.

Management, perseverance and faith of Ania Gambin, vision, consequence, and vigilance of Paweł Stankiewicz, the atmosphere and the rendition while working with Tomasz Gambin, insights, and proficiency of Bogusław Kluge, and the hard work of all people from IMID made this thesis possible.

Publications and resources coauthored by the author while working on this dissertation are listed below.

Publications referred to in Chapters 2.

Publication referred to in Chapter 4:


Publication referred to in Chapter 5:


Publications referred to in Chapters 6:


Conferences, and other resources referred to in Chapter 6.

Before the time working on this dissertation the author coauthored the following publications.


REFERENCES


