
Geometric Sample Area Correction of an IEF Gel in a Newborn Screening Application Using Image Processing Methods

Master of Science (Technology)
Thesis
University of Turku
Department of Computing
Health Technology
2022
Santeri Mäntyniemi

UNIVERSITY OF TURKU
Department of Computing

SANTERI MÄNTYNIEMI: Geometric Sample Area Correction of an IEF Gel in a
Newborn Screening Application Using Image Processing Methods

Master of Science (Technology)

Thesis, 74 p.

Health Technology

February 2022

Hemoglobinopathies are a group of inherited red blood disorders, that cause either abnormal hemoglobin variants or disrupt the production of normal hemoglobins. The clinical significance of these hemoglobinopathies vary from asymptomatic to fatal. The most clinically significant hemoglobin variant is hemoglobin S (HbS), causing a disorder known as sickle cell disease, with approximately 300 000 newborns affected each year, especially in Sub-Saharan Africa.

Isoelectric focusing (IEF) is one of the most used methods for screening hemoglobinopathies in newborns. IEF uses a pH gradient to separate the hemoglobin molecules, based on their isoelectric point, into sharp and distinct bands. These bands are then compared to control samples to identify different hemoglobin variants. Sometimes the IEF gel has distortions, caused by sources relating to the manufacturing of the gel or the laboratory environment, which hinder the interpretation of the gel. The goal of this thesis is to propose a method for straightening images of these IEF gels using image processing methods.

Images of real IEF gels are used to develop the method, which is then evaluated with three different tests. The tests include performance, band quantitation, and stability tests. The real images dataset is used for the evaluation in tests one and two, while artificial worst-case images are produced for test three. The results of these tests demonstrate the capabilities and limits of the proposed method.

The proposed method performed well with the dataset containing real images. Out of the 142 images in the dataset, the method managed to straighten 139 of them successfully. In the second test, it was noticed that the proposed method retained the relative band quantities accurately. These results show that the straightening works as intended and can be used as a helpful tool when identifying different hemoglobinopathies in IEF gels. The method could also be developed further to work with images of other protein gel electrophoresis as well, as long as the specifications for the method are defined again accordingly.

Keywords: hemoglobinopathy, IEF, geometric correction, image processing

Contents

1	Introduction	2
2	Background	7
2.1	Hemoglobinopathies	7
2.1.1	Thalassemias	8
2.1.2	Hemoglobin variants	12
2.2	Methodologies for hemoglobinopathy screening in newborns	17
2.2.1	Isoelectric focusing	17
2.2.2	Other methods	21
2.3	Gel image processing	24
3	Materials and methods	26
3.1	Dataset	26
3.2	Geometric correction method	28
3.2.1	Straightening of bands	28
3.2.2	Alignment of samples	34
3.3	Image correction metrics	47
3.4	Evaluation process	49
3.4.1	Performance testing with real images	49
3.4.2	Comparing the relative band quantities	50
3.4.3	Stability test with synthetic images	50

3.5	Implementation	51
4	Results	52
4.1	Performance test	52
4.2	Comparison of band quantities	55
4.3	Stability test	58
5	Discussion	61
5.1	Performance test	61
5.2	Comparison of band quantities	68
5.3	Stability test	69
6	Conclusions	73
	References	75

Abbreviations

Abbreviation	Term
1D	One-dimensional
2D	Two-dimensional
BGE	Background electrolyte
CZE	Capillary zone electrophoresis
Hb	Hemoglobin
HPFH	Hereditary persistence of fetal hemoglobin
HPLC	High-performance liquid chromatography
IEF	Isoelectric focusing
MCHC	Mean cell hemoglobin concentration
NGS	Next-generation sequencing
SCD	Sickle cell disease
UN	United Nations
UV	Ultraviolet
WHO	World Health Organization

1 Introduction

Hemoglobinopathies are a group of inherited genetic red blood cell disorders, that affect either the production of normal hemoglobins (thalassemias) or result in production of hemoglobin variants [1]. Most notable hemoglobin variant is hemoglobin S (HbS). The presence of HbS is the cause of multiple disorders, forming together a group called “sickle cell disease” (SCD). This causes normally round red blood cells to distort into a crescent, sickle-like shape, hence the name sickle cell. SCD includes all disorders caused by the occurrence of HbS, such as SCD-SS, SCD-SC and HbS combined with beta-thalassemias. These are explained more thoroughly in Chapter 2.

The clinical significance of SCD and other hemoglobinopathies vary from asymptomatic to fatal. Newborn screening is crucial to detect these disorders, since 3 to 6 months after birth some SCDs start to cause serious danger, if not treated [2]. This is because fetal hemoglobin (HbF) protects the newborn from symptoms caused by the HbS. HbF is the main hemoglobin found in fetuses. It has higher oxygen affinity than adult hemoglobin (HbA), allowing the transfer of oxygen from maternal to the fetal circulation [3]. The production of HbF starts at 6 weeks of pregnancy and it remains as the dominant hemoglobin until three months after birth. After that, the level of HbF starts slowly decreasing and the level of HbA starts to increase. A normal healthy adult has 95%-97% of HbA and under 1% of HbF [4]. If the person has normal levels of HbA, then the carrying of the HbS usually causes no symptoms. This is referred as having sickle cell trait (SCD-AS), and only affects

the possible descendants of the carrier. However, if the newborn has issues with the production of HbA, then depending on the type of the SCD, it may lead to serious issues, such as acute splenic sequestration, overwhelming septicemias, and acute chest syndrome [2]. If the underlying SCD is detected early, the patient can be treated, for example, with blood transfusion.

Most common method for screening hemoglobinopathies is isoelectric focusing (IEF), because of its high resolution, inexpensiveness, ease-of-operation and rapid turnaround time [5]. IEF is a form of gel electrophoresis where the hemoglobins are separated into different bands based on their isoelectric point using a pH gradient [6]. A trained expert then compares the locations of these bands to a control samples' bands locations, and based on that, makes decisions which hemoglobin variant it is. A common control sample used is AFSC, which corresponds to a sample containing HbA, HbF, HbS and HbC, where the order describes their concentration from highest to lowest. For example, HbS can be detected in a patient sample by locating a band in the same lane as the control sample's HbS-band. IEF can also be used to detect quantitative disorders in normal hemoglobins by examining the intensity of the bands. A band with greater intensity has higher amounts of that hemoglobin variant.

In a screening application, an IEF gel commonly has two sides, A and B. Typically, control samples are added at the start, middle and at the end of each gel side. This is done, so that the gel side would contain enough control samples for the comparison to be feasible between a patient sample and its nearest control sample. A gel side can be divided further into two gel quarters, from a control sample to another. Thus, the IEF gel consists of two gel sides and four gel quarters. Figure 1.1 shows an example of an IEF gel quarter with AFSC-control samples. In this example, all except one of the patient samples are normal newborn with highest quantity of HbF and some amount of HbA. The one exception is the next to last patient sample,

which has HbS present. The band above HbA in each patient sample is normal and results from partial acetylation of HbF.

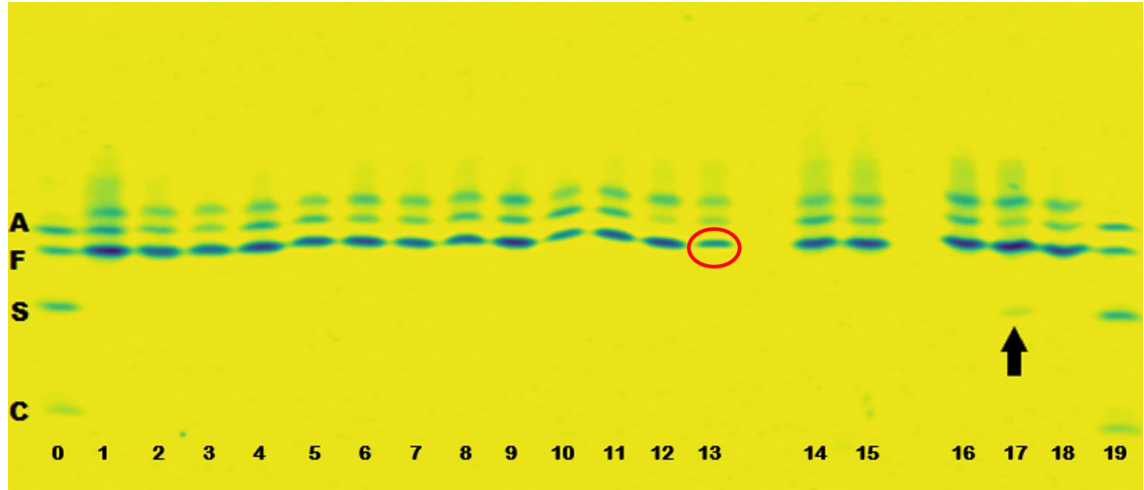


Figure 1.1: Image of an IEF gel quarter with patient samples (samples 1 to 18) between two AFSC-control samples (samples 0 and 19) at the start and at the end. Red circle shows a single band. Arrow points to the sample 17, which contains HbS.

A perfect IEF gel would have straight lanes and bands to enable easy comparison of the patient samples to control samples, but occasionally the gel has distortions. These can be caused by multiple reasons relating to the manufacturing and running the gel or the laboratory environment. These distortions can be divided into two separate computational issues. The first issue is the bending of a single sample's bands, known as the “smiling”-effect, depicted in Figure 1.2. All the bands within a single sample are curved the same way.

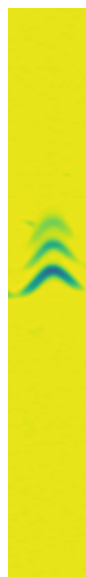


Figure 1.2: Image of a single sample with curved bands

The second issue is the bending of the whole sample area. These lane distortions make "waves" in the gel and can make the gel uninterpretable for the expert. Figure 1.3 shows an example of this lane distortion on a gel side. The goal of this thesis is to find ways to solve these two issues and straighten the image of the gel by using image processing methods.

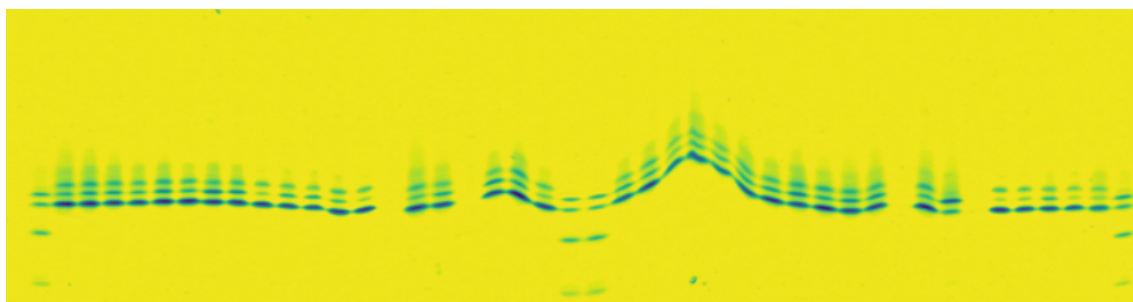


Figure 1.3: Image of an IEF gel side with lane distortions

This thesis was done for PerkinElmer Turku site (Wallac Oy). PerkinElmer is a global corporation that provides solutions to diagnostics, life sciences, food and applied markets [7]. They are the global market leader in newborn screening, providing complete systems, such as newborn screening kits, consumables, instruments and software. Wallac Oy is one of PerkinElmer's largest manufacturing and research and development facilities with nearly 600 employees [8]. All images of the gels in this thesis are provided by PerkinElmer. The gels have been created with their instrument Migele™ Gel Electrophoresis Unit (see Figure 1.4) [9]. Migele™ Gel Electrophoresis Unit was developed by Wallac Oy and published in 2018. It is used to separate proteins, in this case hemoglobins, with IEF.

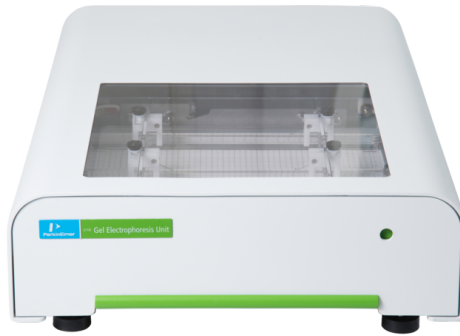


Figure 1.4: Migele™ Gel Electrophoresis Unit [9]

Structure of the thesis is following. In this introduction chapter a context for the thesis is given and a small overview of the problem and the proposed method are stated. Chapter 2 is a background chapter that focuses on giving sufficient theoretical knowledge to the topic. Chapter 3 contains information about the used data and a deeper dive into the proposed method. Chapter 4 presents the results of this thesis and Chapter 5 the discussion about the results. Chapter 6 is the conclusion chapter.

2 Background

2.1 Hemoglobinopathies

Hemoglobin is a respiratory protein located in red blood cells whose main purpose is to transport oxygen from the lungs into tissues and transport carbon dioxide back to the lungs [10]. It consists of four subunits, each having one protein chain (globin) and one heme group. Normal adult hemoglobin (HbA) consists of two alpha (α) chains and two beta (β) chains [4]. Other globulins include delta (δ) chain in normal adults and gamma (γ) chain, epsilon (ϵ) chain, and zeta (ζ) chain in embryonic and fetal parts of life. These last three chains form Hb Grower and Hb Portland in the embryonic stage but are replaced with fetal hemoglobin (HbF) before birth. HbF is replaced with HbA and HbA2 over the first year of life. Table 2.1 shows the normal hemoglobins found at birth and in adults [11].

Table 2.1: Normal hemoglobins [11]

Hemoglobin	Structure	Levels at birth	Levels in adults
F	Two alpha, two gamma	70%-80%	<1%
A2	Two alpha, two delta	0.5%	2.5%
A	Two alpha, two beta	20%-25%	97%

Hemoglobin disorders are called hemoglobinopathies. It is estimated that 7% of world population carry some form of hemoglobinopathy [12] [13]. Over 1000 different hemoglobinopathies have been detected, most of them being asymptomatic, but

some of them can be fatal [4]. These disorders are passed down through families in an autosomal recessive inheritance. This means that two copies of the mutated gene are needed for it to cause issues. Carrying one mutated gene and one healthy gene usually causes no health issues, these people are simply referred to as carriers. Having two different forms of the same gene is referred to as heterozygous, whereas homozygous is referred to as having the same mutated gene from both parents.

As mentioned at the beginning of Chapter 1, hemoglobinopathies can be divided into two categories: quantitative disorders in hemoglobin synthesis and qualitative disorders in hemoglobin structure [4]. In the following Sections 2.1.1 and 2.1.2, the most clinically significant hemoglobinopathies are introduced in both categories.

2.1.1 Thalasseмии

Thalasseмии are disorders in the synthesis of the globin chains. The affected globin chain can be any of the presented at the beginning of Chapter 2.1, but the most clinically significant ones are α -thalassemia and β -thalassemia [13].

Alpha-thalassemia

Alpha-thalassemia is one of the most common hemoglobinopathies, occurring mainly in Africa, Arab countries and South-East Asia [13]. Four genes are responsible for production of alpha-globin, two from each parent ($\alpha\alpha/\alpha\alpha$) [4]. Alpha-thalassemia results from reduced (α^+) or absent (α^0) functionality of one or more of these four genes [14]. Four categories of alpha-thalassemia can be formed, with different severity levels, based on how many genes are affected.

If only one of the four genes is affected ($-\alpha/\alpha\alpha$), it is referred to as the silent carrier [14] or heterozygous α^+ -thalassemia (α -thalassemia minima) [13]. This type of alpha-thalassemia usually causes no symptoms.

Alpha-thalassemia trait is characterized by two dysfunctional alpha genes [14].

It is also referred as α -thalassemia minor and it can be either heterozygous α^0 -thalassemia ($--/\alpha\alpha$) or homozygous α^+ -thalassemia ($-\alpha/-\alpha$), depending on which two genes it affects [13]. Alpha-thalassemia trait causes mild anemia.

When there are three genes affected ($--/-\alpha$), it is called hemoglobin H disease. This disorder causes significant health issues, including microcytic and hypochromic hemolytic anemia [14]. When the body has issues with the production of alpha-globin chains, it tries to compensate that by producing more beta-, gamma-, and delta-globin chains. Four beta-globin chains can form a hemoglobin variant called HbH and four gamma-globin chains can form another variant called Hb Bart's. The name of the disease comes from the presence of HbH, since the production of the beta-globin chains is increased. In newborns, while the production of gamma-globin chains is still active, Hb Bart's might also be present, but after beta-globin chain production takes over, HbH is developed [4].

The most severe form of alpha-thalassemia is called Hb Bart's hydrops fetalis syndrome [14]. In this case, the fetus cannot produce any alpha-globin chains to produce the normal hemoglobins, HbF or HbA. This causes the fetus' blood to contain mainly Hb Bart's (four gamma-globin chains). This syndrome is not compatible with life and these fetuses are usually stillborn or die shortly after birth.

The amount of Hb Bart's in different alpha-thalassemias is 1%-2% [14] in silent carriers, 5%-6% [14] in alpha-thalassemia minor and 80%-90% [13] in Hb Bart hydrops fetalis syndrome. In HbH disease, depending on the age of the patient, either HbH or Hb Bart's can be detected. Hb Bart's hydrops fetalis syndrome can be detected with prenatal screening and the early termination of such pregnancies is recommended. This could be avoided all along with the screening of the parents, because the disease is typically inherited from both parents having the alpha-thalassemia trait. Different alpha-thalassemias can be detected with IEF from the presence of these two abnormal hemoglobin variants. In IEF, higher levels of a cer-

tain hemoglobin variant is demonstrated as a band with higher intensity, in other words as a darker band. The different alpha-thalassemias are summarized in Table 2.2 [13].

Table 2.2: Alpha-thalassemias [13]

Condition	Genotypes	Symptoms
Normal	$(\alpha\alpha/\alpha\alpha)$	No symptoms
Alpha-thalassemia silent carrier	$(-\alpha/\alpha\alpha)$	No symptoms, small changes to red blood cell count
Alpha-thalassemia minor	$(-\alpha/-\alpha)$ or $(--/\alpha\alpha)$	Mild anemia, significant changes to red blood cell count
HbH disease	$(--/-\alpha)$	Variable chronic hemolytic anemia
Hb Bart's hydrops fetalis	$(--/--)$	Life-threatening fetal anemia

Beta-thalassemia

Whereas alpha-thalassemia results from problems in the production of the alpha-globin chain, beta-thalassemia results from partial (β^+) or missing (β^0) beta-globin chain production [15]. It is mostly seen in Mediterranean countries, South-East Europe, Arab nations, and Asia [13]. Two genes are responsible for the production of beta-globin, one from each parent. Beta-thalassemias can be divided into three categories based on the number of genes affected, and whether the affected genes are functioning partially or are completely absent. The three forms of beta-thalassemia from the least severe to the most severe are β -thalassemia minor, β -thalassemia intermediate, and β -thalassemia major [15].

The least severe form of beta-thalassemia, β -thalassemia minor, is also referred to as beta-thalassemia carrier and beta-thalassemia trait [15]. This form of beta-thalassemia is typically asymptomatic but might cause mild anemia. Subjects suffering from this are typically heterozygous, meaning only one gene is affected. Either

the production is partial (β^+/β) or completely absent (β^0/β).

Both β -thalassemia intermediate and β -thalassemia major have both genes as defective, but in the intermediate form there is still some production of beta-globin, whereas in the major form there is no synthesis of beta-globin [15]. In β -thalassemia intermediate the subject is usually mild homozygous (β^+/β^+) or compound heterozygous (β^+/β^0) and in β -thalassemia major they can be homozygous (β^0/β^0) [13].

Having β^+ or β^0 does not tell whether the person has the intermediate or the major form of beta-thalassemia. The difference comes from the severity of the symptoms. In β -thalassemia major, the clinical presentation occurs between 6 and 24 months [15]. The symptoms include severe anemia and multiple other severe diseases. A common sign of β -thalassemia major is also the enlargement of spleen and liver. The patient requires lifelong blood transfusions and, if optimally treated, has a life expectancy of 50 to 60 years [13].

The severity of β -thalassemia intermediate varies. In the most severe form of β -thalassemia intermediate, the symptoms occur later than in the major form, around 2 to 6 years [15]. Patients have mild anemia and only occasionally require blood transfusions, but their growth and development are hindered. Milder forms of β -thalassemia intermediate can be asymptomatic to adult life.

Similar to alpha-thalassemia, where the production of other globin chains is increased due to the lack of alpha-globin chain production, beta-thalassemia results in increased production of alpha-, gamma-, and delta-globin chains [4]. This results in higher levels of HbA2 (two alpha-globin chains and two delta-globin chains) and HbF (two alpha-globin chains and two gamma-globin chains). The higher levels of HbA2 and HbF can be used to identify beta-thalassemias from IEF gels, as they appear as darker bands in beta-thalassemia patients than in normal patient samples. Table 2.3 summarizes the different forms of beta-thalassemia [13].

Table 2.3: Beta-thalassemias [13]

Condition	Genotypes	Symptoms
Normal	(β/β)	No symptoms
Beta-thalassemia minor	(β^+/β) or (β^0/β)	Mild anemia
Beta-thalassemia intermediate	(β^+/β^0) or (β^+/β^+)	Moderate disease Variable transfusion dependency
Beta-thalassemia major	(β^+/β^0) or (β^0/β^0)	Severe illness with long-term transfusion-dependent anemia

2.1.2 Hemoglobin variants

The other category of hemoglobinopathies is the different hemoglobin variants caused by disorders in the hemoglobin structure. There are over 1000 different hemoglobin variants detected and new ones are detected all the time [16]. Some of the most common and severe variants are presented in this chapter. Most common variants worldwide include HbS, HbC and HbE [12] [13]. Previously, different hemoglobin variants were only prevalent in malaria-endemic regions (Mediterranean, Asia, and Sub-Saharan Africa), but due to migration, the different hemoglobin variants have spread all over the world [4]. Conventionally, new variants are named after the geographic origin of the affected person [16]. For example, HbD-Punjab is prevalent in the Punjab region of Northwest India [17]. These different variants can easily be detected with IEF, as they appear as separate bands in the IEF gel compared to the normal hemoglobin types (HbA, HbF, HbA2). Figure 2.1 shows a chart that reflects the relative placement of some hemoglobin variants compared to normal hemoglobins in an IEF gel [18].

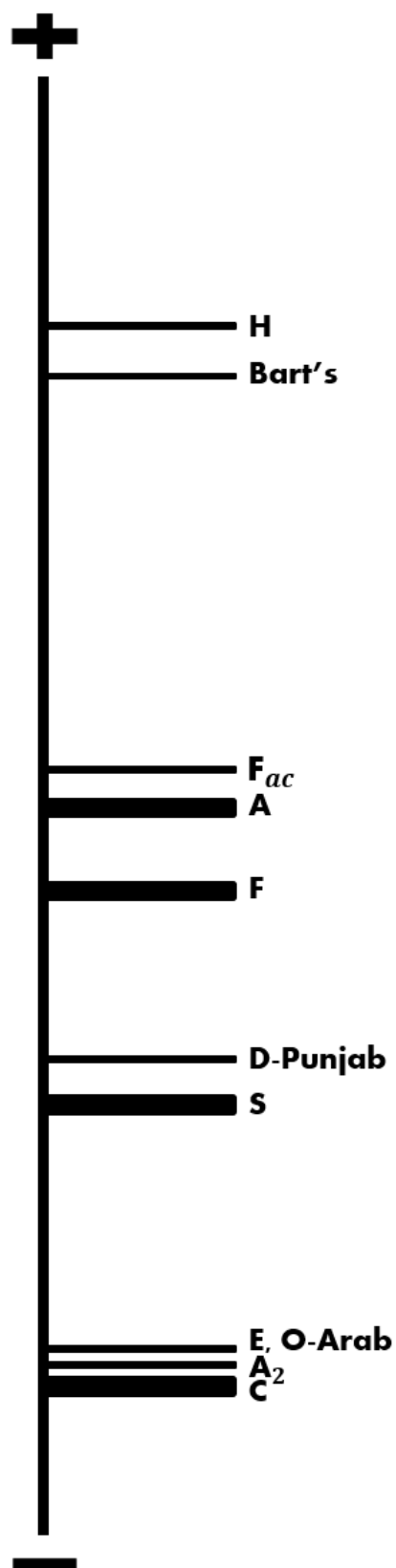


Figure 2.1: Chart of relative placements of some hemoglobin variants to normal hemoglobins in an IEF gel (adapted from [18])

Sickle cell disease

Sickle cell disease, including all occurrences of HbS variant, is the most frequent and clinically significant hemoglobin disorder. World Health Organization (WHO) and United Nations (UN) have declared sickle cell disease as a global public health problem [19]. Approximately 300 000 newborns are affected by sickle cell disease every year, with 75% of them born in Sub-Saharan Africa [20]. High frequencies of sickle cell disease in areas with malaria have been linked to the protective effect of sickle cell trait towards severe malaria [21]. Unfortunately, these asymptomatic carriers of sickle cell trait might pass it forward to their children, who might get the dangerous homozygous form of sickle cell disease SCD-SS, if both of the child's parents have sickle cell trait. Figure 2.2 represents the inheritance of sickle cell disease from two parents having sickle cell trait. SCD-SS have also been referred to as sickle cell anemia, but today's nomenclature suggests not to use that term, since the increased awareness of the symptoms have revealed that the dominant symptoms are vascular obliterations in organs, rather than anemia [13] [22].

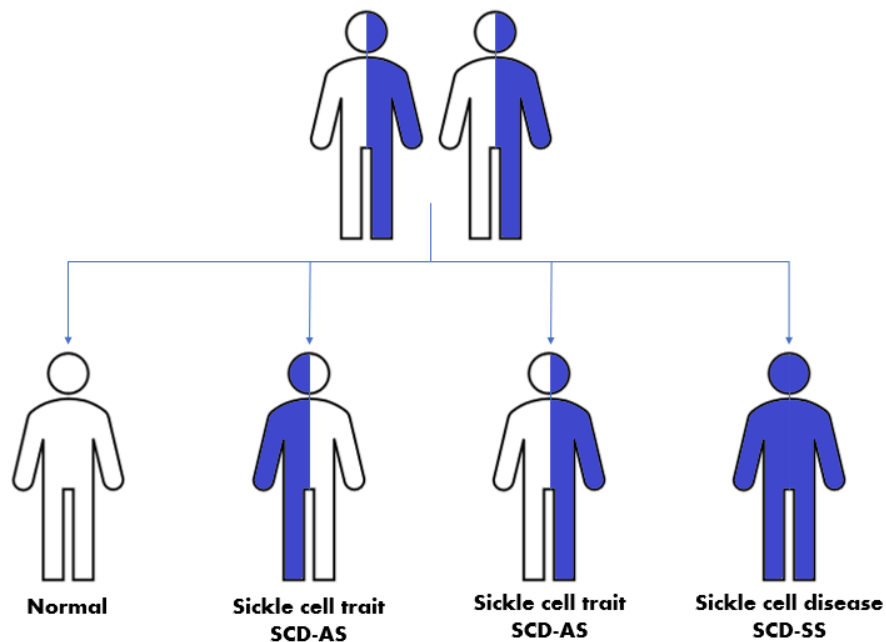


Figure 2.2: Representation of sickle cell inheritance

HbS consists of two alpha-globin chains and two mutated beta-globin chains. The mutation is caused by single amino acid substitution in the beta-globin gene [23]. In addition to homozygous sickle cell disease (SCD-SS) and sickle cell trait (SCD-AS), there is a wide range of other mixed heterozygous SCDs such as SCD-SC, SCD-SD, SCD- β^+ -thalassemia and SCD- β^0 -thalassemia [4].

In SCD-SC, the patient is compound heterozygous for both HbS and a variant HbC. SCD-SC results in lower frequencies of acute crises, higher life-expectancy, and overall weaker symptoms of sickle cell disease. Hemoglobin S combined with beta-thalassemia, on the other hand, can be as severe as SCD-SS, depending on the type of thalassemia. In SCD- β^+ -thalassemia, HbA is still produced and that results in milder form of sickle cell disease. In SCD- β^0 -thalassemia, similarly to SCD-SS, there is no HbA present, and it is very dangerous. Other rarer heterozygous mixes with HbS include SCD-S/D-Punjab and SCD-S/O-Arab. These have similar severe symptoms as SCD-SS and SCD- β^0 -thalassemia. [23]

There are also multiple factors that affect the severity of sickle cell disease [2]. One is the high levels of HbF. As mentioned in Chapter 1, HbF protects the newborn from symptoms caused by HbS, until the level of HbF starts normally decreasing after around 3 to 6 months after birth. For some people, the high levels of HbF persist naturally into adulthood. This condition is referred to as hereditary persistence of fetal hemoglobin (HPFH) [4]. HPFH combined with HbS results in less severe or even completely absent SCD-related complications. It also protects from symptoms caused by thalassemias.

Alpha-thalassemia has also been noticed to mitigate the severity of SCD [2]. Heterozygous α^+ -thalassemia has been detected to lower the mean cell hemoglobin concentration (MCHC) in SCD-SS, which results in less sickling of the red blood cells. Homozygous α^+ -thalassemia lowers the MCHC even more, but any consequent sickling reduction may be offset by an increase in total hemoglobin.

Other variants

After HbS, the most common hemoglobin variants are HbC and HbE. As mentioned in previous chapter, HbC can be present together with HbS where it will actually result in milder symptoms of SCD. HbC can also occur without HbS. Heterozygous HbC, or HbC trait, causes no issues and is usually asymptomatic, but homozygous HbC, or HbC disease, results in nearly 100% quantity of HbC in hemoglobin and results in severe chronic hemolytic anemia and pain crisis [4].

HbE resembles beta-thalassemia, in the sense that it causes deficiency in the production of beta-globin chains [4]. It is very common in South-East Asia [13]. Heterozygous HbE, or HbE trait, results in symptoms similar to beta-thalassemia minor. Homozygous HbE, or HbE disease, results in moderate anemia. HbE can be combined with β^0 -thalassemia, which will cause severe anemia and other symptoms similar to beta-thalassemia intermediate.

Rarer, but still severe hemoglobin variants include HbD-Punjab and HbO-Arab. HbD-Punjab is a sub-type of HbD and it is also referred in literature as HbD-Los Angeles [17]. It has a frequency of 2% in the population of the Punjab region in India. It is also common in Italy, Belgium, Austria and Turkey. Heterozygous HbD-Punjab is asymptomatic and in very rare homozygous form HbD-Punjab causes mild anemia. HbD-Punjab becomes severe when present together with HbS or beta-thalassemia.

HbO-Arab is a variant mostly prevalent in Balkans, the Middle East and Africa [24]. The heterozygous form of HbO-Arab results in mild anemia and symptoms similar to beta-thalassemia minor. Again, as with most of the variants already mentioned, HbO-Arab becomes severe when combined with HbS or beta-thalassemia. Homozygous HbO-Arab is extremely rare and there are only a few known cases.

2.2 Methodologies for hemoglobinopathy screening in newborns

Traditionally, hemoglobinopathies are identified using gel electrophoresis [11]. In today's newborn screening applications, gel electrophoresis has been replaced with more robust options. The two most used methods are isoelectric focusing (IEF) and high-performance liquid chromatography (HPLC) [1] [11]. In the following sections, these methods are introduced. IEF is presented more thoroughly than others in Section 2.2.1, since it is the method that Migele™ Gel Electrophoresis Unit uses and therefore is the most relevant to this thesis. Section 2.2.2 is dedicated to a brief review of HPLC and other methods used in newborn screening for hemoglobinopathies.

2.2.1 Isoelectric focusing

Isoelectric focusing (IEF) is a form of gel electrophoresis. Electrophoresis is a general term referring to methods that use an applied electrical field to separate charged molecules [25]. This method was first introduced by Arne Tiselius in 1937 [26]. Typical electrophoresis system consists of two electrodes, anode and cathode, that have opposite charges and a supporting medium known as background electrolyte (BGE) [6]. In gel electrophoresis, the support medium is gel, commonly composed of either polyacrylamide or agarose [25].

IEF uses a pH gradient to separate the hemoglobin molecules to their isoelectric point [27]. Isoelectric point is the level of pH where the net charge of the hemoglobin is zero and it stops migrating. IEF uses a stable pH gradient, compared to the traditional method of using a single and a constant pH value [28]. The pH gradient, provided by carrier ampholytes, is lower at the anodic side and higher at the cathodic side. During IEF, the hemoglobin molecule runs along the pH gradient, with changing velocity, until it reaches its isoelectric point and stops. The electrical field

will keep the molecules at their respective isoelectric points, where the molecules form a sharp distinct band. If the molecules start to migrate away, they gain net charge again, and are forced back to their isoelectric point. When the device is shut down and the electrical field removed, the molecules start to diffuse again.

The key difference between IEF and traditional gel electrophoresis is that in IEF there is a clear end point for the molecules, which is the isoelectric point, whereas in traditional gel electrophoresis, there is no end point, and the potential difference has to be ceased before the molecules diffuse into the anodic compartment [28]. Figure 2.3 depicts a comparison between the conventional gel electrophoresis and IEF [28]. In the figure, a mixture of two proteins A and B is presented, with isoelectric points of 8 and 5, respectively. Titration curves of the proteins are shown on the right of each drawing. With traditional gel electrophoresis, the two proteins move with constant net negative charge and constant velocity to a specified point (pH=9) of their titration curve. This leads to a net charge of -3 for protein A and -1 for protein B. There is no force in this system that prevents diffusion, and as mentioned above the system has to be stopped in time, before the proteins are eluded into the anodic compartment. In IEF, on the other hand, the proteins are continuously forced to their isoelectric point, which results in very sharp and focused bands. The isoelectric point is the point in the titration curve, where the net charge is zero.

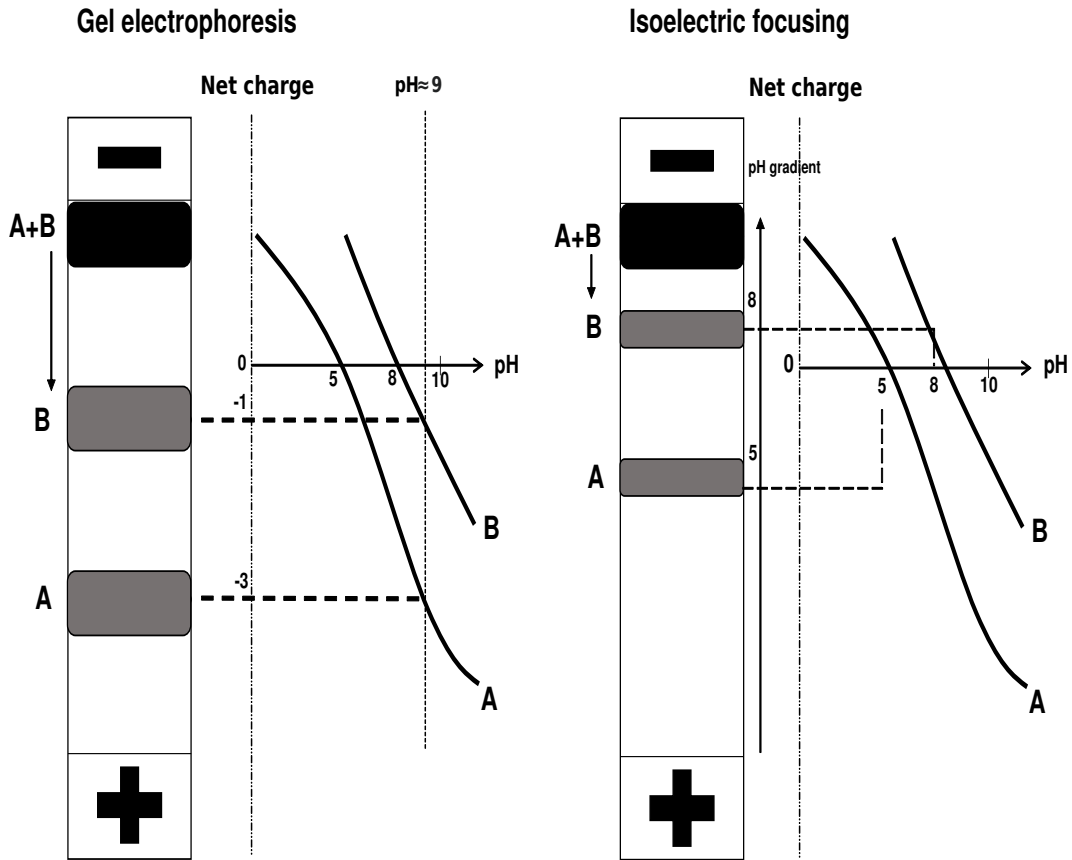


Figure 2.3: Comparison between a traditional gel electrophoresis and isoelectric focusing (IEF). (adapted from [28])

There are multiple reasons why IEF is often used as the initial screening test for hemoglobinopathies. First is the capability of using IEF to simultaneously screen large amounts of samples. For example, MigeleTM Gel Electrophoresis Unit can screen up to 90 samples at the same time, by using a large gel with two sides and a setup containing one cathode and two anodes. Figure 2.4 shows this kind of IEF setup with two anodes and one cathode. The gel has two sides, A and B, and the samples are inserted to the cathodic end of both sides, where they will migrate towards the anolyte. IEF is also very cost-effective since it can run high volumes of samples with low costs [29].

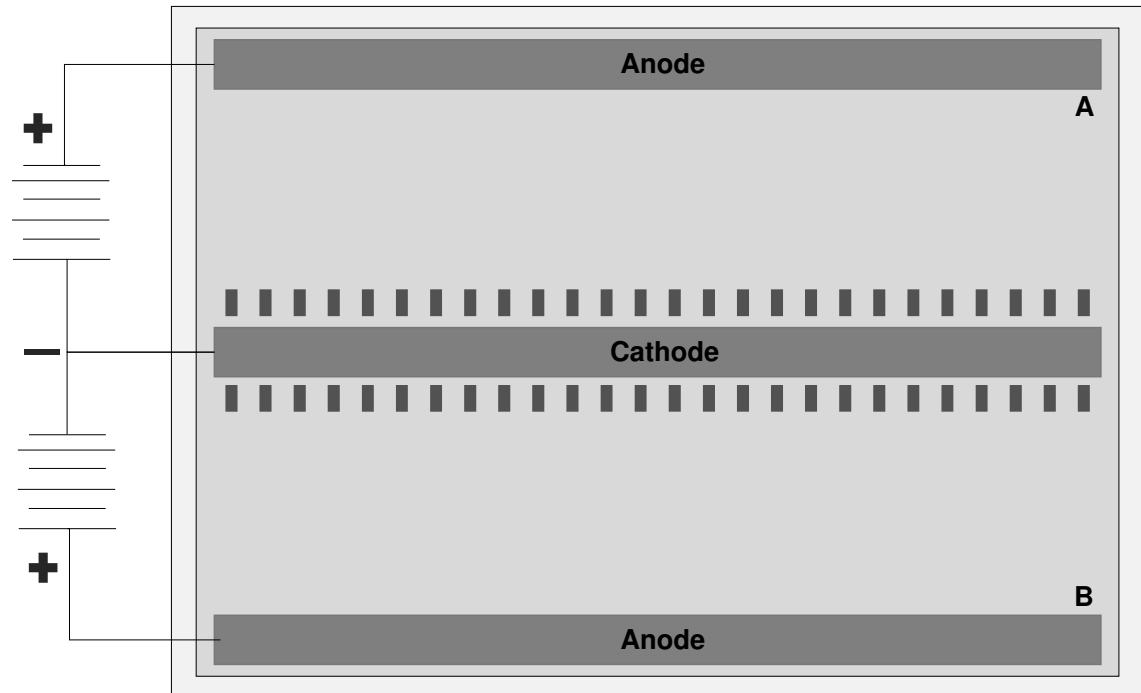


Figure 2.4: Typical IEF setup with two anodes and one cathode

Second reason is the relatively low run time. Depending on the size of the gel, IEF can be run within 60 to 90 minutes, excluding sample preparation [9]. Time efficiency can also be improved by running multiple IEF units simultaneously. Up to four Migele™ Gel Electrophoresis Units can be used with one circulating water bath. They can also be stacked on top of each other to save space.

Lastly, the sharp bands created with IEF are easy to interpret. One can quickly detect any abnormalities when comparing the samples to a normal newborn sample or a control sample. Also, changes in the intensity of normal hemoglobin bands can be detected, such as the absence of HbA band. However, some hemoglobin variants will separate into the same or almost the same locations, which will make confident identifying of the variant difficult. This is why some alternative method, such as HPLC, is needed to confirm the result.

2.2.2 Other methods

Other methods for hemoglobin separation and hemoglobinopathy detection include traditional gel electrophoresis with alkaline or acid gels, HPLC and capillary zone electrophoresis (CZE) [30]. One viable method to utilize in the future is next-generation sequencing (NGS), which is used for large-scale DNA screening [31]. The costs of this method are still too high to be used for extensive screening in endemic developing countries, but in the future, NGS might be used largely to diagnose hemoglobinopathies.

The separation process of traditional gel electrophoresis was already reviewed in Section 2.2.1. Similar to IEF, some hemoglobin variants will migrate to the same zone with this method, but they can be further identified by using a two-tiered electrophoresis protocol [30]. This includes first using a constant alkaline pH, for example, with cellulose acetate ($\text{pH} \approx 8.6$), and then using a constant acidic pH, for example, with citrate agar ($\text{pH} \approx 6.0$). The traditional gel electrophoresis is cheaper than IEF, but very time-consuming and labor-intensive.

Like IEF, HPLC is based on the net charge of the molecule at a specific pH [30]. In HPLC, the hemoglobin is eluted by buffers with a pH gradient [29]. Retention time is the time between injection of the sample into the column and the elution. The hemoglobin is quantified by integrating the area under the curve of the produced chromatogram. Figure 2.5 represents a chromatogram of HPLC, where x-axis represents time in minutes and the y-axis represents the response in volts [29]. Together, with the information about the retention time and the relative quantification, the automated systems in HPLC can identify the hemoglobin variant. The advantages of HPLC are that it is fully automated and can identify the hemoglobin very accurately. However, compared to IEF, HPLC is very expensive, requires technical expertise to calibrate, maintain and operate, and is highly sensitive to bad samples [30].

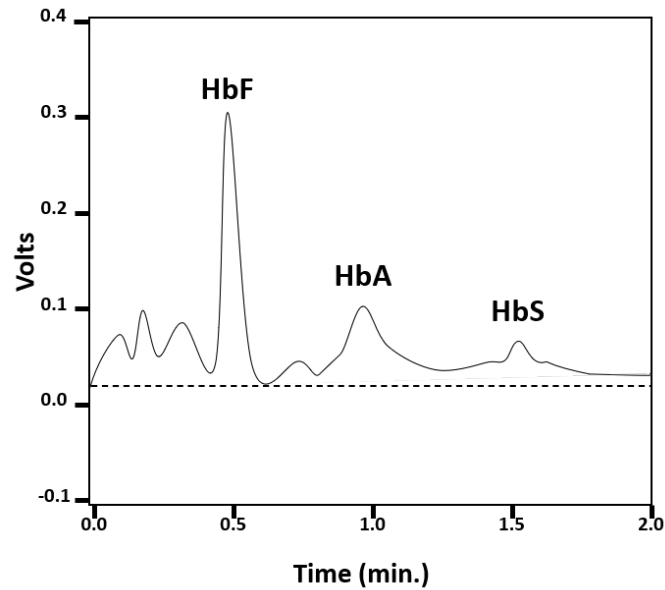


Figure 2.5: Scheme of a HPLC chromatogram representing a FAS-pattern (HbF/HbA/HbS) [29]

Capillary zone electrophoresis (CZE) combines traditional electrophoresis with chromatography [30]. Instead of a gel support medium, alkaline buffers are used to separate the hemoglobin molecules in small silica glass tubes (capillaries), based on the pH of the buffer. The molecules migrate toward the cathode inside the capillary via electro-osmotic flow and the optical density is detected by ultraviolet (UV) system. Much like HPLC, CZE calculates the time between the injection of the sample and the detection of a specific molecule. CZE is also automatic, and it can identify the variant from the defined migration zones and by the integration of the peaks [29]. In neonatal systems, there are 13 different zones, N1 to N13, where HbF migrates to N7 and HbA to N10. Similar to HPLC, CZE is expensive, and it requires high level of expertise for sample processing and interpretation [30].

Because of the difficulty of identifying a hemoglobin variant with certainty, multiple screening methods are often used together to confirm the results. This also solves the issue with protein-based separation methods overlapping hemoglobin variants in

particular zones [30]. Figure 2.6 shows a model created by Association of Public Health Laboratories that depicts the basic hemoglobinopathy testing process [11]. The primary screening method is typically IEF or HPLC, depending on the laboratory. The second screening method depends on the first screening method and on the hemoglobin variant detected.

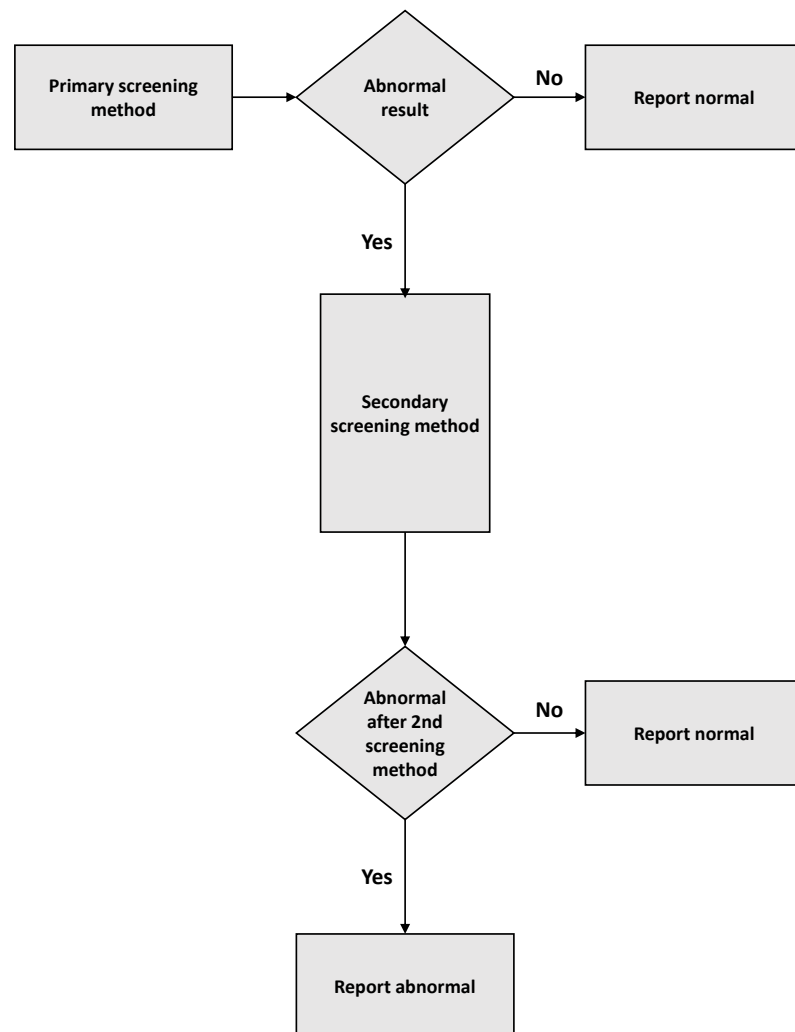


Figure 2.6: Basic hemoglobinopathy testing process [11]

2.3 Gel image processing

In this section, some of the image processing techniques used in the proposed method of this thesis are introduced. In addition to visually inspecting a two-dimensional (2D) image, the image can be represented and analyzed through its one-dimensional (1D) profile. This vector form can be created, for example, by calculating the mean of the pixel intensities in each row of the image. Another way for creating the vector form is by summing all the pixel intensity values of each row. This technique is useful for representing and analyzing single samples, since the vector form can be used to detect the bands' locations. Figure 2.7 shows a control sample and its 1D-signal, that is created by calculating the mean value for each row of the image. In the vector form, the y-axis represents the pixel intensity values between 0 and 1, where 0 represents total white and 1 represents total black. The x-axis represents the y-axis location of the original image. After locating the bands, the relative quantities of the bands can be calculated.

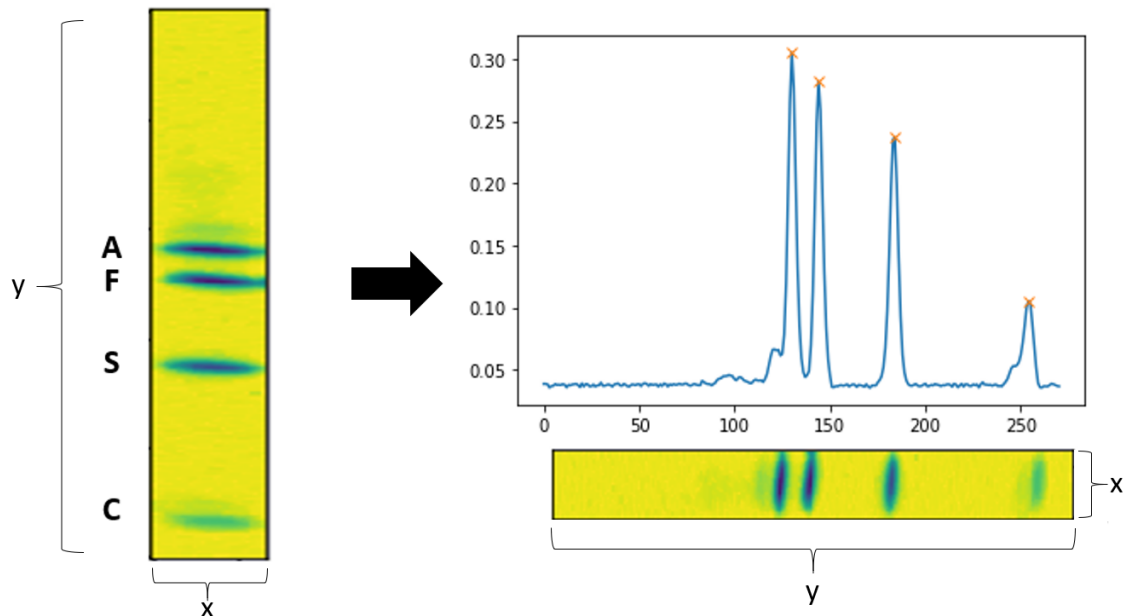


Figure 2.7: 1D-signal of a control sample created by calculating the mean pixel intensity values for each row

It can be assumed that a control sample in an IEF gel image contains four bands representing HbA, HbF, HbS and HbC. These bands can be located by finding the peaks of the signal. The peaks are the local maxima in the signal, found by comparing each point of the signal to its direct neighbors. If the two neighbors have smaller amplitude, then that point is registered as local maxima. In a control sample's case, the four highest local maxima correspond to the four bands in the sample.

After finding the peaks, the peak areas for each peak can be calculated. Peak areas are calculated by multiplying the peak height and the width of the peak at half-height. The relative quantity of a peak is calculated by dividing the peak area by the sum of all the peak areas. Equations (2.1) and (2.2) show the calculations for the percentage quantity of a control sample's band A compared to all the four bands.

$$A_{Area} = h \times w_{\frac{1}{2}} \quad (2.1)$$

$$A_{percentage} = \frac{A_{Area}}{A_{Area} + F_{Area} + S_{Area} + C_{Area}} \times 100 \quad (2.2)$$

In addition to assessing in what areas the bands appear, the relative band quantities are crucial for the interpretation of the samples in IEF gels. The quantity of each hemoglobin variant has to stay the same inside a sample, to avoid any errors in the hemoglobinopathy screening process. For this reason, the above equations are used to evaluate whether the proposed method changes the relative quantities of the bands. The peak areas are calculated from the 1D-signals for each control sample in the original image and then calculated again after the straightening. Then these two values are compared, to ensure that the method does not change the relative quantities.

3 Materials and methods

3.1 Dataset

The dataset used in this thesis consists of images of IEF gel sides that were created with PerkinElmers' MigeleTM Gel Electrophoresis Unit. The images are not publicly available, therefore PerkinElmer provided them for this thesis. The dataset contains images of both gel sides from 71 different large IEF gels, so in total there are 142 images. The images are all grayscale, but are mapped to viridis colormap in figures, because it makes the images easier and faster to interpret [32]. In this colormap, yellow represents white and dark purple black, with different shades of green and blue in between.

There are a few assumptions made about the images. First assumption is that the gel side always contains four AFSC-control samples. One is located at the start of the sample area, two are located at the middle, and one is located at the end of the sample area. The presence of good quality control samples is important for the proposed method, since they are used as the reference locations for the HbA-, HbF-, HbS- and HbC-bands. It is also assumed that the locations for the control samples are known. In the dataset used in this thesis, the majority of the gel sides have 40 samples, with control samples at positions 0, 19, 20 and 39.

Second assumption is that the position of each sample is given in the dataset with the images. This means that the x-coordinates for the starting and the ending

position of each sample are known. The given x-coordinates have been manually chosen for each sample. A gel side can also have empty spots without a sample.

Third assumption is that most of the samples are normal newborn. This assumption is based on the fact that the images provided are from a newborn screening program and different hemoglobinopathies are rare, so most of the samples are normal newborn with highest quantity of HbF, then HbA and a fade band above HbA of acetylated HbF.

These assumptions were used as the base when starting to develop the method for straightening a distorted image. Figure 3.1 depicts an example image in the dataset which fulfills the assumptions. The image contains four control samples at correct positions, the starting and ending points for each sample are highlighted, and most of the samples are normal newborn.

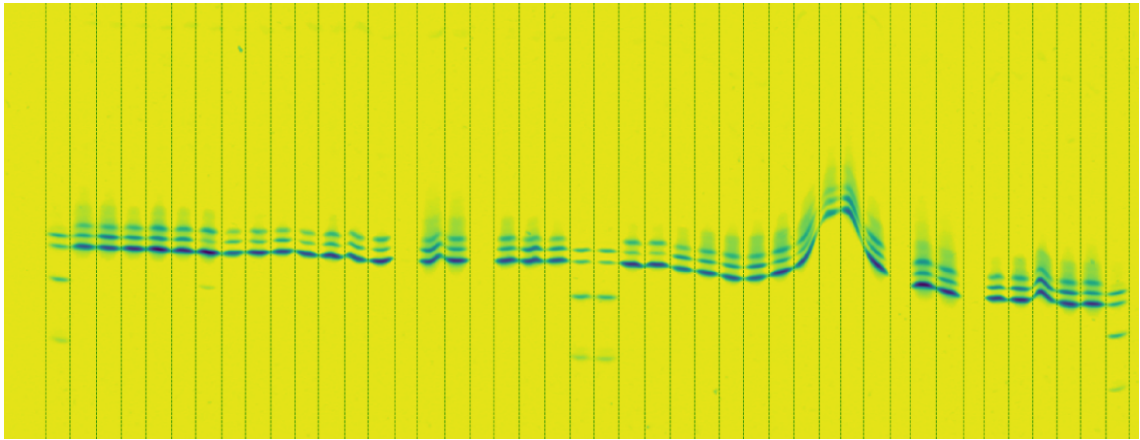


Figure 3.1: Example image in the dataset which fulfills the assumptions. The starting and ending positions for each sample are highlighted.

3.2 Geometric correction method

This thesis' proposed method attempts to solve the two separate issues first introduced in Chapter 1. To recap, the first issue is the bending of bands within a single sample and the second issue is the bending of the whole sample area. The geometric correction method includes solving both of these issues by first straightening every curved band, and then aligning all samples with the first control sample. The different stages of the method are presented thoroughly in the following two sections.

3.2.1 Straightening of bands

The straightening of curved bands within a sample is done independently for each sample. The method includes first finding the position of a single band, and then shifting each pixel column of the sample up or down so that the band is straight. Because all the bands within a sample are curved the same way, locating a single band and straightening it causes all the bands within the sample to be straightened.

The tracing of a single band starts by representing each pixel column of the sample as a 1D-signal. In an ideal case, a single band could be located by finding the pixel with the greatest intensity in each column, detected as the highest peak in the 1D-signal. However, this is almost never the case, because the highest peak can alter between the bands for adjacent columns, or even be completely outside the bands, because of an artifact such as a dust particle with greater intensity.

To filter out the artifacts, the search area for locating a single band is reduced. This filtering process relies on the assumption that most of the highest peaks in each column are close together, with only a few outliers. Figure 3.2 shows the process of reducing the search area in four images. First image shows the original sample, from which the 1D-signals for each column are computed. Second image shows the found highest peaks plotted on top of the original image in red. This example shows the far from perfect case, where the highest peaks are spread across multiple bands. In

the third image, the 10th and 90th percentiles of the pixel row coordinates, shown as horizontal lines, are calculated from the found highest peaks. The last image shows the final search area, after padding has been added. Ten pixels of padding is added at the top and at the bottom of the search area, because the range between the 10th and 90th percentiles can be very small, if the highest peaks are already very close to each other.

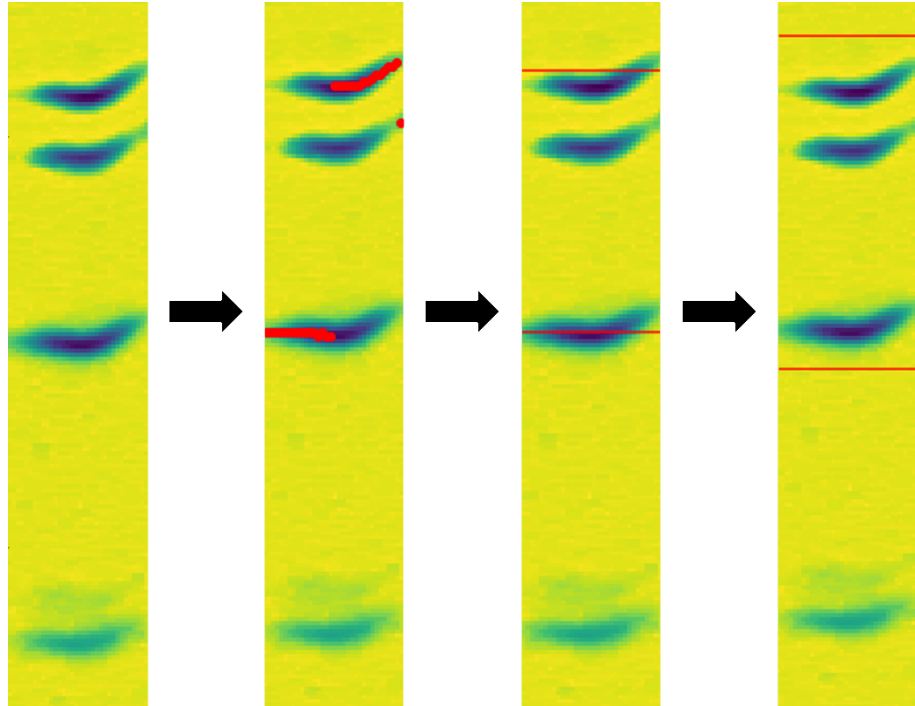


Figure 3.2: Process of reducing the search area to avoid artifacts

As the example shows in Figure 3.2, only finding the highest peaks is not always sufficient for locating a single band. Therefore, after reducing the search area, the 1D-signal for each column is created again, but this time, the highest peaks are searched inside the search area. An assumption is made that subsequent pixels in a single band are always one of the four highest peaks. Therefore, the coordinates for the four highest peaks are stored. Then the process iterates through a loop which compares the distances between the previous column's peaks to the following column's peaks. Starting point is chosen to be the peak in the first column that is

closest to the median of the first five columns' highest peaks. This starting point is then compared to the next column's four highest peaks, and the one closest to it is chosen as the next point. This process is repeated for all the columns. Figure 3.3 shows the four highest peaks found, and the chosen points after the distance comparison process. Last image is the outcome after the chosen points are shifted either up or down so that they form a straight horizontal line. In the first image, the four highest peaks are plotted with different colours: highest peak in red, second highest peak in green, third highest peak in magenta, and fourth highest peak in cyan. In the chosen points, it can be seen that closest to the starting point, which is the median of the first five highest peaks, is the highest peak. Then, when comparing which peak is the closest in the next column, the closest point is always the highest peak until around halfway through the band. After that, the closest peak is the second highest peak, until the last pixel, when the closest peak is the third highest peak.

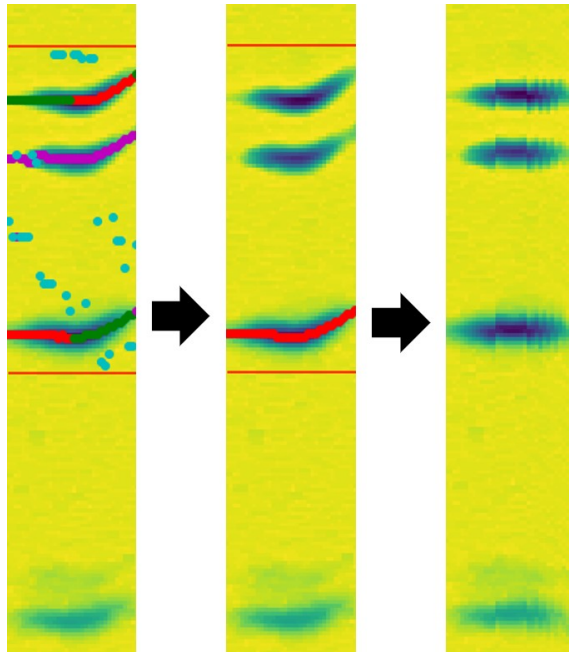


Figure 3.3: Locating the four highest peaks and using them to trace a single band. Then the chosen points are shifted up or down so that they form a horizontal line.

In a rare case, when the next pixel in a band is none of the four highest peaks, the previous point is set as the next point as well. This rare case is detected by calculating the distance to the nearest point, and if it is over a certain threshold, it is assumed that the nearest point is in another band.

The last step in the straightening of bands is applying a 6th degree polynomial function to the chosen points. This results in smoother bands, especially for samples that have significantly steep bands. Figure 3.4 shows the effect of the 6th degree polynomial fit for a band with a very large slope. Figure 3.4 (a) shows the result without the polynomial fit and Figure 3.4 (b) shows the result with the polynomial fit. The fitting of the polynomial is done in all samples, but in most cases, it does not change the chosen points, since most of the bands are already fairly straight.

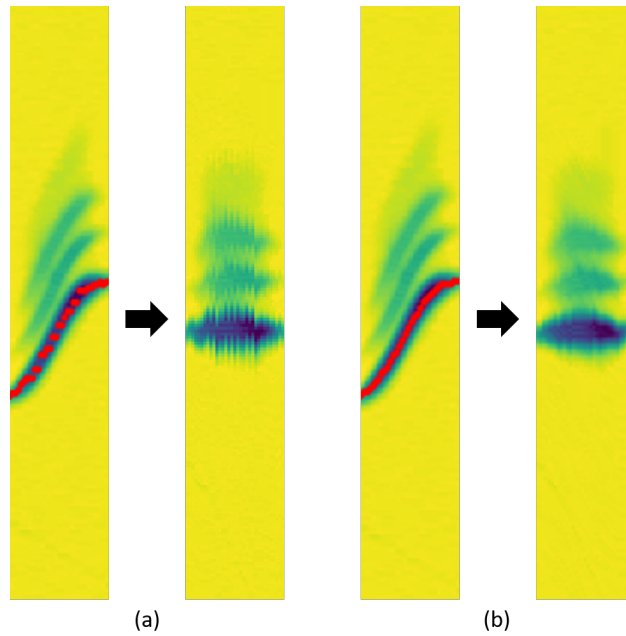


Figure 3.4: The effect of polynomial fit to the points found in a band

The algorithm for straightening the bands is shown as pseudocode in Listing 1. Figure 3.5 shows what the image looks like after the straightening of all the bands. After this process, the next step is to align all the samples with the first control sample.

Listing 1 Algorithm for straightening the bands

FindPeakCoordinate(column, N) : Find the coordinate for the N-th highest peak

Percentile(list, N): Computes the N-th percentile of the list

Median(list): Computes the median of the list

FindClosestPeak(value, list): Compute which element in the list is closest to the given value

Polynomial(x, y, deg): Fit a polynomial of degree deg to points (x,y)

MovePixels(column, N): Moves pixels in a column N-times. Up if N is positive and down if N is negative

```

1: for sample = 0, ..., i do
2:   peakMax = []
3:   for pixelColumn = 0, ..., j do
4:     peakMax[j] = FindPeakCoordinate(pixelColumn, 1)
5:   end for
6:   10thPerCoordPadding = Percentile(peakMax[0...j], 10) - 10
7:   90thPerCoordPadding = Percentile(peakMax[0...j], 90) + 10
8:   fourHighestPeaks = []
9:   for pixelColumn = 0, ..., j do
10:    filteredPixelColumn = pixelColumn[10thPerCoordPadding:90thPerCoordPadding,]
11:    fourHighestPeaks[j] = FindPeakCoordinate(filteredPixelColumn, 1:4)
12:  end for
13:  highestPeaksMedian = Median(fourHighestPeaks[0:4][0])
14:  closestPeakToMedianCoordinate = FindClosestPeak(highestPeaksMedian, fourHighestPeaks[0][0:3])
15:  maxCoordinates = []
16:  maxCoordinates[0] = closestPeakToMedianCoordinate
17:  for pixelColumn = 1, ..., j do
18:    closestPeakCoordinate = FindClosestPeak(maxCoordinates[j-1], fourHighestPeaks[j][0:3])
19:    maxCoordinates[j] = closestPeakCoordinate
20:    if absolute(maxCoordinates[j-1] - maxCoordinates[j]) >= distanceThreshold then
21:      maxCoordinates[j] = maxCoordinates[j-1]
22:    end if
23:  end for
24:  sampleFit = Polynomial(0:length(maxCoordinates), maxCoordinates, 6)
25:  polynomialCoordinates = sampleFit(0:length(maxCoordinates))
26:  polynomialCoordinatesMedian = Median(polynomialCoordinates)
27:  for polynomialCoordinates = 0, ..., c do
28:    coordinateDifference = c - polynomialCoordinatesMedian
29:    sample[i], c = MovePixels(sample[i], c, coordinateDifference)
30:  end for
31: end for

```

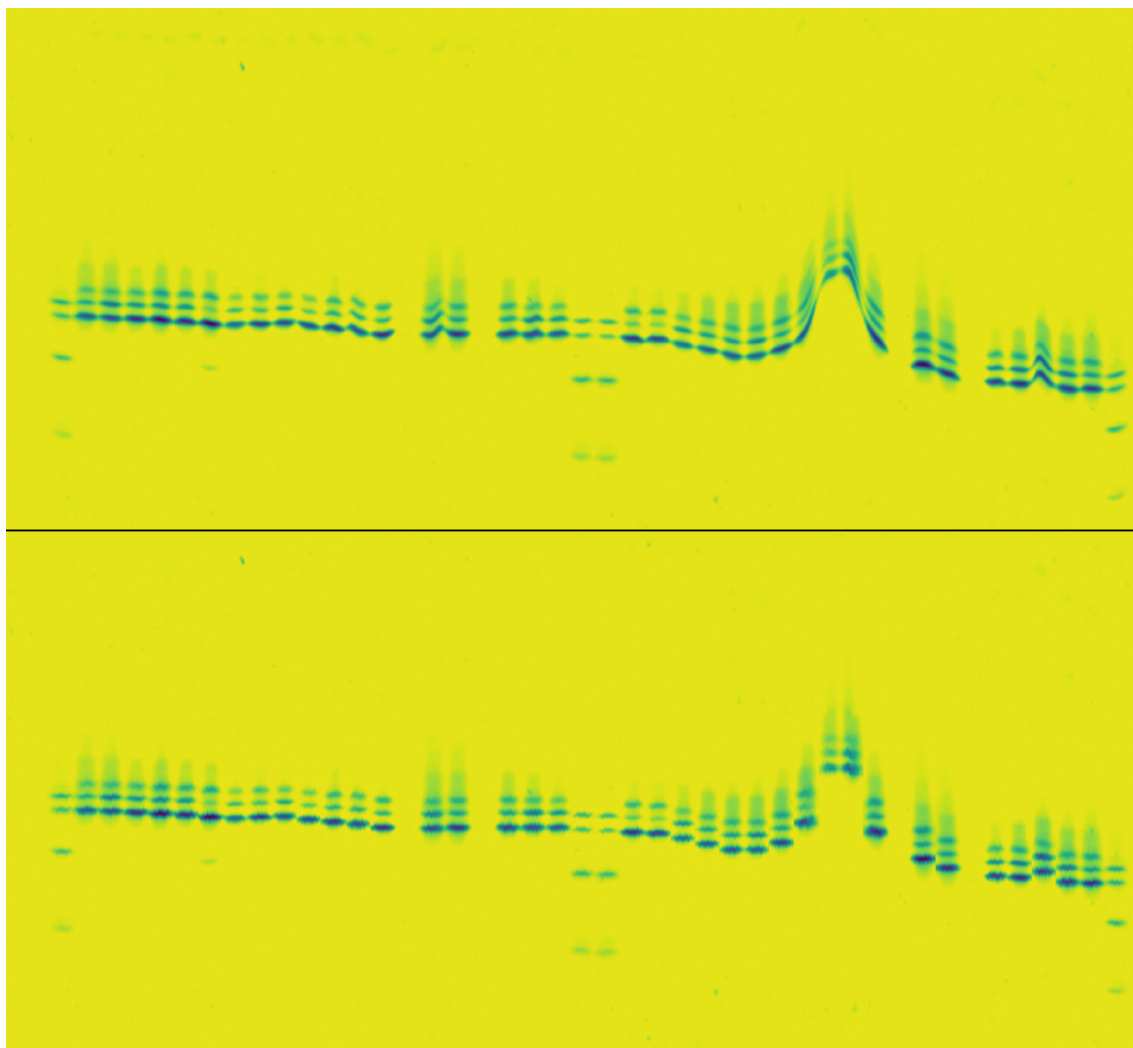


Figure 3.5: Image of a gel side before and after straightening each band

3.2.2 Alignment of samples

The first step in the alignment process is identifying the HbF-band from the first control sample. This step is important, since control samples' bands are used as a reference, to which the other samples are aligned. Similar to the straightening of the bands, a 1D-signal is created from the sample. However, in this case, the 1D-signal is created from the whole sample, rather than from each pixel column. The mean value for each row of the image is calculated and then stored as one vector. This process was shown in Figure 2.7. If the control samples were manufactured or produced in a laboratory according to widely used standards, it could be assumed that the second highest peak would be HbF, but due to variety in the mixture of the control samples, it cannot be relied on, since the quantities of the hemoglobin variants in the control samples might change. Since it is known that the order of the bands in a control sample is always HbA, HbF, HbS, and HbC, the 1D-signal can be scanned from left to right, and the second peak over a certain threshold can be identified as HbF. This is shown as pseudocode in Listing 2.

Listing 2 Algorithm for finding the HbF-band of the first control sample

Mean(array, axis): Calculate the mean along the specified axis

FindPeaks(signal, threshold) : Find all peaks above a certain threshold

1: signal = *Mean*(controlSamples[0], axis=1)

2: peaks = *FindPeaks*(signal, threshold)

3: firstControlSampleFLocation = peaks[1]

Next, the whole sample area is moved either up or down, so that the newly identified HbF-band of the first control sample is aligned with a preselected y-coordinate of 300, which for this dataset was found to be the most feasible location. This ensures that the preselected y-coordinate can be used as a reference location, since it is certain that it is the row that contains HbF-band of the first control sample. The moving of the whole sample area also establishes that all the images straightened

will have the sample area at the same position. Listing 3 shows this process as pseudocode and Figure 3.6 shows the image before (A) and after (B) moving the whole sample area to the preselected y-coordinate.

Listing 3 Algorithm for moving the whole sample area, so that the HbF-band of the first control sample is aligned with the specified y-coordinate

MoveSampleUp(sample, N): Move sample N-pixels up

MoveSampleDown(sample, N): Move sample N-pixels down

```

1: coordinateDifference = firstControlSampleFLocation - yCoord
2: if coordinateDifference >= 0 then
3:   image = MoveSampleUp(image, coordinateDifference)
4: else
5:   image = MoveSampleDown(image, absolute(coordinateDifference))
6: end if

```

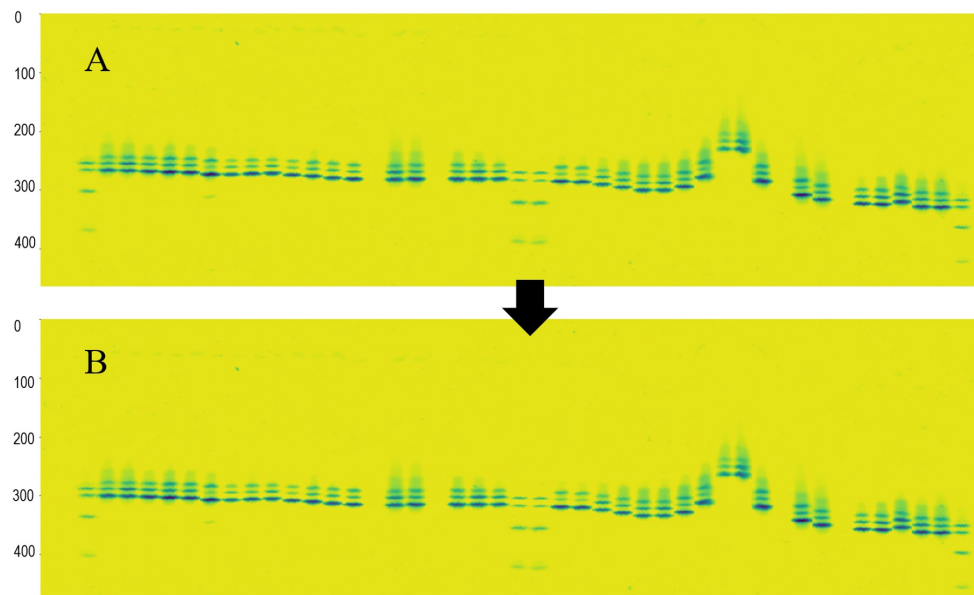


Figure 3.6: Before (A) and after (B) moving the whole sample area down to a preselected y-coordinate 300

Then the 1D-signals are created for the rest of the samples. For each of these signals, the location of the highest peak is stored. The highest peak found is used for aligning the sample with the HbF-band of the first control sample, since in a normal newborn sample, the highest peak is most often HbF. Listing 4 shows the pseudocode for this function. The highest peak in a 1D-signal of a sample can also be some other band, but these rare cases are handled later.

Listing 4 Algorithm for finding the greatest intensity band in each sample and aligning them with the HbF-band of the first control sample

Mean(array, axis): Calculate the mean along the specified axis

FindPeakCoordinate(signal, N) : Find the coordinate for the N-th highest peak

MoveSampleUp(sample, N): Move sample N-pixels up

MoveSampleDown(sample, N): Move sample N-pixels down

```

1: peakMax = []
2: for sample = 0, ..., i do
3:   signal = Mean(sample, axis=1)
4:   peakMax[i] = FindPeakCoordinate(signal, 1)
5: end for
6: for sample = 1, ..., i do
7:   coordinateDifference = peakMax[i] - yCoord
8:   if coordinateDifference >= 0 then
9:     sample = MoveSampleUp(sample, coordinateDifference)
10:  else
11:    sample = MoveSampleDown(sample, absolute(coordinateDifference))
12:  end if
13: end for

```

Figure 3.7 shows as an example three 1D-signals created from patient samples. It can be interpreted from the signals that sample 1 has HbA as the highest peak. Sample 2, on the other hand, has HbF as the highest peak. An abnormal variant can also be identified close to HbC. Sample 3 also has HbF as the highest peak. Additionally, it has no clear HbA-band, but it has a visible variant, most likely HbS. The diffusion around HbF is caused by long storage time for the sample, which causes these "aging bands". Over time, the hemoglobin molecules start to degrade, which is accelerated in high-humidity environments [33]. Figure 3.8 shows the outcome after aligning every sample with the HbF-band of the first control sample. In this example, the process performs as intended, since all the patient samples have HbF-band as the highest peak in their 1D-signals.

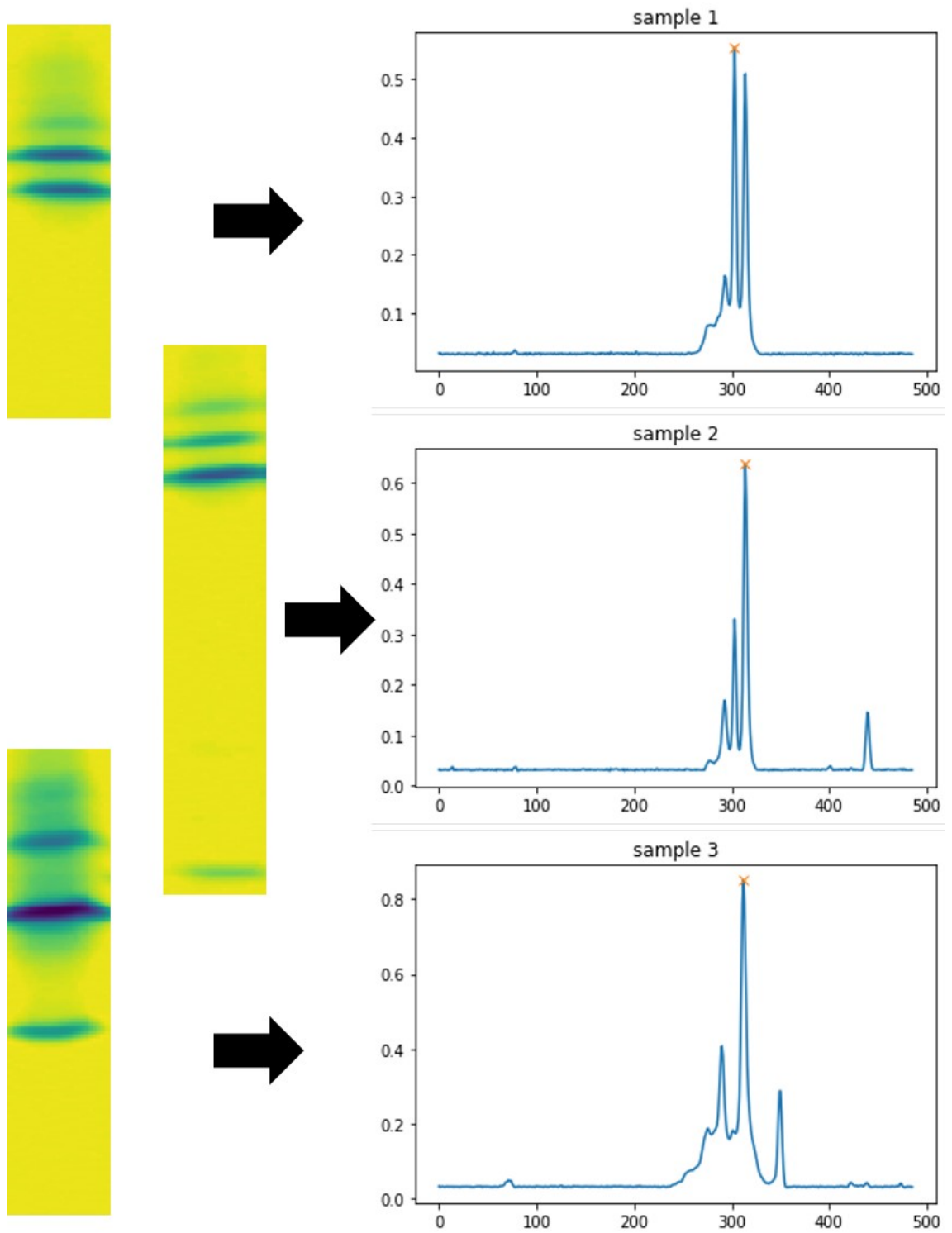


Figure 3.7: Example of 1D-signals created from three patient samples. The sample images in this figure are cropped.

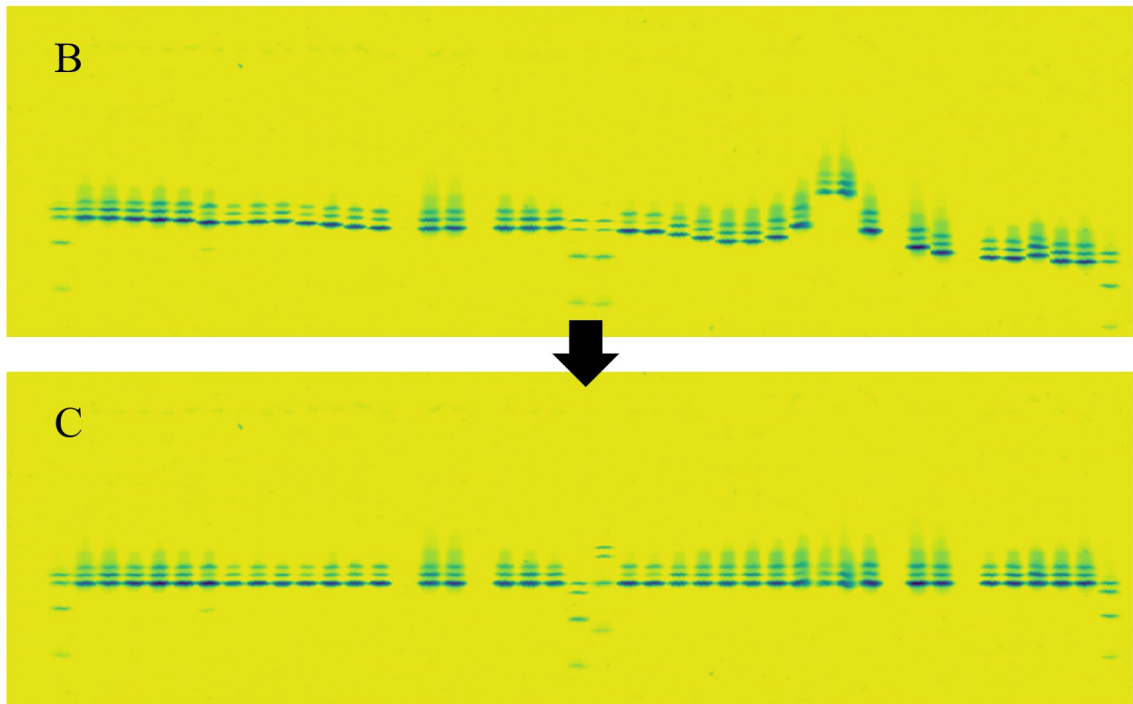


Figure 3.8: Before (B) and after (C) aligning all samples with the HbF-band of the first control sample by their highest peak

Control samples, on the other hand, are often aligned falsely, since they do not generally have the HbF-band with the greatest intensity. This is fixed next by finding the location of the HbA-band in the control samples and aligning them with the first control sample's HbA-band. The identification of HbA-band is done the same way as in the step of identifying the HbF-band of the first control sample. Control samples are aligned with the HbA-band, since the accuracy of correctly identifying HbA-band in a control sample is higher than correctly identifying a HbF-band. Listing 5 shows this function as pseudocode and Figure 3.9 shows the image before (C) and after (D) aligning the control samples.

Listing 5 Algorithm for finding the HbA-band from each control sample and then aligning them with the HbA-band of the first control sample

Mean(array, axis): Calculate the mean along the specified axis

FindPeak(signal, threshold) : Find all peaks above certain threshold

MoveSampleUp(sample, N): Move sample N-pixels up

MoveSampleDown(sample, N): Move sample N-pixels down

```

1: controlSampleALocations = []
2: for controlSample = 0, ..., i do
3:   signal = Mean(controlSample, axis=1)
4:   peaks = FindPeaks(signal, threshold)
5:   controlSampleALocations[i] = peaks[0]
6: end for
7: for controlSample = 1, ..., i do
8:   coordinateDifference = controlSampleALocations[i] - controlSampleALocations[0]
9:   if coordinateDifference >= 0 then
10:    controlSample = MoveSampleUp(controlSample, coordinateDifference)
11:   else
12:    controlSample = MoveSampleDown(controlSample, absolute(coordinateDifference))
13:   end if
14: end for

```

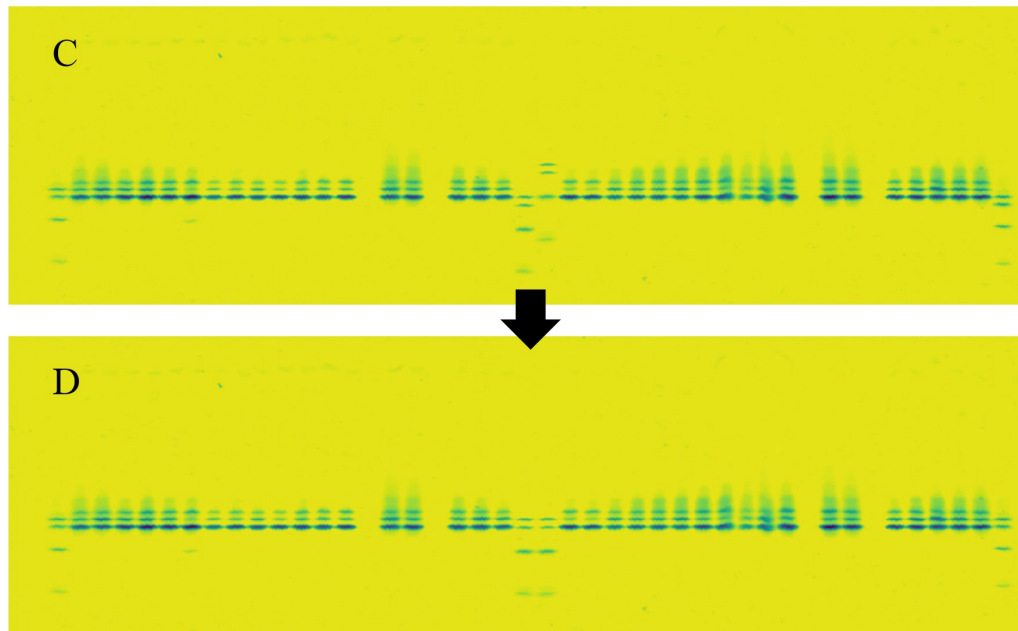


Figure 3.9: Before (C) and after (D) aligning all control samples with the first control sample

In addition to falsely aligning control samples, the alignment of all the samples by their highest peak causes false corrections also to patient samples that do not have the HbF-band with the greatest intensity. For example, adult samples have higher intensity of HbA than HbF, and even though the images in the dataset used in this thesis are from a newborn screening program, some samples contain more HbA than HbF. These samples, falsely aligned with HbF based on their highest peak, can be identified as samples (1) having only HbA-band and no HbF-band, and (2) having high amounts of HbA and some amounts of HbF. Rare cases when a sample contains some other than HbA or HbF in highest quantity are discussed later in Chapter 4.

The two cases when a sample contains more HbA than HbF are solved with two different approaches. First, the locations for the highest and the second highest peaks for all the patient samples are stored. Then samples that have only HbA and no HbF are identified by calculating the ratio between the highest and the second highest peaks. If the second highest peak is of certain percentage or less of the highest, which depends on the imaging system used, it is assumed that the sample contains only one distinct band, which can be identified as HbA, and then that sample is aligned with the HbA-band of the first control sample. Figure 3.10 shows an example of this.

The case where a sample contains some amounts of HbF, but higher amounts of HbA, can be detected by checking whether the y-coordinate for the location of the second highest peak is larger than the highest peak, meaning that the second highest peak is "under" the highest peak, when the y-coordinate 0 is at the top of the image and increases downwards. The second highest peak is searched between 10 and 30 pixels under the highest peak, because it is the typical distance between the HbA- and HbF-bands. Figure 3.11 shows an example of this. Listing 6 shows as pseudocode the algorithm that detects and fixes these two cases.

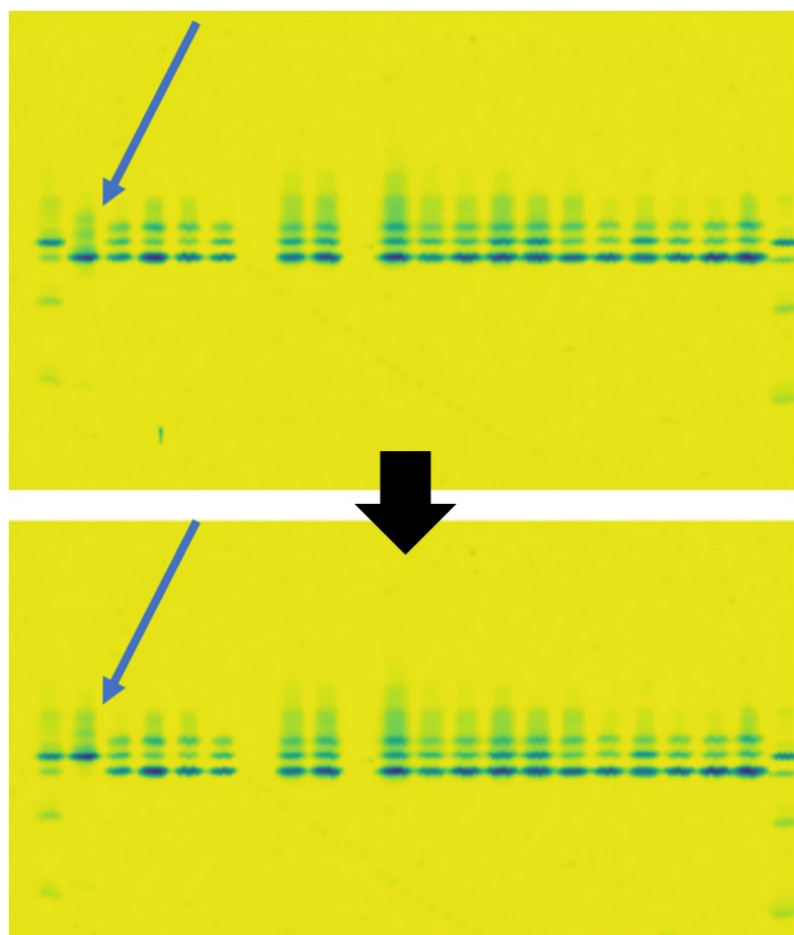


Figure 3.10: The first image shows each sample's greatest intensity band aligned with the first control sample's HbF-band. The sample arrowed is detected as only having one distinct band and therefore aligned with the HbA-band of the first control sample.

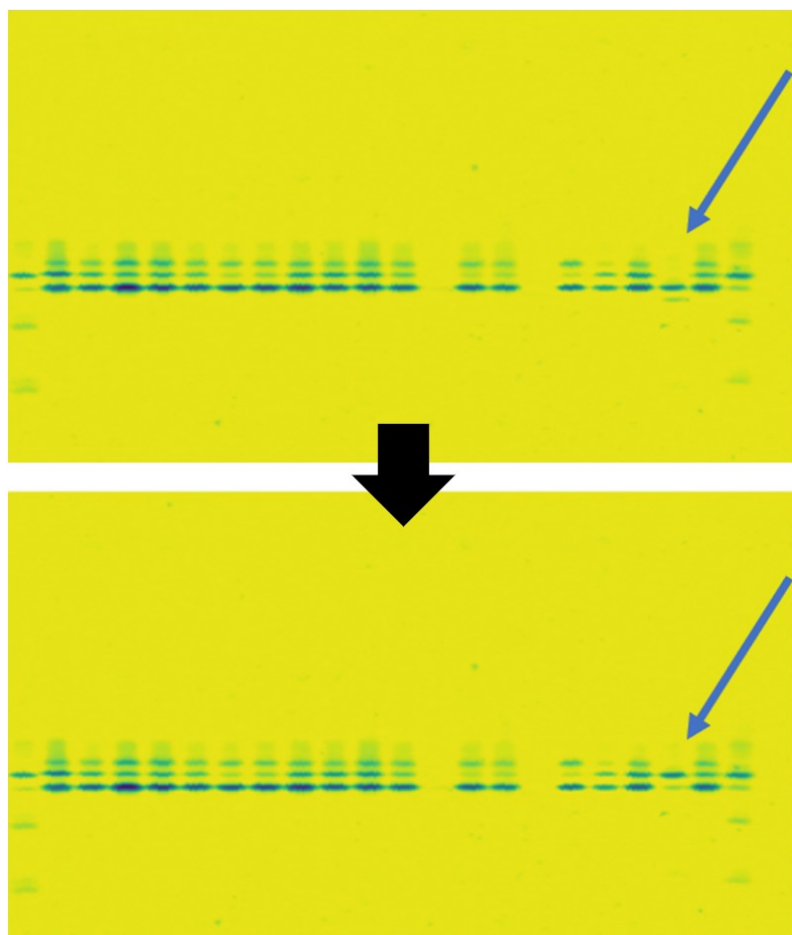


Figure 3.11: The first image shows each sample's greatest intensity band aligned with the first control sample's HbF-band. The sample arrowed is detected as having higher amounts of HbA than HbF and therefore aligned with the HbA-band of the first control sample.

Listing 6 Algorithm for detecting cases where the HbA-band of a sample has the greatest intensity and then correcting its alignment with the HbA-band of the first control sample

Mean(array, axis): Calculate the mean along the specified axis

FindPeakCoordinate(signal, N) : Find the coordinate for the N-th highest peak

MoveSampleUp(sample, N): Move sample N-pixels up

MoveSampleDown(sample, N): Move sample N-pixels down

```

1: for sample = 0, ..., i do
2:   if sample not controlSample then
3:     signal = Mean(sample, axis=1)
4:     highestPeak = FindPeakCoordinate(signal, 1)
5:     secondHighestPeak = FindPeakCoordinate(signal, 2)
6:     if signal[secondHighestPeak]/signal[highestPeak] <= ratioThreshold then
7:       coordinateDifference = highestPeak - controlSampleALocations[0]
8:       if coordinateDifference >= 0 then
9:         sample = MoveSampleUp(sample, coordinateDifference)
10:      else
11:        sample = MoveSampleDown(sample, absolute(coordinateDifference))
12:      end if
13:    else if highestPeak+30 > secondHighestPeak > highestPeak+10 then
14:      coordinateDifference = highestPeak - controlSampleALocations[0]
15:      if coordinateDifference >= 0 then
16:        sample = MoveSampleUp(sample, coordinateDifference)
17:      else
18:        sample = MoveSampleDown(sample, absolute(coordinateDifference))
19:      end if
20:    end if
21:  end if
22: end for

```

The process of aligning the samples is summarized in Figure 3.12.

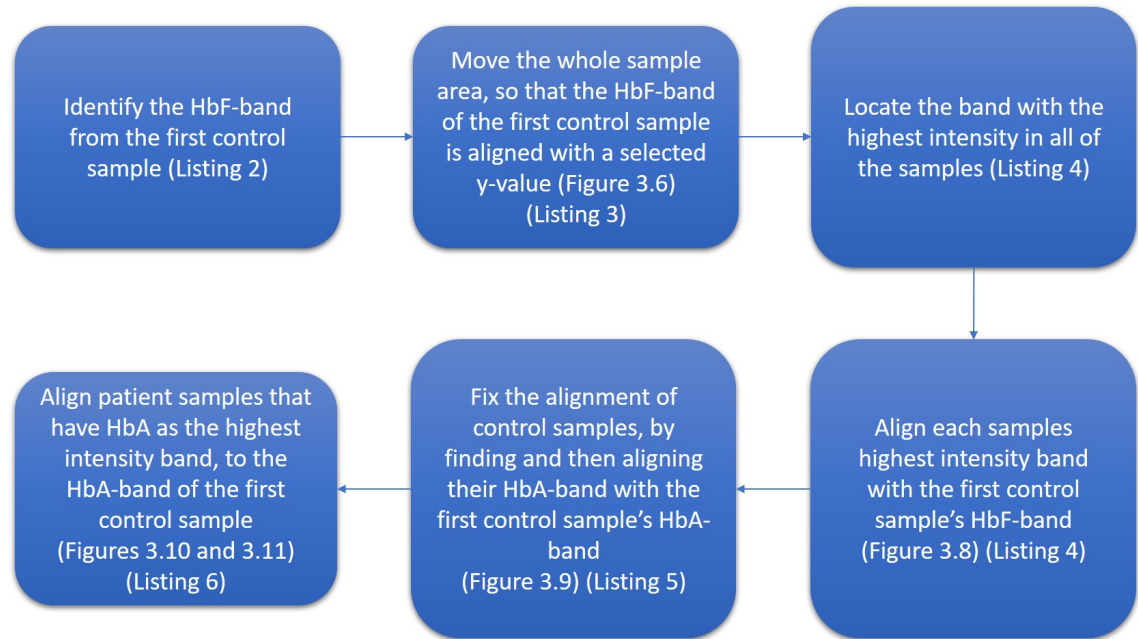


Figure 3.12: Workflow for aligning the samples

After straightening the curved bands and aligning the samples, the last step in the whole process of straightening the sample area is applying a Gaussian blur to the image. It is done especially to smoothen the bands that might still have rough edges after shifting the pixel columns up and down. The Gaussian blur is applied with a Gaussian kernel of width and height of 5. Figure 3.13 shows the effect of Gaussian blur to a few example samples.

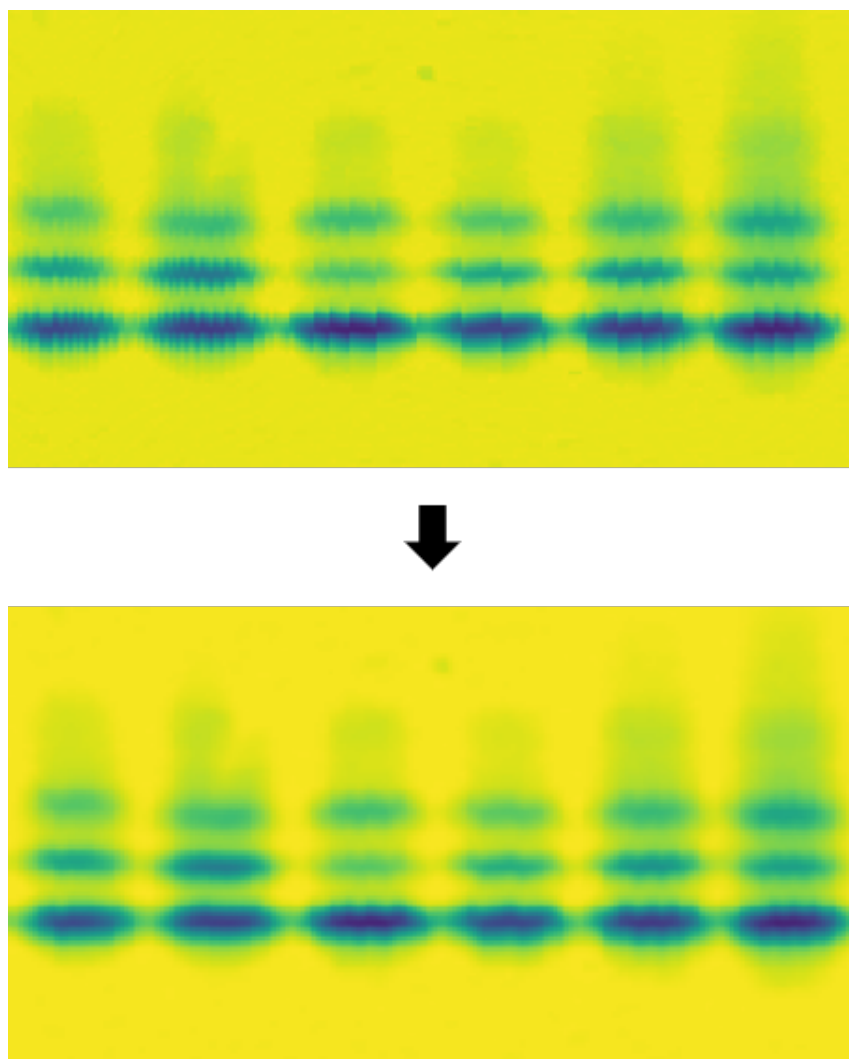


Figure 3.13: Effect of Gaussian blur

3.3 Image correction metrics

Two metrics were created to assess how much an image has been straightened. The first metric is the amount of the pixel columns moving up or down inside each sample, when straightening the bands, which is stored as the absolute sum of the movements for each sample. The second metric is the number of pixels each sample is moved up or down relative to the first control sample.

Figure 3.14 shows these two metrics plotted as bar charts with the image before and after straightening. For both bar charts, the x-axis is the sample number. The first bar chart shows the absolute sum of pixel column movements inside each sample. The value has been normalized between 0 and 1, with min-max normalization, using 0 as minimum value and 801 as maximum value, which is the highest amount of correction in all the samples of the whole dataset. The wave can be detected from the first bar chart, since samples from 28 to 31 have large values, which means they were originally very curved and have been straightened a lot. The second bar chart shows the number of pixels each sample has been moved up or down relative to the first control sample. The wave in the gel can be detected by the negative bars in samples 29 and 30, which means they have been aligned the opposite direction as rest of the samples. The whole sample area can be seen descending from left to right, since this bar chart shows the bars ascending from the left to right.

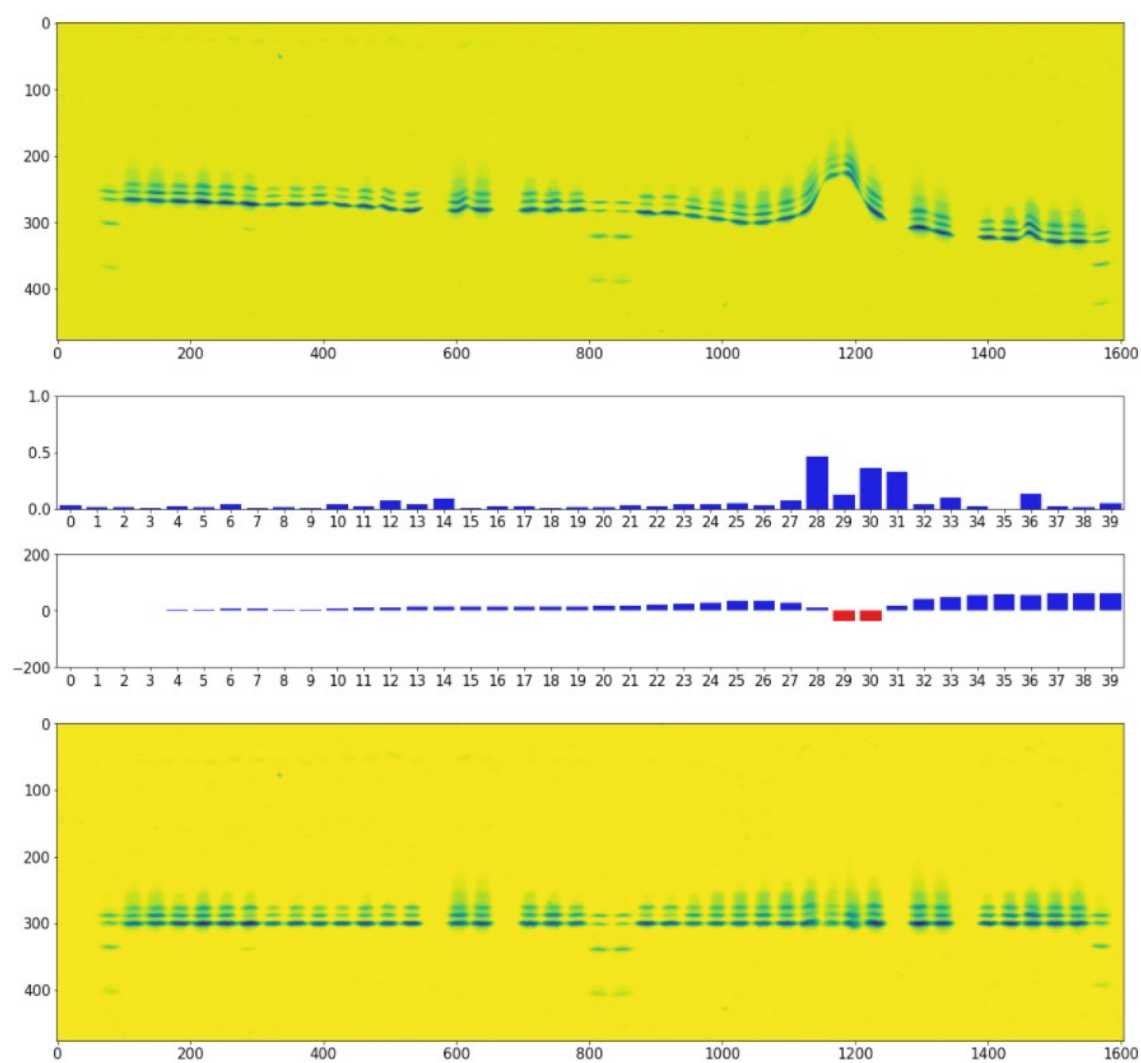


Figure 3.14: Two metrics plotted with the image before and after straightening

3.4 Evaluation process

The proposed method is evaluated with three different tests. The first test estimates how many of the real images in the dataset were straightened correctly by examining whether there were any errors in the straightening of curved bands or in the alignment of samples. The second test is based on the requirement that the straightening method should retain the relative quantities of the bands within a sample. This is done by comparing the relative peak areas of the 1D-signal created from each control sample before and after the straightening. The last evaluation method is a stability test, which is used to test the method with extreme synthetic images created to assess the limits of the proposed method.

3.4.1 Performance testing with real images

The images in the dataset are divided into three levels, based on the severity of the distortions in the gel. First distortion level contains images that have straight sample area or only slight distortions. Second distortion level contains images that have small waves and bending in the bands. The third distortion level contains images that have large waves in the sample area, bands extremely curvy, or control samples affected, so they cannot be used as a reference. The performance of the method is evaluated for images in each distortion level using two variables, straightenability (image is straightened correctly or incorrectly) and interpretability (all samples can or cannot be interpreted after straightening). Straightened correctly includes all images that are straightened without errors. If an image contains at least one sample that is incorrectly straightened due to the proposed method, the image is marked as straightened with errors. Interpretability of an image means that, after straightening, all the samples are still good quality for interpretation. Uninterpretable images contain at least one sample that cannot be interpreted after straightening.

3.4.2 Comparing the relative band quantities

In the second test, the areas of the peaks from each control sample's 1D-signal are calculated before and after the straightening. The peak areas are then compared to see whether the relative quantities of the peaks have changed due to the proposed method. This test is only done for control samples, because they are the most important samples that need to remain unchanged, since they are used as the reference for patient samples. It is assumed that, if the control samples remain unchanged, then the method also keeps patient samples unchanged. The Equations (2.1) and (2.2) introduced in Section 2.3 are used for calculating the peak areas.

3.4.3 Stability test with synthetic images

Stability test is used to test the limits of the proposed method on a set of synthetic images. This is done, because some extreme cases are so rare that they do not occur in the dataset of real images. Five following situations were created synthetically: a gel containing lots of empty sample positions which mimics a second-tier screening gel, a gel with only one control sample in the middle which simulates gel run protocol deviations where the control material runs out, a gel having an abnormal variant with greatest intensity above HbF-band, and another gel where the abnormal variant is below HbF-band. The fifth image is a gel with added random Gaussian noise, which simulates random image artifacts such as dust particles. The standard deviation of the distribution of the Gaussian noise was based on the normal variation present in the background of the images. The standard deviation of the pixel intensities was calculated for the background in each image. The mean of these standard deviations was 0.005. In the stability test, the random Gaussian noise was added with a distribution that has a mean of 0 and a standard deviation of 0.03, which is approximately six times the normal variation.

3.5 Implementation

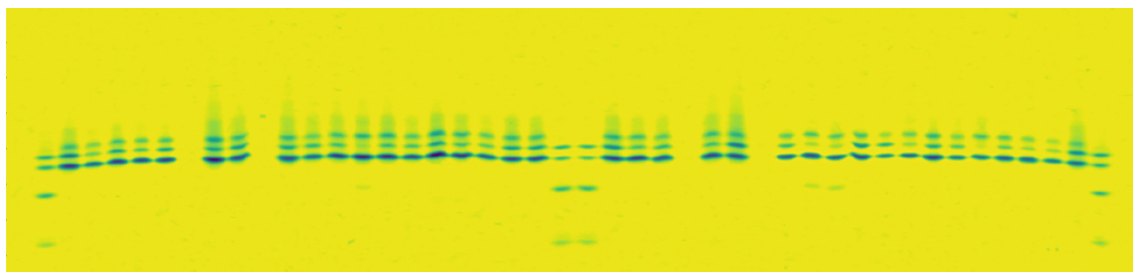
For this thesis, Python 3.6.7 was used as the programming language in a JupyterLab 0.35.4 environment. The following packages were used for data analysis and algorithm development: pandas 0.23.4, NumPy 1.15.4, OpenCV 4.5.4, seaborn 0.11.2, and Matplotlib 3.0.2. All the calculations were done using Intel® Core™ i9-9940X processor. The mean runtime for straightening one image was 0.319 seconds.

4 Results

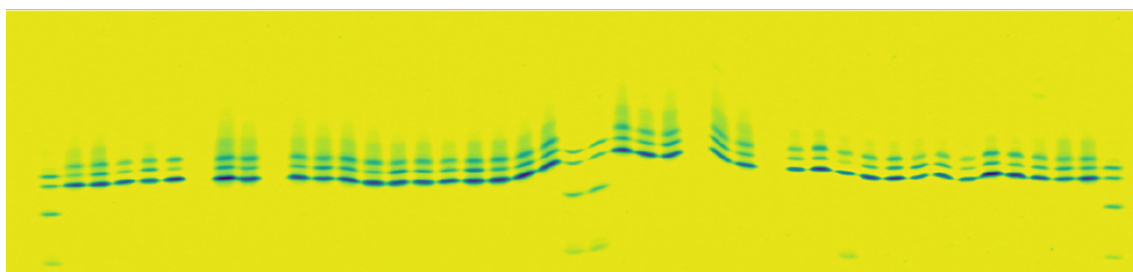
4.1 Performance test

The dataset contains 142 images of IEF gel sides, which are divided into three categories, introduced in Section 3.4.1, based on the level of distortion. The first level 'low' contains 103 images of almost straight gels, second level 'medium' contains 28 images of gels with small distortions, and the third level 'high' contains 11 images of gels with a lot of distortion. Examples for all three categories are depicted in Figure 4.1.

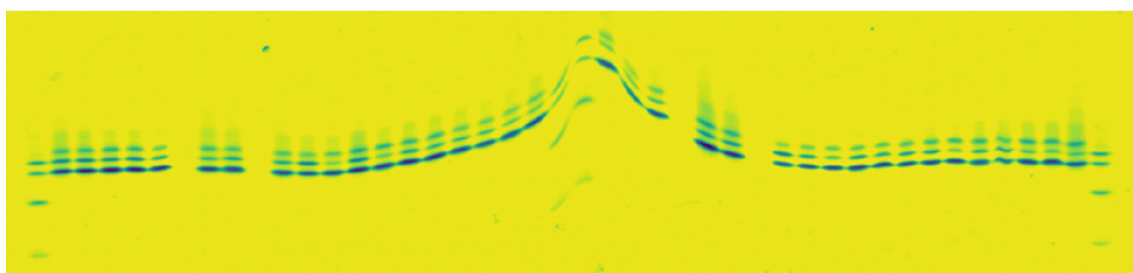
It is important to note that the performance of the proposed method is evaluated for each image, and not for each sample. Also, the correctness of the straightening is purely based on visual inspection of the straightened image and is not absolute. If a straightened image contains at least one sample, with an error, the image is flagged. In Chapter 5, the cause for each error is investigated and explained thoroughly and categorized based on whether the error was in the straightening of bands or in the alignment of samples. Additionally, even though an image is marked as uninterpretable, most of the samples are still straightened correctly and can be interpreted.



(a)



(b)



(c)

Figure 4.1: Example images of the different distortion levels. Level one (a) has straight sample area or only slight distortions. Level two (b) has small waves and bending in bands. Level three (c) has large waves and extremely curved bands that also affect the control samples.

The results for the performance evaluation are presented in Table 4.1. The proposed method was able to straighten correctly 139 gel side images out of the total 142. From the three not straightened correctly, one was interpretable while the remaining two were not.

Table 4.1: Contingency table of the performance test

	Straightened correctly			Straightened with errors			Total
Distortion level	Low	Medium	High	Low	Medium	High	
Interpretable	102	25	6	1	0	0	134
Not interpretable	0	3	3	0	0	2	8
Total	139			3			142

4.2 Comparison of band quantities

The peak areas were calculated from the 1D-signals created for each control sample in all the images. In total, the peak areas were calculated for 564 different control samples from 142 different images containing 2256 bands. The peak areas were calculated both before and after straightening. The difference in the peak areas is represented as either increase or decrease in the relative peak quantities as percentage points. The median and the standard deviation of the differences are depicted in Table 4.2 for each band separately and for all the bands together. The biggest differences were in the quantitation of the F-band before and after correction, while the A-band quantitation had the least. However, the standard deviation of the differences was larger with the S- and C-bands. Figure 4.2 shows the differences in a histogram for all the bands together.

Table 4.2: Differences in percentage points of the relative band quantities

Band	Median	Standard deviation
A	-0.135	1.340
F	-0.200	1.242
S	0.175	2.127
C	0.171	1.841
All	0.001	1.686

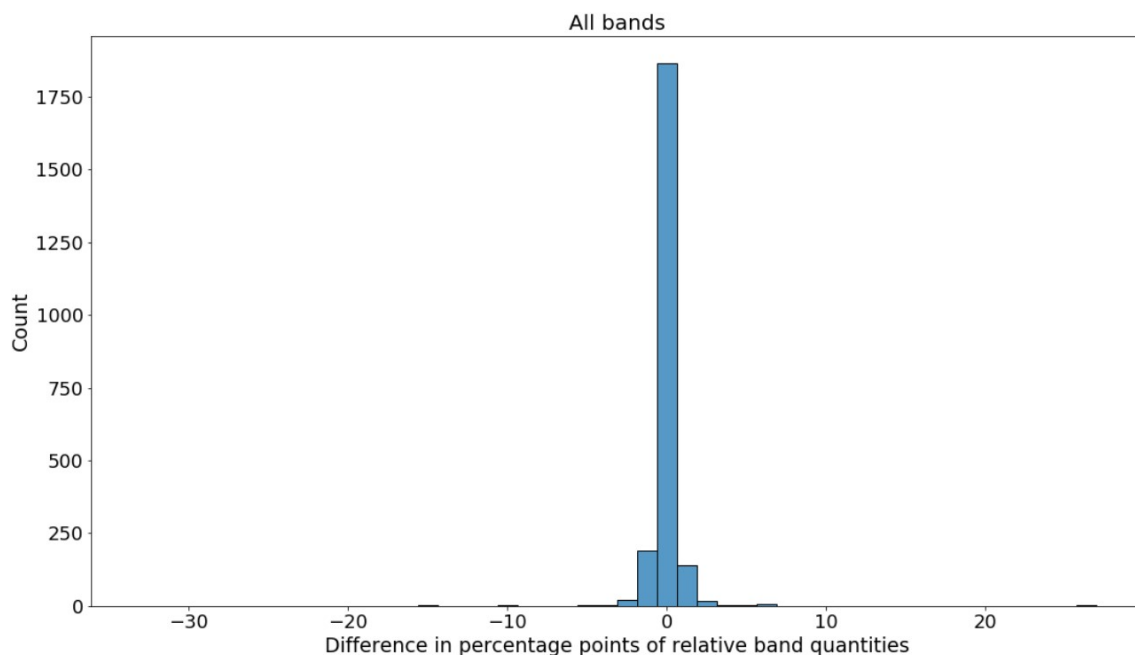
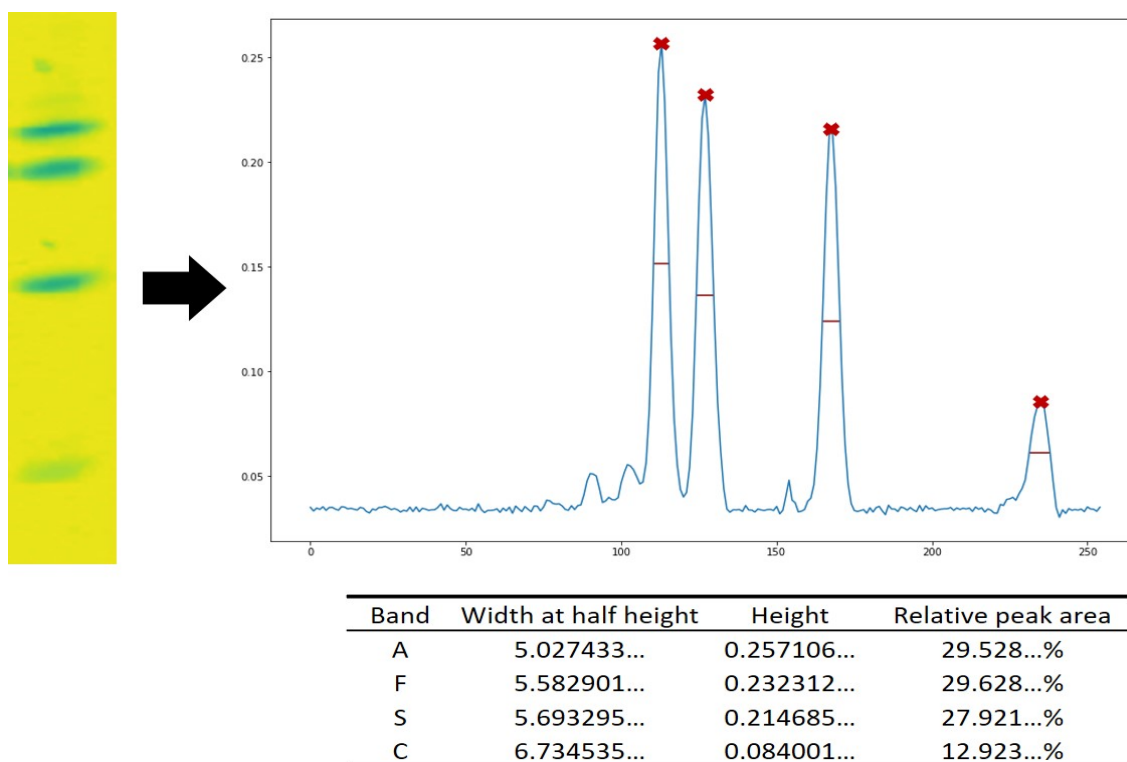
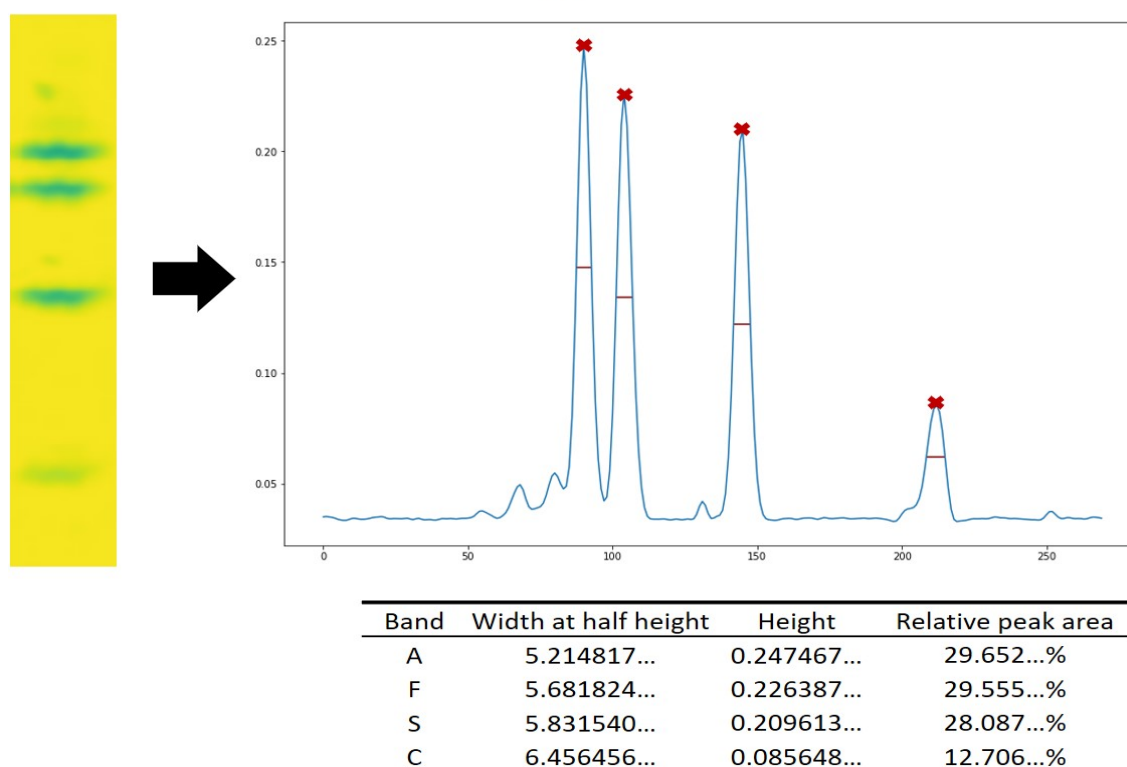


Figure 4.2: Histogram of the differences in the relative band quantities caused by the straightening

Figure 4.3 shows an example of peak area calculations from one control sample's 1D-signal before and after straightening. The width at half height is measured from the red horizontal lines and the height is measured from the peaks marked with 'x'. The differences in relative band quantities for this example: A increased ≈ 0.1 percentage points, F decreased ≈ 0.1 percentage points, S increased ≈ 0.1 percentage points, and C decreased ≈ 0.2 percentage points.



(a)



(b)

Figure 4.3: Example of peak area calculations from one control sample's 1D-signal before (a) and after (b) straightening. The sample images in this figure are cropped.

4.3 Stability test

Five synthetic images were created to represent rare cases that were not present in the dataset containing real images. The synthetic images included a gel with added random Gaussian noise, a very empty gel, a gel having only one control sample in the middle, and a gel having a sample with some other band than HbA or HbF with the greatest intensity. The last case was tested with an image having an abnormal variant with greatest intensity above the HbF-band and another image having the greatest intensity variant below the HbF-band. Figures 4.4, 4.5, 4.6, 4.7, and 4.8 show the outcome of straightening these five images.

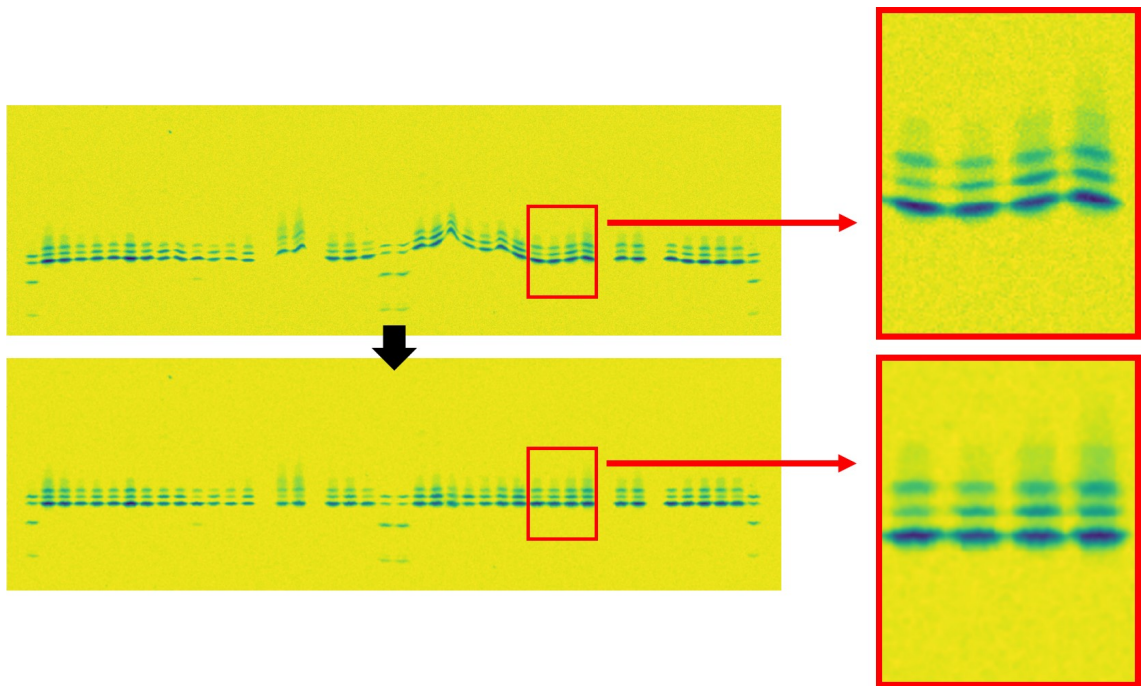


Figure 4.4: Example of a gel with added random Gaussian noise with a distribution that has a mean of 0 and standard deviation of 0.03

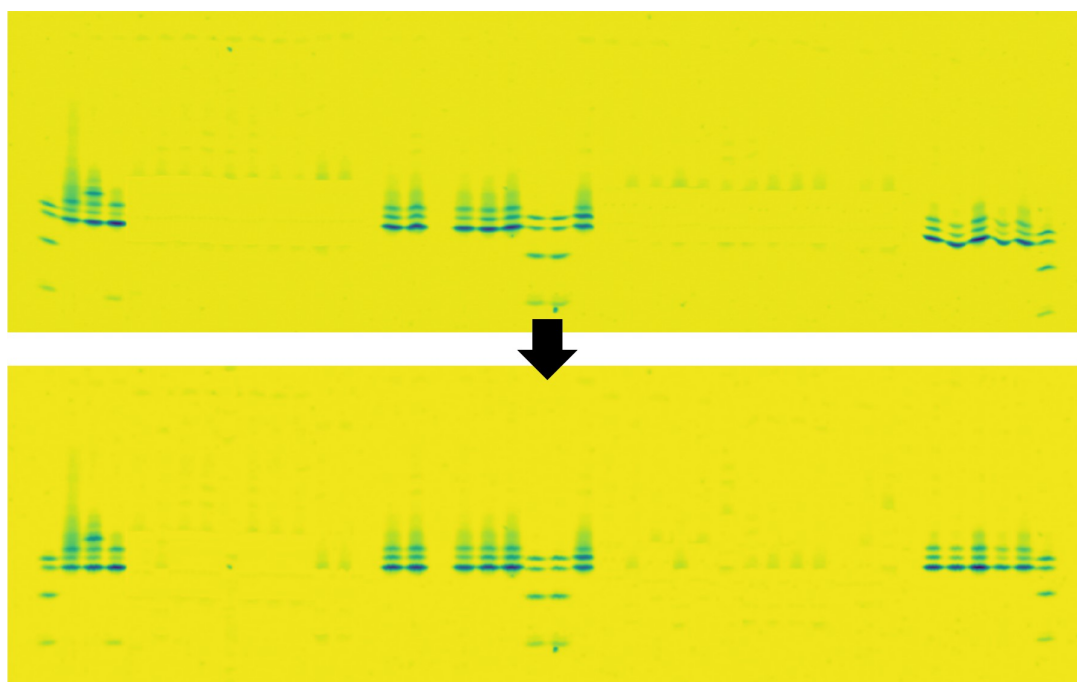


Figure 4.5: Example of a gel with a lot of empty positions

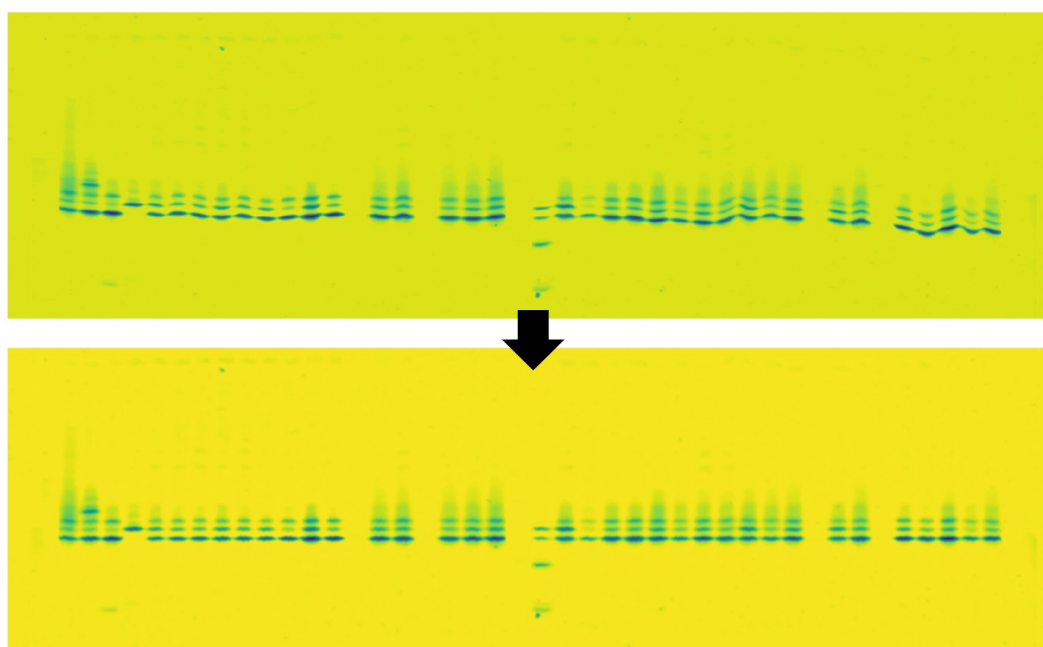


Figure 4.6: Example of a gel with only one control sample in the middle

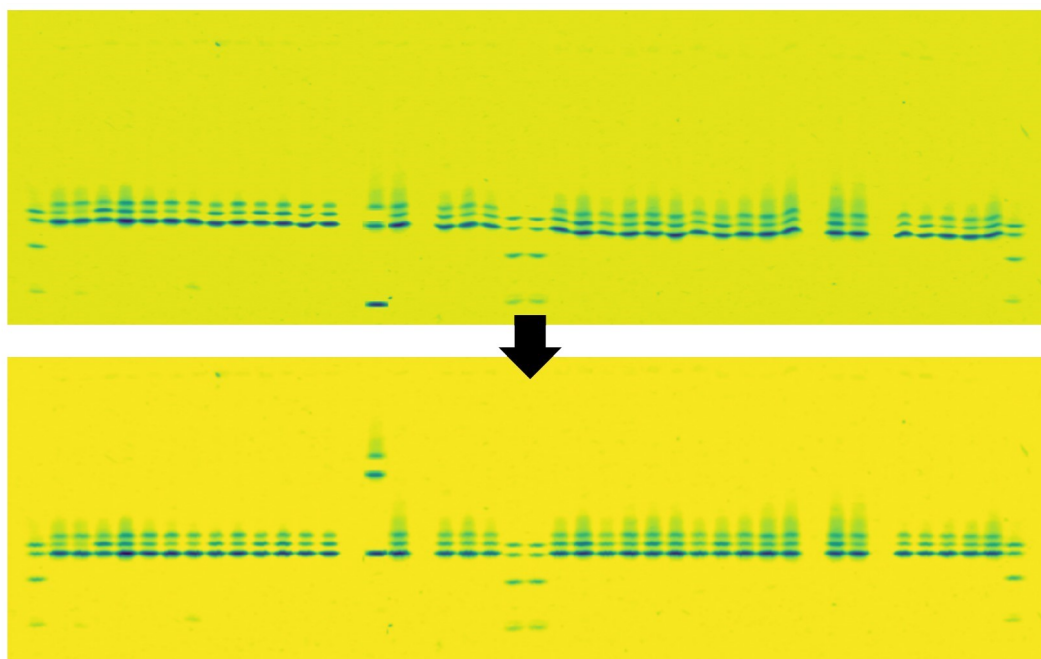


Figure 4.7: Example of an abnormal variant below HbF-band

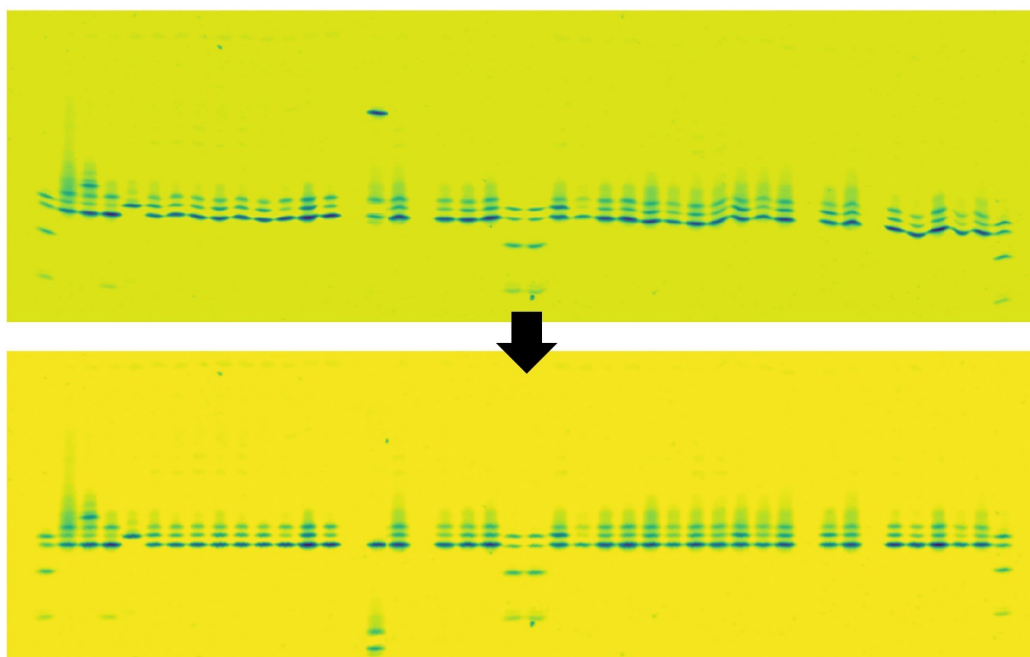


Figure 4.8: Example of an abnormal variant above HbF-band

5 Discussion

5.1 Performance test

From Table 4.1 it is clear that most of the images were straightened correctly and can be interpreted after straightening. Figures 5.1 and 5.2 show two examples belonging to this category. These are shown as examples that the proposed method functions as intended even with samples that contain abnormal variants, as long as the HbA- or the HbF-band has the greatest intensity in the sample. In Figure 5.1, two samples with abnormal variants are circled. First one of them has a variant that is not in the same lane as the control sample's HbS- and HbC-bands, so it must be some other variant. The second sample has two abnormal variants that can be identified as HbS and HbC. It also has low quantities of HbA, so the sample could be identified as SCD-SC. Figure 5.2 has an abnormal variant band above the normal hemoglobins, so the sample could be identified as some form of alpha-thalassemia, where the level of Hb Bart's has been elevated.

Out of the 103 images categorized with low-level distortion, 102 were straightened correctly and can be interpreted. Also, 25 out of 28 images with medium-level distortion and 6 out of 11 images with high-level distortion were in this category. It can be concluded that the level of distortion affects the success of the straightening and the interpretability after the straightening.

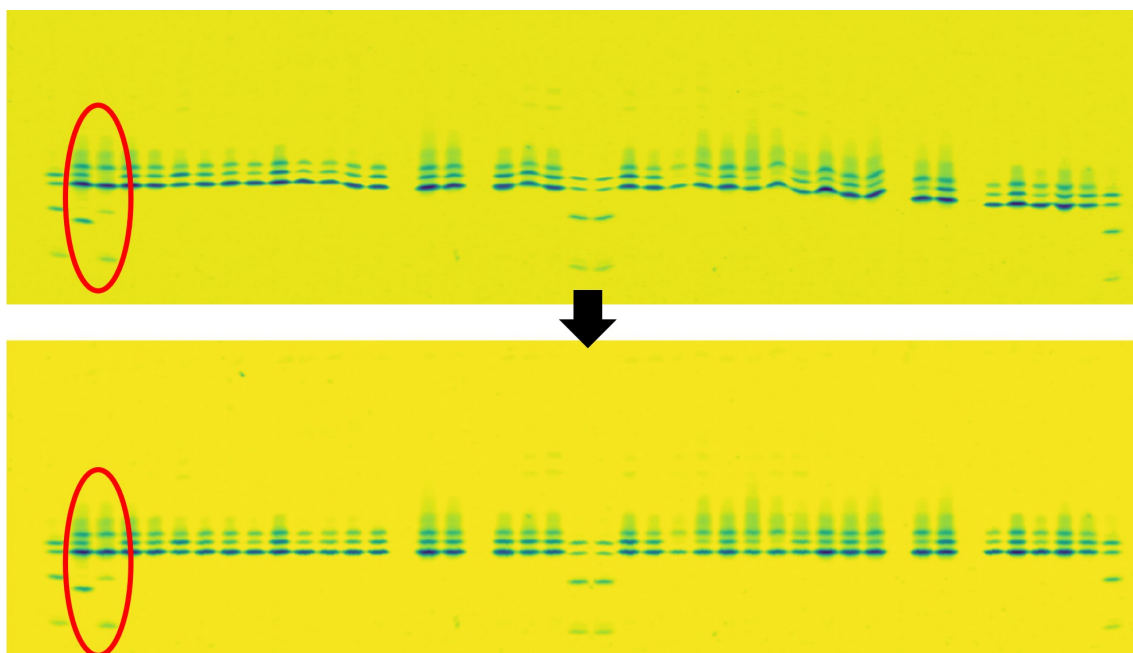


Figure 5.1: Example of an image that is straightened without errors and interpretable. Two samples containing abnormal variants circled.

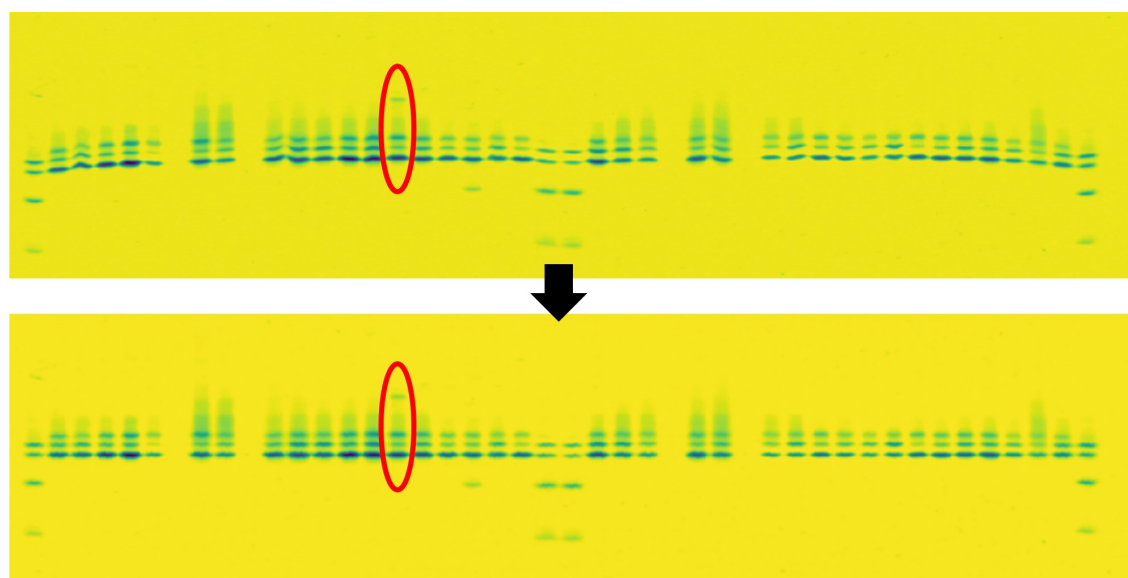


Figure 5.2: Example of an image that is straightened without errors and interpretable. A sample with Hb Bart's circled.

Six of the correctly straightened images cannot be interpreted after the straightening. Three of those had medium-level distortions and three had high-level distortions. Images are in this category for two reasons. First reason is that some samples have very curved bands, that even after forcing them straight, they do no longer represent a sample with good quality. Figure 5.3 shows an example of a gel, with high-level distortion, that has extremely curved samples, which even after straightening, are uninterpretable.

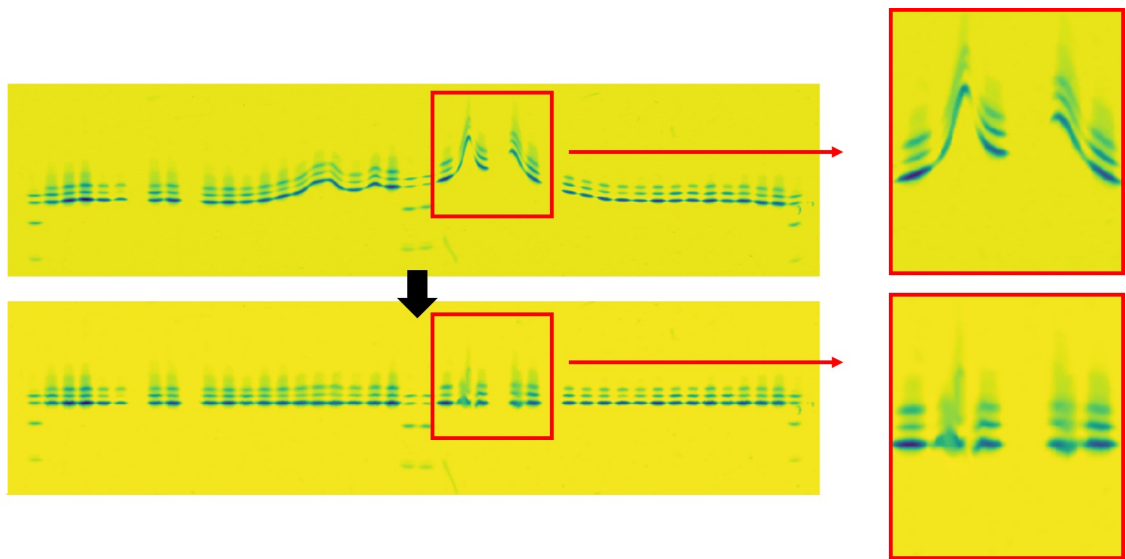


Figure 5.3: Example of an extremely curved band that even after straightening is not interpretable

Second reason is that the distortion in the gel changes the control samples' bands relative distance, therefore they cannot be used as a reference location when identifying hemoglobin variants. Figure 5.4 shows a typical example of this, where the bending of the whole sample area changes the relative distance between the HbS- and HbC-bands in the control sample at position 20. The quality of the next patient sample at position 21 is also lowered. Both gel errors represent extreme situations where the gel run would have to be repeated, at least for the skewed samples. The

straightening method, however, enables partial interpretation to be done for the samples that are on the same gel side and are not skewed.

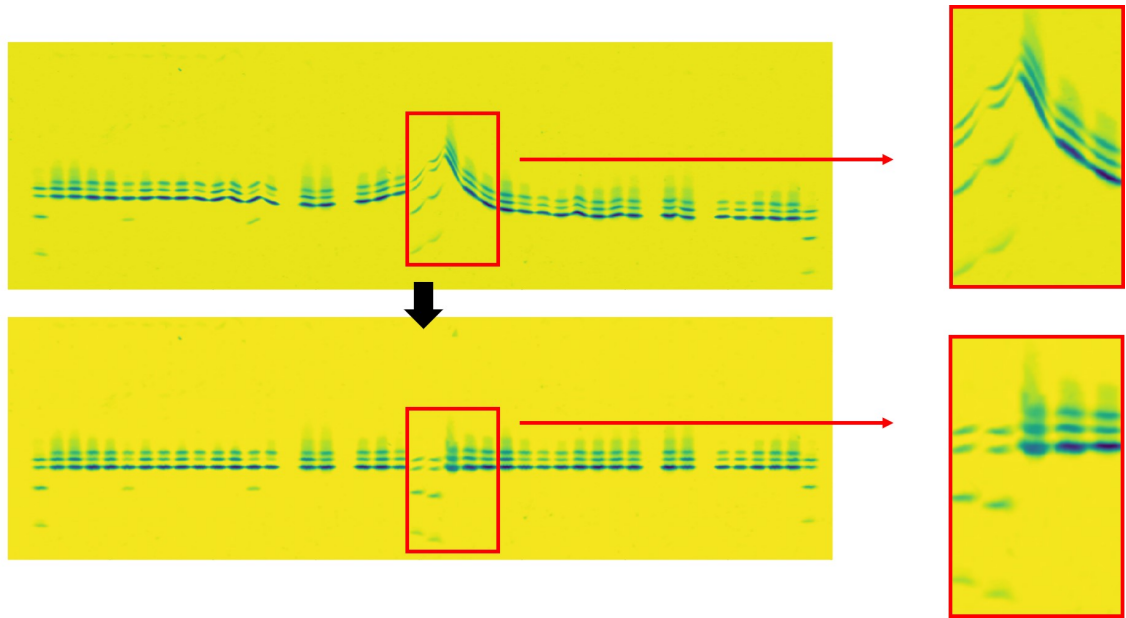


Figure 5.4: Example of change in the relative distance of the control sample's bands caused by distortion

One image, shown in Figure 5.5, was straightened with errors, but it was still categorized as interpretable. This is because even though the first control sample remains curved after the straightening, the rest of the samples are straightened okay. The error is caused by the dust artifact above the HbS-band. When comparing the distances between the four highest peaks in subsequent columns, the method starts to trace the artifact instead of tracing a single band. After the artifact ends, the four highest peaks are too far away in other bands, so the method selects the previous peak as the next peak, for the rest of the sample. The chosen points in the first control sample are shown in Figure 5.6.

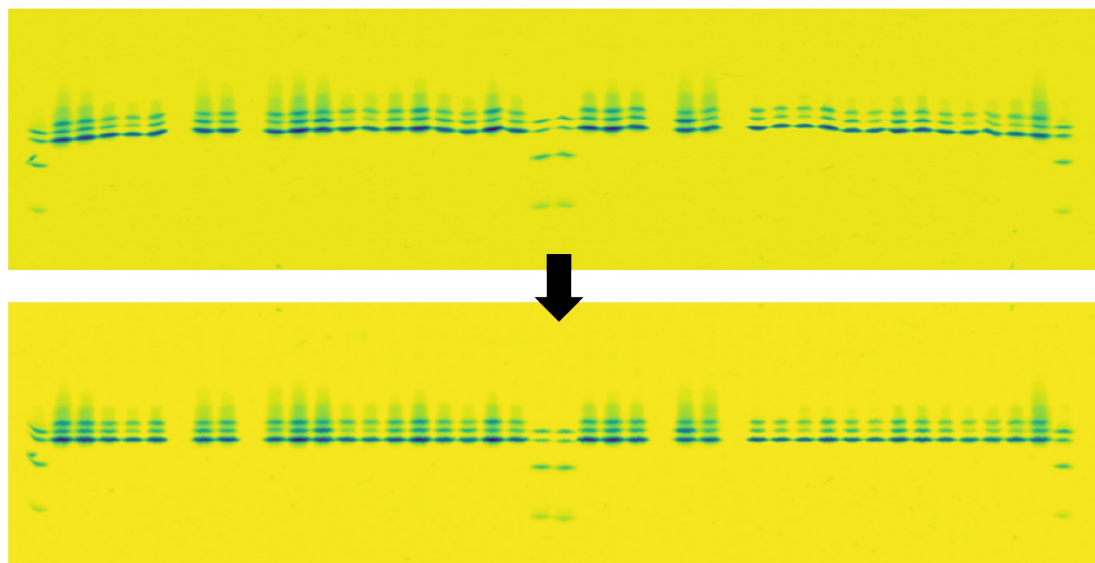


Figure 5.5: Example of an error in the straightening of the first control sample caused by an artifact

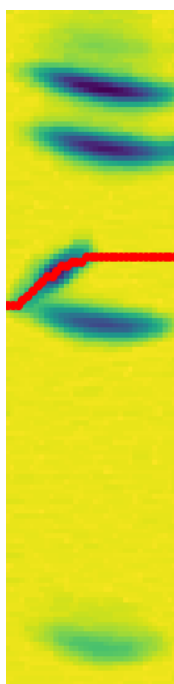


Figure 5.6: Error in band detection in the first control sample of Figure 5.5

Two images were straightened with errors and were not interpretable. Figure 5.7 shows first of them. From left to right, the second control sample is still curved after straightening, due to errors in the nearest peak detection in the straightening of bands method. The error is caused by the fact that one pixel in the band, that is being traced, is not one of the four highest peaks, and because the four highest peaks are too far away, the method sets the next point to the same as the previous one, similarly to the example in Figure 5.6. The chosen points for the sample are shown in Figure 5.8. Even if that control sample would have been straightened correctly, the other control sample in the middle has altered distance between the bands.

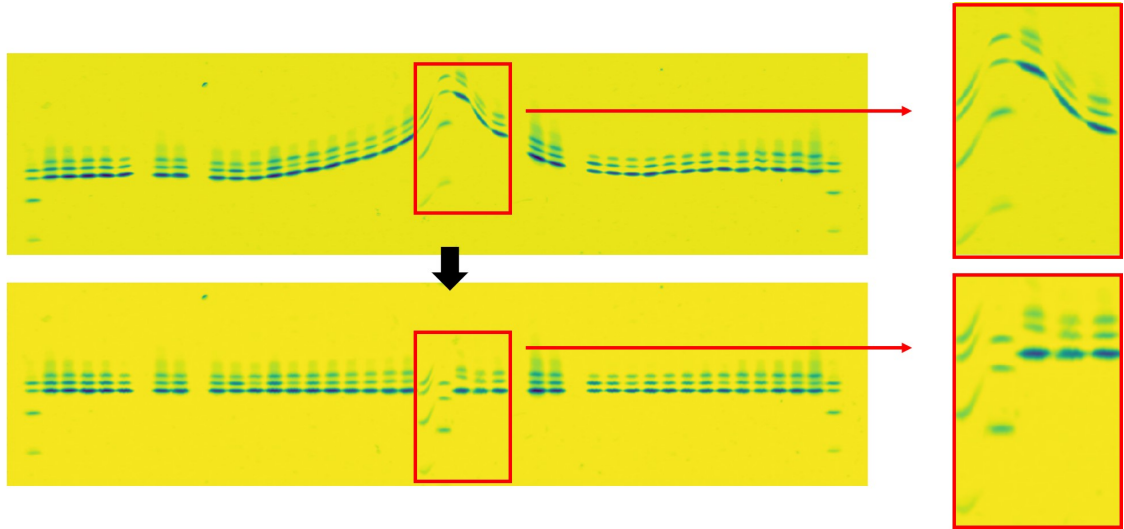


Figure 5.7: Example of an image that is straightened with errors and not interpretable

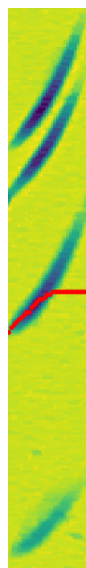


Figure 5.8: Error in band detection in the second control sample of Figure 5.7

The second image in this category, shown in Figure 5.9, cannot be interpreted or straightened, because of the poor quality of the gel, caused by mistakes in the manufacturing of the gel. The few last samples have blended together, so they cannot be interpreted. Also, the given x-coordinates for the positions of those samples are not accurate, so the straightening method does not function as intended.

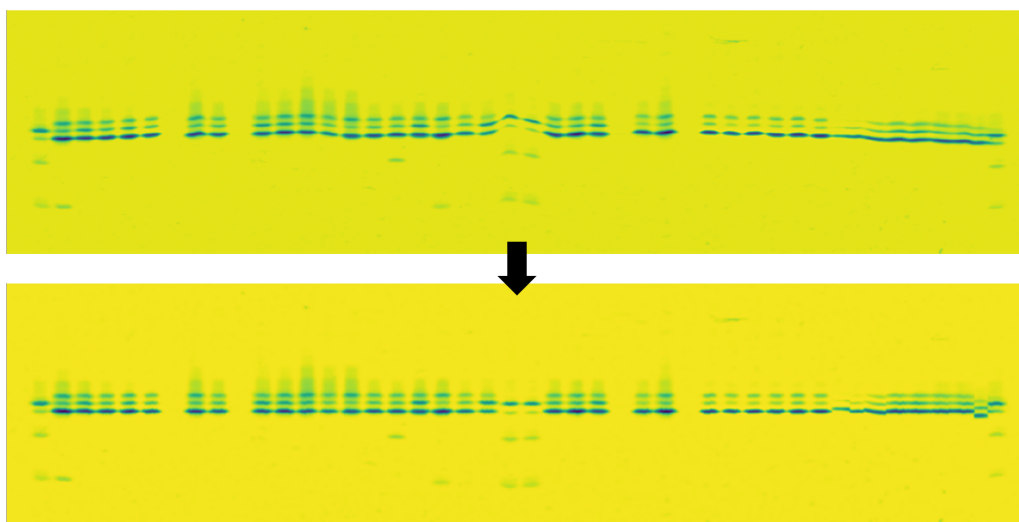


Figure 5.9: Example of a low-quality gel that has samples blended together

5.2 Comparison of band quantities

From the Table 4.2 and Figure 4.2, it is clear that most of the control samples' bands retained the relative quantities after the straightening. The few outliers, with a large increase or decrease in the difference, are caused by the peak area calculation process rather than the straightening method itself. When a sample has very curved bands, the 1D-signal's four peaks no longer represent the bands, since the bending causes the peaks to go on top of each other and lose information in the 1D-signal. Figure 5.10 shows an example of a very curved sample, whose 1D-signal's calculated peak areas are not accurate representation of the peaks' actual quantities.

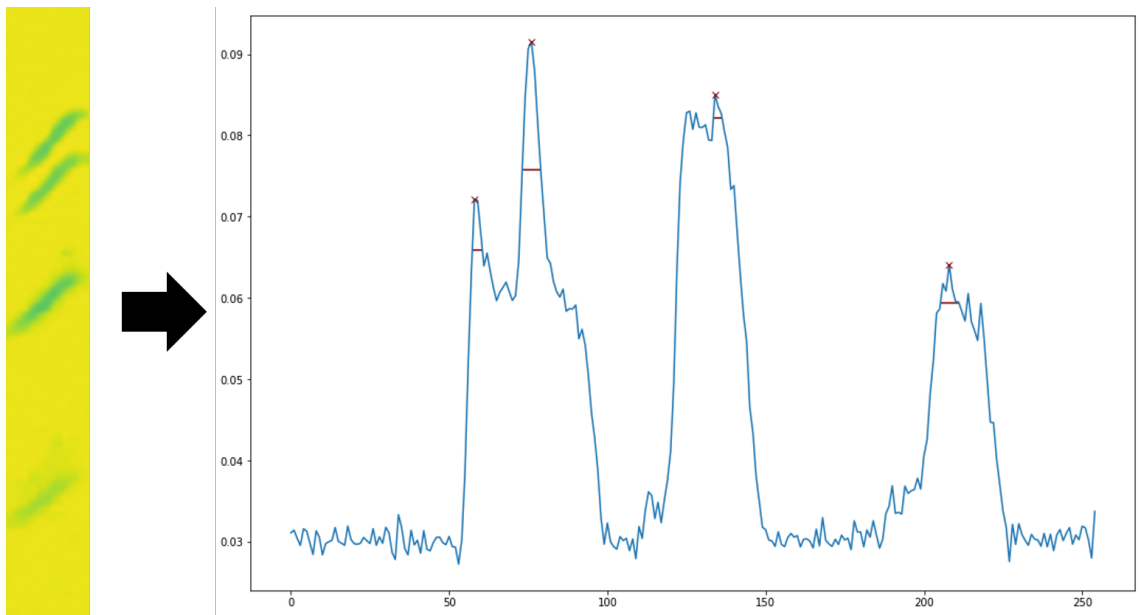


Figure 5.10: Example of a curved sample that causes errors in the peak area calculations

5.3 Stability test

Figure 4.4 showed that the method still straightened correctly after applying random Gaussian noise that has a distribution with a mean of 0 and a standard deviation of 0.03. The mean of the standard deviations of all the 142 original images backgrounds was 0.005, which means that the method still functions as intended after the standard deviation of the pixel intensities in the image is six times greater. Using 0.06 as the standard deviation in the distribution of the random Gaussian noise, which is 12 times greater than the original standard deviation of the backgrounds, results in errors in the band straightening, while the alignment of samples still functions without errors. This is shown in Figure 5.11. The errors are due to the four highest peaks not being in the bands anymore, since the noise increases the intensity of the background randomly.

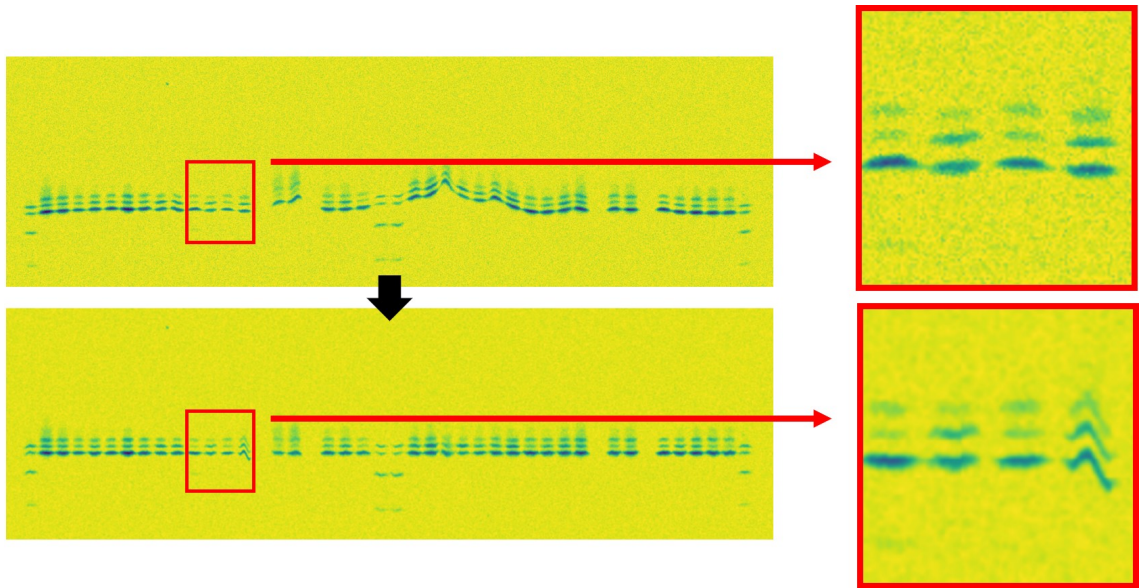


Figure 5.11: Example of a gel with added random Gaussian noise that has a distribution with a mean of 0 and standard deviation of 0.06

Figure 4.5 with empty gel and Figure 4.6 with only one control sample in the middle were straightened without issues. This is because the straightening of curved bands is done independently for each sample, so the empty positions in the gel do not affect the process. Also, the alignment of the samples is done with only one control sample, which can be located anywhere in the image, as long as it is of good quality.

In Figures 4.7 and 4.8, the method misaligned the samples that had some other band than HbF- or HbA-band having the greatest intensity. In both examples, the abnormal variant was falsely aligned with the HbF-band. This was expected since the method aligns each sample's greatest intensity band with the HbF-band of the first control sample and can only fix falsely aligned HbA-bands that have the greatest intensity.

The image correction metrics, introduced in Section 3.3, can be used to detect these abnormalities. Figure 5.12 shows these metrics plotted with the image having greatest intensity band below HbF-band, before and after straightening. From the second bar chart, sample 14 can be detected to have been moved up exceptionally large amounts. This can be used to flag the sample as abnormal. Figure 5.13 shows the same bar charts for the image that has a sample with the greatest intensity band above HbF-band. From the second bar chart in this figure, sample 14 can easily be visually detected moving the opposite direction as the other samples.

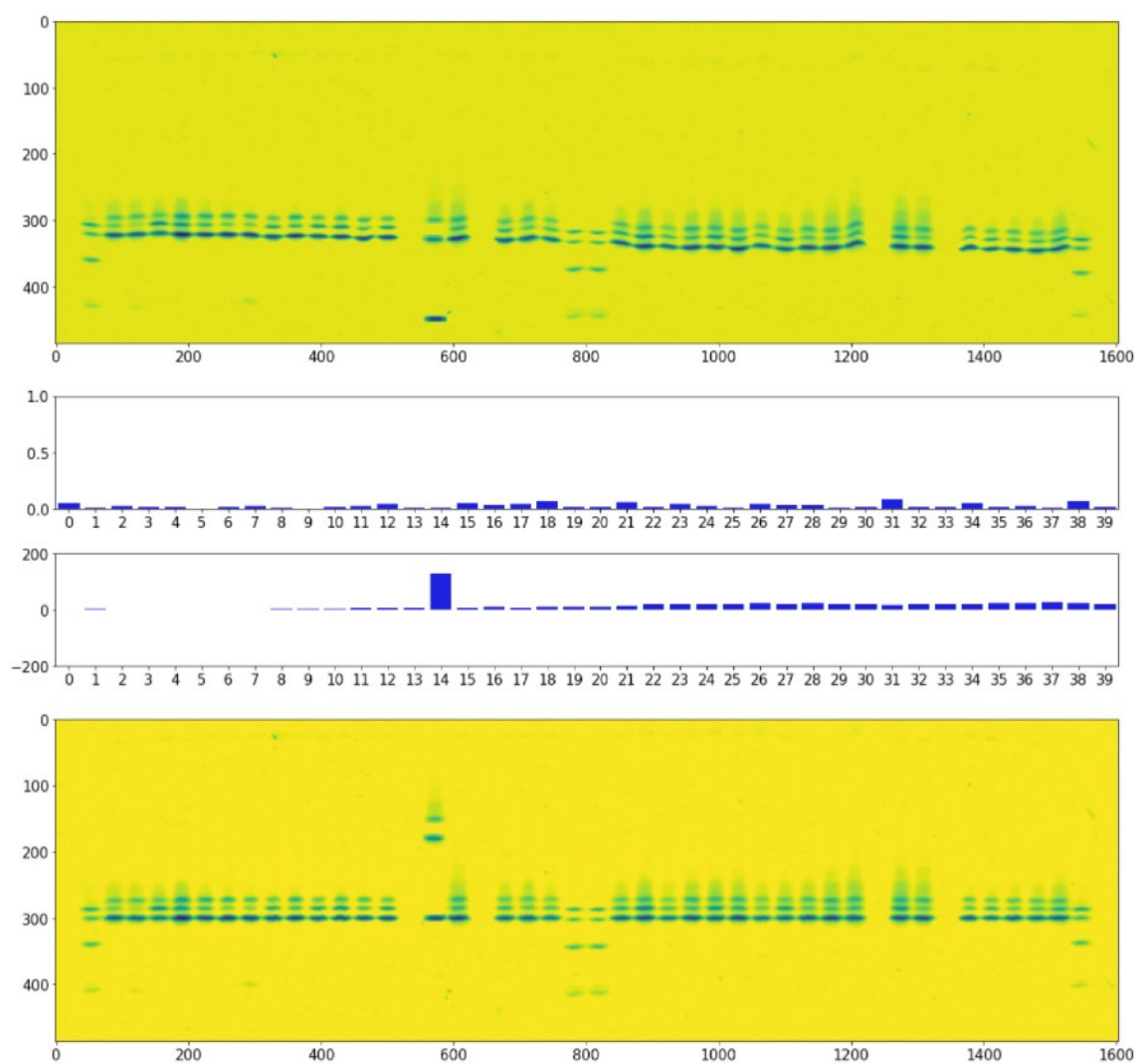


Figure 5.12: Image correction metrics used to detect the abnormal variant below the HbF-band in sample 14

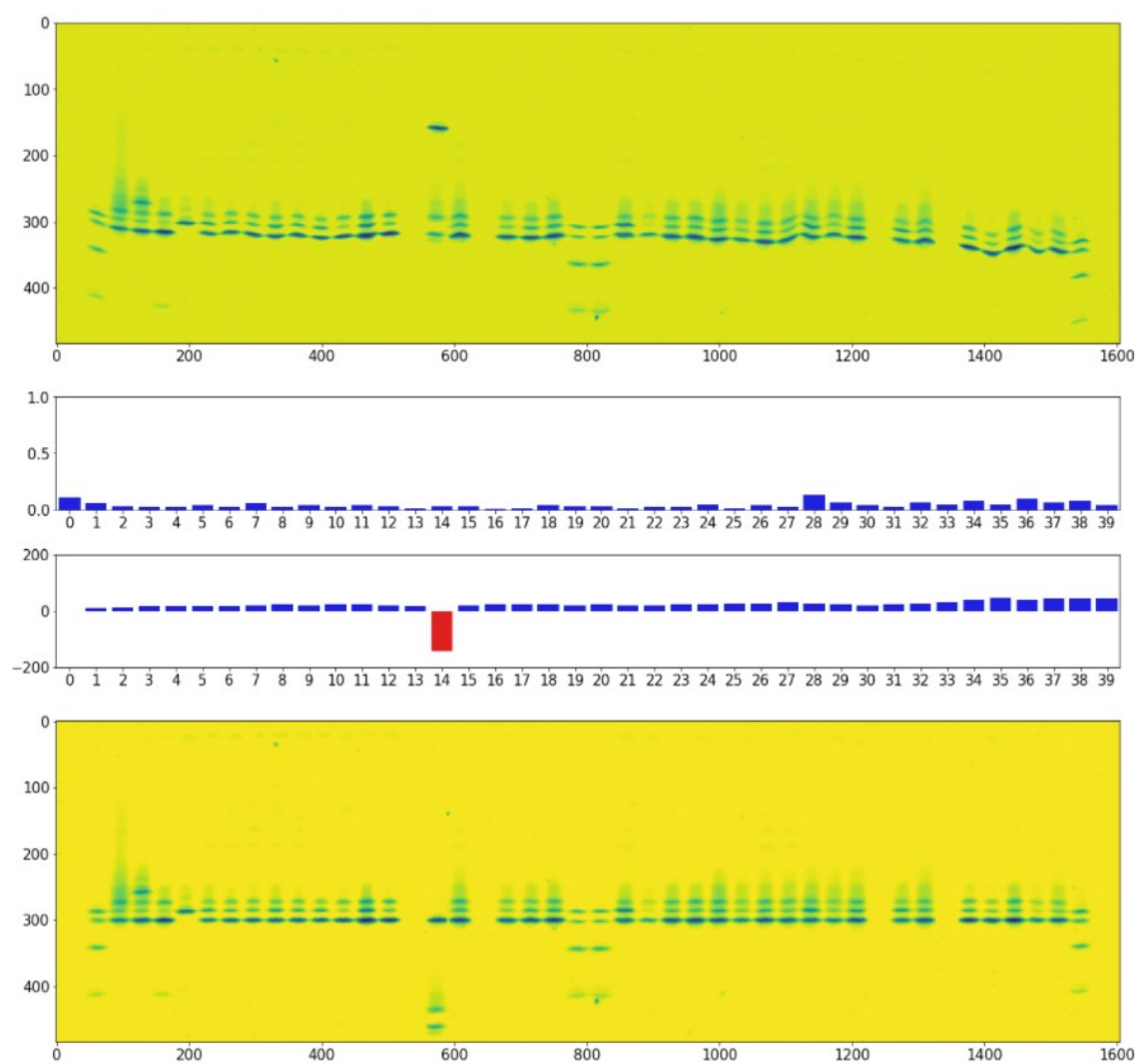


Figure 5.13: Image correction metrics used to detect the abnormal variant above the HbF-band in sample 14

6 Conclusions

Hemoglobinopathies are hemoglobin disorders, which either produce abnormal hemoglobin variants or affect the production of normal hemoglobins. Most hemoglobinopathies are asymptomatic, but some of them are clinically significant and can be fatal. It is approximated that 7% of world population carry some form of hemoglobinopathy. Newborn screening is important for detecting these disorders early, to ensure the right treatment.

The most used method for screening hemoglobinopathies in newborns is isoelectric focusing (IEF). IEF separates the different hemoglobin molecules in the blood into sharp distinct bands by their isoelectric point using a pH gradient. Samples with abnormal variants or changes in the quantities of normal hemoglobins can easily be visually detected when comparing to a good control sample. However, the visual interpretation process becomes difficult when the IEF gel has distortions. The distortions are caused by multiple reasons related to the manufacturing of the gel or the laboratory environment. The goal of this thesis was to propose a method for straightening images of these distorted gels.

The proposed method consists of two phases. First, all bands within each sample are straightened individually, and then each sample is aligned with the first control sample. Out of the 142 images used in the performance test, the method managed to straighten 139 of them successfully and only three had errors. The performance test showed that the method functions well, especially with good quality gels, but also

with gels that have high distortions. However, even if the method functions correctly, the physical distortions in the gel can be so severe that they can cause samples to lose image quality to a point that they are unusable, even after straightening.

The peak area comparison from the control samples' 1D-signals before and after straightening showed that the method retained the relative band quantities. This is important since the intensity of the bands is crucial when identifying different hemoglobin variants, and therefore they need to remain unchanged during the straightening process.

The stability test showed the weaknesses of the method. The method cannot correctly align samples that have some other band than HbF or HbA with the highest intensity. Therefore, these bands are falsely aligned with the HbF-band. At the moment, the only method to detect these occurrences is manually detecting them from the visualizations of the image correction metrics. Luckily, these cases are very rare and most of the samples in a newborn screening application are normal newborn with HbF-band as the greatest intensity or some composition where HbA-band has the greatest intensity.

Limitation for this thesis was relatively small and unilateral dataset. The method was optimized to function with this particular dataset and might fail with some rare cases that were not covered in the stability test. Also, more difficult distortions, not present in the dataset, might cause issues for the method.

Tests could be done in the future with a larger dataset of images, to truly assess the functionality and capabilities of the method. Also, the method could be expanded to work with other than IEF gel images as well. For example, similar distortions can be found with other protein gel electrophoresis as well [34].

The method proposed in this thesis enables more accurate quantitation of the bands and therefore relieves the trained experts work when identifying different hemoglobinopathies from IEF gels in a newborn screening environment.

References

- [1] J. M. Benson and B. L. Therrell, “History and current status of newborn screening for hemoglobinopathies”, *Seminars in Perinatology*, vol. 34, no. 2, pp. 134–144, 2010, Newborn Screening, ISSN: 0146-0005. DOI: <https://doi.org/10.1053/j.semperi.2009.12.006>.
- [2] G. R. Serjeant, “The natural history of sickle cell disease”, *Cold Spring Harbor Perspectives in Medicine*, vol. 3, no. 10, 2013. DOI: [10.1101/cshperspect.a011783](https://doi.org/10.1101/cshperspect.a011783).
- [3] D. P. Kaufman, *Physiology, fetal hemoglobin*, Mar. 2021. [Online]. Available: <https://www.ncbi.nlm.nih.gov/books/NBK500011/>, (accessed: 12.10.2021).
- [4] A. Dasgupta and A. Wahed, “Chapter 23 - hemoglobinopathies and thalassemia”, in *Clinical Chemistry, Immunology and Laboratory Quality Control (Second Edition)*, A. Dasgupta and A. Wahed, Eds., Second Edition, Elsevier, 2021, pp. 457–487, ISBN: 978-0-12-815960-6. DOI: <https://doi.org/10.1016/B978-0-12-815960-6.00005-4>.
- [5] M. N. Reddy and R. A. Francios, “Rapid quantitation of hemoglobin S by isoelectric focusing”, *Annals of clinical and laboratory science*, vol. 24, no. 5, pp. 401–406, 1994. [Online]. Available: <https://pubmed.ncbi.nlm.nih.gov/7529472/>.

- [6] L. Farmerie, R. R. Rustandi, J. W. Loughney, and M. Dawod, “Recent advances in isoelectric focusing of proteins and peptides”, *Journal of Chromatography A*, vol. 1651, p. 462 274, 2021, ISSN: 0021-9673. DOI: <https://doi.org/10.1016/j.chroma.2021.462274>.
- [7] PerkinElmer, *Complete solutions for newborn screening*, Brochure, Aug. 2019. [Online]. Available: https://rh.perkinelmer.com/wp-content/uploads/2021/06/Newborn-screening-complete-solutions-brochure_1244-1216-20_16.8.2019-3.pdf, (accessed: 01.10.2021).
- [8] —, “PerkinElmer Finland”, [Online]. Available: <https://www.perkinelmerfinland.fi/>, (accessed: 05.10.2021).
- [9] —, (). “Migele™ Gel Electrophoresis Unit: Detection of hemoglobinopathies in newborns”, [Online]. Available: <https://rh.perkinelmer.com/products/migele-gel-electrophoresis-unit/#Instrument>. (accessed: 01.10.2021).
- [10] A. J. Marengo-Rowe, “Structure-function relations of human hemoglobins”, *Baylor University Medical Center Proceedings*, vol. 19, no. 3, pp. 239–245, 2006, PMID: 17252042. DOI: 10.1080/08998280.2006.11928171.
- [11] Centers for Disease Control and Prevention and Association of Public Health Laboratories. (Dec. 2015). “Hemoglobinopathies: Current practices for screening, confirmation and follow-up”, [Online]. Available: https://www.cdc.gov/ncbddd/sicklecell/documents/nbs_hemoglobinopathy-testing_122015.pdf. (accessed: 05.10.2021).
- [12] D. Weatherall and J. Clegg, “Inherited haemoglobin disorders: An increasing global health problem”, *Bulletin of the World Health Organization*, vol. 79, no. 8, pp. 704–712, Feb. 2001. [Online]. Available: <https://apps.who.int/iris/handle/10665/268402>.

-
- [13] E. Kohne, “Hemoglobinopathies”, *Deutsches Ärzteblatt International*, vol. 108, no. 31-32, pp. 532–540, 2011. DOI: 10.3238/arztebl.2011.0532.
- [14] R. Galanello and A. Cao, “Alpha-thalassemia”, *Genetics in Medicine*, vol. 13, no. 2, pp. 83–88, Jan. 2011. DOI: 10.1097/gim.0b013e3181fcb468.
- [15] R. Galanello and R. Origa, “Beta-thalassemia”, *Orphanet Journal of Rare Diseases*, vol. 5, no. 1, May 2010. DOI: 10.1186/1750-1172-5-11.
- [16] C. S. Thom, C. F. Dickson, D. A. Gell, and M. J. Weiss, “Hemoglobin variants: Biochemical properties and clinical correlates”, vol. 3, no. 3, a011858–a011858, Feb. 2013. DOI: 10.1101/cshperspect.a011858.
- [17] L. de Souza Torres, J. V. Okumura, D. G. H. da Silva, and C. R. Bonini-Domingos, “Hemoglobin D-Punjab: Origin, distribution and laboratory diagnosis”, *Revista Brasileira de Hematologia e Hemoterapia*, vol. 37, no. 2, pp. 120–126, 2015, ISSN: 1516-8484. DOI: <https://doi.org/10.1016/j.bjhh.2015.02.007>.
- [18] P. Basset, Y. Beuzard, M. Garel, and J. Rosa, “Isoelectric focusing of human hemoglobin: its application to screening, to the characterization of 70 variants, and to the study of modified fractions of normal hemoglobins”, *Blood*, vol. 51, no. 5, pp. 971–982, May 1978, ISSN: 0006-4971. DOI: 10.1182/blood.V51.5.971.971.
- [19] B. Modell, “Global epidemiology of haemoglobin disorders and derived service indicators”, vol. 2008, no. 6, pp. 480–487, Jun. 2008. DOI: 10.2471/blt.06.036673.
- [20] V. M. Pinto, M. Balocco, S. Quintino, and G. L. Forni, “Sickle cell disease: A review for the internist”, vol. 14, no. 7, pp. 1051–1064, Aug. 2019. DOI: 10.1007/s11739-019-02160-x.

- [21] A. C. Allison, “Protection afforded by sickle-cell trait against subtertian malarial infection”, vol. 1, no. 4857, pp. 290–294, Feb. 1954. DOI: 10.1136/bmj.1.4857.290.
- [22] M. Bender, *Sickle cell disease*, 2021. [Online]. Available: <https://www.ncbi.nlm.nih.gov/books/NBK1377/>, (accessed 14.10.2021).
- [23] J. Howard and P. Telfer, “Sickle cell disease in clinical practice”, in *Sickle cell disease in clinical practice*, ser. In Clinical Practice. Springer, 2015, ch. 1, pp. 3–28, ISBN: 9781447124733.
- [24] S. Dror, “Clinical and hematological features of homozygous hemoglobin O-Arab [beta 121 Glu -> Lys]”, vol. 60, no. 3, pp. 506–507, Nov. 2012. DOI: 10.1002/pbc.24414.
- [25] T. Isbir, D. Kirac, B. Demircan, and B. Dalan, “Gel electrophoresis”, in *Brenner’s Encyclopedia of Genetics (Second Edition)*, S. Maloy and K. Hughes, Eds., Second Edition, San Diego: Academic Press, 2013, pp. 165–167, ISBN: 978-0-08-096156-9. DOI: <https://doi.org/10.1016/B978-0-12-374984-0.00580-5>.
- [26] A. Tiselius, “A new apparatus for electrophoretic analysis of colloidal mixtures”, *Trans. Faraday Soc.*, vol. 33, pp. 524–531, 0 1937. DOI: 10.1039/TF9373300524.
- [27] H. Wajcman and K. Moradkhani, “Abnormal haemoglobins: Detection & characterization”, *Indian J. Med. Res.*, vol. 134, pp. 538–546, Oct. 2011.
- [28] P. G. Righetti, “Isoelectric focusing: Theory, methodology and applications”, in. Elsevier Biomedical Press, 1983, ch. 1, pp. 1–86.
- [29] C. Frömmel, “Newborn screening for sickle cell disease and other hemoglobinopathies: A short review on classical laboratory meth-

- ods—isoelectric focusing, hplc, and capillary electrophoresis”, *International Journal of Neonatal Screening*, vol. 4, no. 4, 2018. DOI: 10.3390/ijns4040039.
- [30] P. T. McGann and C. Hoppe, “The pressing need for point-of-care diagnostics for sickle cell disease: A review of current and future technologies”, *Blood cells, molecules & diseases*, vol. 67, pp. 104–113, Sep. 2017. DOI: 10.1016/j.bcmd.2017.08.010.
- [31] A. Achour, T. T. Koopmann, F. Baas, and C. L. Harteveld, “The evolving role of next-generation sequencing in screening and diagnosis of hemoglobinopathies”, *Frontiers in Physiology*, vol. 12, p. 1025, 2021. DOI: 10.3389/fphys.2021.686689.
- [32] Y. Liu and J. Heer, “Somewhere over the rainbow: An empirical assessment of quantitative colormaps”, in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, 2018, pp. 1–12, ISBN: 9781450356206. DOI: 10.1145/3173574.3174172.
- [33] B. W. Adam, D. L. Chafin, and V. R. De Jesús, “Stabilities of hemoglobins A and S in dried blood spots stored under controlled conditions”, *Clinical Biochemistry*, vol. 46, no. 12, pp. 1089–1092, 2013, ISSN: 0009-9120. DOI: <https://doi.org/10.1016/j.clinbiochem.2013.05.043>.
- [34] M. Starita-Geribaldi, A. Houri, and P. Sudaka, “Lane distortions in gel electrophoresis patterns”, *Electrophoresis*, vol. 14, no. 1, pp. 773–781, 1993. DOI: 10.1002/elps.11501401121.