

University of South Bohemia



Faculty of Agriculture



Dissertation Thesis

Mgr. Jan Říha

2012

University of South Bohemia
Faculty of Agriculture

Department of Genetics, Animal Breeding and Nutrition

Microsatellite Markers
and
Genetic Diversity Issues in Cattle

České Budějovice 2012

Mgr. Jan Říha

Student: Mgr. Jan Říha

Supervisor: Prof. Ing. Jindřich Čítek, CSc.

Specialist Supervisor: Prof. Ing. Josef Dvořák, CSc., dr. h. c.

Doctoral Study Programme: Biotechnologies

Doctoral Study Branch: Agricultural Biotechnologies

Presented dissertation thesis was compiled in years 2009 – 2012 at the department of Genetics, Animal Breeding and Nutrition, Faculty of Agriculture, University of South Bohemia. I declare, that I created this thesis by myself and I quoted and used properly all of materials, papers and another sources used for this work.

Acknowledgments :

Thanks to Prof. Ing. Jindřich Čítek, CSc. and Prof. Ing. Josef Dvořák, CSc., dr. h. c. for their enthusiastic supervision, guidance and support.

Whoever speculates upon four things, a pity for him! He is as though he had not come into the world, what is above, what is beneath, what before, what after. (Talmud, Chagigah 1:2)

Table of Contents

| | |
|---|-----------|
| 1 General Introduction | 11 |
| 2 Objectives of the Thesis | 12 |
| 3 Literature Review | 13 |
| 3.1 Evolution of Cattle | 13 |
| 3.2 Bovine genome | 13 |
| 3.3 Breeds characterisation | 14 |
| 3.3.1 Charolais | 14 |
| 3.3.2 Aberdeen Angus | 14 |
| 3.3.3 Hereford | 15 |
| 3.3.4 Holstein | 15 |
| 3.3.5 Limousin | 16 |
| 3.3.6 Piedmontese | 16 |
| 3.3.7 Simmental | 17 |
| 3.3.8 Fleckvieh cattle | 17 |
| 3.4 Population genetic measures | 19 |
| 3.4.1 Basic genetic diversity measures | 19 |
| 3.4.2 Heterozygosity | 19 |
| 3.4.3 The Hardy–Weinberg equilibrium | 20 |
| 3.4.4 Linkage disequilibrium | 20 |
| 3.4.5 Effective population size | 21 |
| 3.4.6 Genetic distances | 21 |
| 3.4.7 Wright’s inbreeding coefficient and F-statistics | 23 |
| 3.4.8 Microsatellite panel effectiveness measures | 24 |
| 3.5 Population genetic structure using machine learning methods | 26 |
| 3.5.1 Machine learning – state of art | 26 |
| 3.5.2 Classification | 28 |
| 3.5.2.1 ZeroR | 29 |
| 3.5.2.2 Decision trees | 29 |
| 3.5.2.3 Decision rules | 31 |
| 3.5.2.4 Naive Bayes classifier | 33 |
| 3.5.2.5 BayesNet classifier | 34 |
| 3.5.2.6 Instance based classifiers | 35 |
| 3.5.2.7 MCMC algorithm | 36 |
| 3.5.2.8 Neural Networks | 37 |
| 3.5.2.9 Support Vector Machines | 38 |
| 3.5.2.10 Metalearning algorithms | 39 |
| 3.5.2.11 Bagging | 40 |
| 3.5.2.12 Voting | 40 |
| 3.5.2.13 Evaluation of classification models | 40 |
| 3.5.3 Clusterization | 42 |
| 3.5.3.1 Hierarchical clustering | 43 |
| 3.5.3.2 K-means algorithm | 44 |
| 3.5.3.3 EM algorithm | 44 |
| 3.5.3.4 Markov chains | 45 |
| 3.6 Microsatellites in Cattle Studies | 47 |
| 3.7 Diversity in Studied Cattle Breeds | 48 |

| | | |
|-------------|--|-----------|
| 3. 7. 1 | Hereford | 48 |
| 3. 7. 2 | Holstein | 49 |
| 3. 7. 3 | Piedmontese | 49 |
| 3. 7. 4 | Simmental | 50 |
| 3. 7. 5 | Fleckvieh | 50 |
| 3. 7. 6 | Limousin, Charolais, Aberdeen Angus | 50 |
| 4 | Material and Methods | 52 |
| 4. 1 | Microsatellite loci | 52 |
| 4. 2 | Datasets | 52 |
| 4. 2. 1 | General Dataset - Purebred Individuals | 52 |
| 4. 2. 2 | Crossbred dataset | 53 |
| 4. 2. 3 | Machine learning datasets | 53 |
| 4. 2. 3. 1 | Genotype dataset | 54 |
| 4. 2. 3. 2 | Allele-length dataset | 54 |
| 4. 2. 3. 3 | Allele-frequency dataset | 54 |
| 4. 3 | Used methods | 55 |
| 4. 3. 1 | Description of the genetic diversity and characterisation of the selected cattle breeds in the Czech Republic | 55 |
| 4. 3. 1. 1 | Number of observation | 55 |
| 4. 3. 1. 2 | Availability | 56 |
| 4. 3. 1. 3 | Within-population inbreeding coefficient | 56 |
| 4. 3. 1. 4 | Allele and genotype frequencies | 56 |
| 4. 3. 1. 5 | Observed heterozygosity (H_0) | 56 |
| 4. 3. 1. 6 | Expected heterozygosity (H_e) | 56 |
| 4. 3. 1. 7 | Genetic distances | 57 |
| 4. 3. 1. 8 | Euclidean distance | 57 |
| 4. 3. 1. 9 | Nei's standard genetic distance | 57 |
| 4. 3. 1. 10 | Goldstein distance | 58 |
| 4. 3. 1. 11 | Slatkin ASD distance | 58 |
| 4. 3. 1. 12 | Shriver DSW distance | 58 |
| 4. 3. 1. 13 | Shared Allele Distance | 58 |
| 4. 3. 2 | Estimation and validation of paternity testing by microsatellite loci in selected cattle breeds | 59 |
| 4. 3. 2. 1 | Polymorphism information content | 59 |
| 4. 3. 2. 2 | Paternity exclusion (PE1) | 59 |
| 4. 3. 2. 3 | Paternity exclusion - one parental genotype unavailable (PE2) | 59 |
| 4. 3. 2. 4 | Parentage exclusion (PE3) | 60 |
| 4. 3. 2. 5 | Combined Exclusion Probability | 60 |
| 4. 3. 3 | Creation of the software support for routine genotyping of microsatellite loci under the reference laboratory conditions | 60 |
| 4. 3. 4 | Proving of usability of machine learning methods in cattle breed discrimination task | 61 |
| 4. 3. 4. 1 | G-metric classifier | 62 |
| 4. 3. 4. 2 | Evaluation of classification models | 63 |
| 5 | Results and Discussion | 64 |
| 5. 1 | Description of the genetic diversity and characterisation of selected cattle breeds in the Czech Republic | 64 |
| 5. 1. 1 | Summary results of genetic variability for microsatellite data by breeds | 64 |

| | | |
|-------------|--|-----|
| 5. 1. 1. 1 | General dataset | 64 |
| 5. 1. 1. 2 | Crossbred dataset | 65 |
| 5. 1. 1. 3 | Czech Fleckvieh | 66 |
| 5. 1. 1. 4 | Aberdeen Angus | 67 |
| 5. 1. 1. 5 | Holstein | 68 |
| 5. 1. 1. 6 | Piedmontese | 69 |
| 5. 1. 1. 7 | Blonde d'Aquitaine | 70 |
| 5. 1. 1. 8 | Czech Simmental | 71 |
| 5. 1. 1. 9 | Charolais | 72 |
| 5. 1. 1. 10 | Hereford | 73 |
| 5. 1. 1. 11 | Galloway | 74 |
| 5. 1. 1. 12 | Limousin | 75 |
| 5. 1. 1. 13 | Summary results of genetic variability for breeds | 76 |
| 5. 1. 2 | Genetic distances | 77 |
| 5. 1. 2. 1 | Euclidean and Nei 1972 genetic distance | 77 |
| 5. 1. 2. 2 | Goldstein and Shriver genetic distance | 78 |
| 5. 1. 2. 3 | Slatkin and SharedAllele genetic distance | 79 |
| 5. 1. 3 | Phylogenetic trees | 79 |
| 5. 1. 3. 1 | Euclidean NJ and UPMGA trees | 80 |
| 5. 1. 3. 2 | Nei 1972 NJ and UPGMA trees | 81 |
| 5. 1. 3. 3 | Goldstein NJ and UPGMA trees | 83 |
| 5. 1. 3. 4 | Shriver NJ and UPGMA trees | 84 |
| 5. 1. 3. 5 | Slatkin NJ and UPGMA trees | 86 |
| 5. 1. 3. 6 | Shared Allele NJ and UPGMA trees | 87 |
| 5. 2 | Estimation and validation paternity testing by microsatellite loci in selected cattle breeds | 89 |
| 5. 3 | Creation of the software support for routine genotyping of microsatellite loci under the reference laboratory conditions | 90 |
| 5. 3. 1 | Network model | 90 |
| 5. 3. 2 | Application model | 91 |
| 5. 3. 3 | Key processes | 91 |
| 5. 3. 4 | Database | 94 |
| 5. 3. 5 | Data operations and SQL queries | 95 |
| 5. 3. 6 | Algorithms | 97 |
| 5. 3. 7 | Security | 100 |
| 5. 3. 8 | GUI | 101 |
| 5. 3. 9 | Protocols | 103 |
| 5. 4 | Proving of usability of machine learning methods in cattle breed discrimination task | 106 |
| 5. 4. 1 | ZeroR | 107 |
| 5. 4. 2 | J48 | 109 |
| 5. 4. 2. 1 | General Dataset | 109 |
| 5. 4. 2. 2 | Allele Length Dataset | 110 |
| 5. 4. 2. 3 | Allele Frequency Dataset | 111 |
| 5. 4. 2. 4 | Discussion of J48 results | 112 |
| 5. 4. 3 | JRip | 113 |
| 5. 4. 3. 1 | General Dataset | 113 |
| 5. 4. 3. 2 | Allele Length Dataset | 114 |

1 General Introduction

Large use of artificial insemination and the global trade of semen have a strong influence on the genetic diversity of cattle breeds. Knowledge of the genetic dynamics in a breed is necessary for avoiding unfavourable trends, like severe reduction of genetic diversity in the population. A considerable number of genetic distance studies for several livestock species was carried out during the past decade by research teams from all over the world. Most studies are based on microsatellite loci, although a number of other polymorphic systems like protein polymorphisms, blood group, or other molecular marker systems, were used alternatively or additionally. Under the coordination of FAO, an initiative called Measurement of Domestic Animal Diversity (MoDAD) was started to provide technical recommendations for such studies in farm animals (FAO - Measurements of Domestic Animal Diversity). To define species-specific standards, the International Society for Animal Genetics (International Society for Animal Genetics) formed a FAO/ISAG advisory group on animal genetic diversity in 1995, which set up recommended species specific lists of microsatellite loci (about 30 per species) for cattle, chicken, sheep and swine to be used in diversity studies (Clave, 2003). Molecular characterization of animal genetic resources may contribute to a rational approach to (Hanotte and Jianlin, 2005) by giving a high priority to breeds that are taxonomically most distinct (Barker, 1999). Moreover, information on diversity and population structure can provide a more rational basis for making the conservation policies and for planning the genetic improvement in future.

With growth development of the informatics and routine genotyping, new problems were formulated based on genetic diversity concept. These problems are connected with large datasets manipulations, their effective evaluation, interpretation of results, creations of new algorithms for specific tasks (traceability, identification, breed discrimination, probabilistic founding of potential parents etc.). Part of mentioned problems concerning genetic diversity of microsatellites in cattle create main aim of this thesis.

2 Objectives of the Thesis

- Description of the genetic diversity and characterisation of the selected cattle breeds in the Czech Republic.
- Estimation and validation paternity testing by microsatellite loci in selected cattle breeds.
- Creation of the software support for routine genotyping of microsatellite loci under the reference laboratory conditions.
- To prove of usability of machine learning methods in cattle breed discrimination task.

3 Literature Review

3.1 Evolution of Cattle

Cattle are the most common type of large domesticated ungulates. They are a prominent modern member of the subfamily *Bovinae* and are the most widespread species of the genus *Bos*. Prehistoric cattle originated many millions of years ago in India, and by early Pleistocene times had migrated to Europe, North Africa, and the rest of Asia. Mitochondrial DNA-, allozyme-, and microsatellite-based studies have demonstrated that the main subdivision of cattle into *Bos taurus* and *B. indicus* corresponded to a deep bifurcation (200 000–1 000 000 years ago), which predates archaeological estimates of cattle domestication (roughly 12 000 year ago) (Baker and Manwell, 1980; Loftus et al., 1994; Bradley et al., 1996; Machugh et al., 1997; Troy et al., 2001). Therefore, it cannot be ruled out that the aurochs were domesticated not only in Southwest Asia, but also in Europe. Indeed, archaeological and genetic evidence suggest that modern cattle might result from two domestication events of aurochs (*Bos primigenius*) in southwest Asia, which gave rise to taurine (*Bos taurus*) and zebuine (*Bos indicus*) cattle, respectively (Loftus et al., 1994; Troy et al., 2001; Helmer et al., 2005). Studies of variation in mitochondrial DNA (mtDNA) sequences showed that these two cattle subspecies are highly diverged from each other and reflect two independent domestications in Africa (Bradley et al., 1998; Hanotte et al., 2002) and East Asia (Mannen et al., 2004) from different aurochs subspecies. However, (Troy et al., 2001) studied mtDNA diversity in modern cattle from Europe, the Near East and Africa and in extinct aurochs and determined a Near-Eastern origin in all European cattle. Similar findings were also reported by (Achilli et al., 2008) whose findings also support a single Neolithic domestication event for *B. taurus* in the Near East, 9–11 thousand years ago. Most cattle in North Eastern Asia that are classified as *Bos taurus* (Phillips, 1961) appeared in this region as domesticated between 5000 and 4000 years B.P. i.e. several thousand years after primary aurochs domestication in West Asia (Payne and Hodges, 1997). It is supposed that the domestic cattle in North Eastern Asia originated from local wild cattle or perhaps from migrants from the early domestic center of the Near East (Mannen et al., 2004). Based on the mitochondrial DNA (mtDNA) diversity studies, (Mannen et al., 1998) suggested that multiple strains of ancestral aurochs were adopted in geographically and temporally separate stages of the domestication process.

3.2 Bovine genome

The genome of the cattle (*Bos taurus*) is similar in size to the genomes of humans and other mammals, containing approximately 3 billion DNA base pairs. The breed of cattle selected for initial sequencing was Hereford, which is used in beef production. Sequencing began in December 2003 and a first draft was completed in October 2004. Sequencing of additional cattle breeds, including the Holstein, Angus, Jersey, Limousin, Norwegian Red and Brahman, allows tracking of the DNA differences among these breeds to assist in the discovery of traits improving meat and milk production and to model human diseases as well.

3.3 Breeds characterisation

3.3.1 Charolais

The Charolais breed was developed in the district around Charolles in Central France (Czech Beef Breeders Association; American International Charolais Association). The breed became established there and achieved considerable regard as a producer of highly rated meat in the markets at Lyon and Villefranche in the 16th and 17th centuries. However, it can be speculated that Charolais dates back to Roman times in ancient Italy and entered the France during invasions of Romans to France and England. First written reference to white steers appears in a French document from the year 878 A.D. Due to historical accident and political peculiarity the forebearers of today's Charolais were isolated around Charolles in east central France from the fourteenth century until 1772. This forced segregation greatly benefited the development of the Charolais breed. The Charolais strain was kept fairly pure, and of necessity, the breeders selected only the best of the white cattle. After the region was reunited with France in 1772, the Charolais cattle began moving throughout France. Two major branches of the breed ensued, the original Charolais and the Nivernais which was centered in the French province of Nievre. In 1864, a Nivernais breeder, Count Charles de Bouille, set up a breed herd book. In 1882 the Charolais breeders followed suit and began registering cattle in the province of Saône-et-Loire. To avoid pedigree confusions the two books merged in 1919 with the older Nievre Herdbook assimilating the Charolais book (Felius, 1995).

The French have long selected their cattle for size and muscling. They selected for bone and power to a greater extent than was true in the British Isles. The French breeders stressed rapid growth in addition to cattle that would ultimately reach a large size. These were men that wanted cattle that not only grew out well but could be depended upon for draft power. Little attention was paid to refinement, but great stress was laid on utility.

Charolais cattle is white or creamy white in color, but the skin carries appreciable pigmentation. Charolais is a naturally horned beef animal. But through the breeding-up program, where naturally polled breeds were sometimes used as foundation animals, polled Charolais have emerged as an important part of the breed. Charolais cattle breed has large with mature bulls weighing from 900 to well over 1100 kg and cows weighing from 600 to over 900 kg (Purdy et al., 2008).

3.3.2 Aberdeen Angus

The Aberdeen-Angus belongs to one of three distinct and well-defined breeds of polled cattle in the United Kingdom. Polled cattle apparently existed in Scotland before recorded history because the likeness of such cattle is found in prehistoric carvings of Aberdeen and Angus. Some historians feel that the Aberdeen-Angus breed and the other Scottish breeds sprang from the aboriginal cattle of the country and that the breeds as we find them today are indigenous to the districts in which they are still found. Although little is known about the early origin of the cattle that later became known as the Aberdeen-Angus breed, it is

thought that the improvement of the original stock found in the area began in the last half of the 18th century. Two strains known as Angus doddies and Buchan humlies were used in the formation of what later became known as the Aberdeen-Angus breed of cattle (doddies, humlies are mentioned as polled in the old Scottish writings) (Purdy et al., 2008).

Apparently little attention was given to the breeding of cattle before the middle of the 18th century, but in the last half of that century, crossing and recrossing these strains of cattle eventually led to a distinct breed that was not far different from either type, since the two strains were originally of rather similar type and color pattern. At the beginning of the 19th century when good herds of Shorthorn cattle were established in Scotland, the Shorthorn were used in the improvement of native stock. It is often suggested that some Shorthorn blood found its way into the Aberdeen-Angus breed prior to the time the Herd Book was closed. On the other hand, the tribes from which the Aberdeen-Angus breed were drawn were supplying England with beef cattle for generations before the beef Shorthorn was used for improvement. Aberdeen-Angus cattle breed has mature bulls weight ranging from 900 to 1050 kg and cows weigh from 550 to 700 kg ([American Angus Association](#)).

Although originally black, within the breed, there is a strain known as Red Angus that was gaining in popularity in the late 20th century, particularly for purposes of outcrossing and crossbreeding.

3.3.3 Hereford

The Hereford is one of the UK's oldest native beef breeds, originating in the County of Herefordshire in the mid 1700's as a product of necessity to produce beef for the expanding food market created by Britain's industrial revolution. To succeed in Herefordshire, farmers must have cattle which could efficiently convert their native grass to beef. There was no breed in existence at the time to fill that need, so the farmers of Herefordshire founded the beef breed that became known as Herefords selected for a high yield of beef and efficiency of production. Herefords in the 1700's and early 1800's in England were much larger than today. Gradually, the type and conformation changed to less extreme size and weight to get more smoothness, quality and efficiency. The herd book was opened in 1846 and since 1886 has been closed to any animal whose sire or dam had not previously been recorded, so for over 120 years, the purity of the breed has remained intact (Purdy et al., 2008; Felius, 1995).

3.3.4 Holstein

The Holstein cow originated in Europe. The major historical development of the well-known and highly selected Holstein breed occurred in the Netherlands, and more specifically in the two northern provinces of North Holland and West Friesland, which lay on either side of the Zuider Zee. The original stocks were the black and white animals of the Batavians and Friesians, typical of migrant European tribes who settled in the Rhine Delta region about 2000 years ago (Del Bol et al., 2001).

Nowadays, Holstein breed is used as a major milk production breed all over the world. As recent breeding methods were always applied, Holstein breed reflects as a result all of advantages (high production, precisely selected breeding animals, etc.) as well as disadvantages (spread genetic diseases, higher level of inbreeding, more costly and time consuming animal treatment, etc.) of this effort (Purdy et al., 2008; [Holstein Association USA](#)).

3.3.5 Limousin

The history of Limousin cattle may very well be as old as the European continent itself because cattle found in cave drawings estimated to be 20,000 years old in the Lascaux Cave near Montignac, France, have a striking resemblance to today's Limousin. Limousin cattle is native to the south central part of France in the regions of Limousin and Marche. As a result of their homeland environment (rugged and rolling with rocky soil and a harsh climate), Limousin cattle evolved into a breed of unusual sturdiness, health and adaptability. The lack of natural resources enabled the region to remain relatively isolated and the farmers free to develop their cattle with little outside genetic interference. During these early times, Limousin were kept as work animals in addition to their beef qualities. Once in the 1700s and again in the mid-1800s, an attempt was made by a small number of French Limousin breeders to crossbreed their cattle in hopes of gaining both size and scale. In 1840, several breeders crossbred their Limousin with oxen of Agenaise variety. Unfortunately, these crossbred cattle proved not to be economical in the majority of the region thus Limousin breeders concentrated upon improving the breed through a very tough natural selective process resulting in an outstanding herd of "purebred" Limousin. The widespread use of natural selection made it important to record the bloodlines of the outstanding Limousin bulls and females as well. So, in November of 1886, the first Limousin herd book was established. Through the late 1800s and early 1900s, Limousin breeders paid close attention to morphological characteristics as the breed developed. The medium size of these cattle as compared to other European breeds was, and is still, an outstanding breed trait. They also selected for the dark golden-red hide with wheat colored underpinnings. French records also show a great deal of emphasis was stressed upon deep chest, a strong top-line, well-placed tailhead and strongly-muscled hindquarter. The end result was an efficient, hardy, adaptable animal that was extremely well-suited for its only intended purpose - to produce beef (Feliuss, 1995; Purdy et al., 2008; [Czech Beef Breeders Association](#)).

3.3.6 Piedmontese

The Piedmontese belongs to the cattle breeds of the Northern Italy Lowland group, the ancestral origin of which is referred to *Bos brachyceros* and to a mixing of *B. brachyceros* and *Bos primigenius* (Baker and Manwell, 1980). Typical of a triple-purpose breed, it was selected in the 1970s for improvement of milk production, through milk performance recording of productivity, while maintaining beef characteristics. In the 1980s, the Breed Society ([National Association of Piedmontese Cattle Breeders](#)) decided to give up milk recording activity and modified the breeding goal to improve beef traits only. The particular characteristic of the Piedmontese cattle breed is in fact muscular hypertrophy, better known as the „double muscle factor“. Milk production of the breed is, however, still more than sufficient to suckle the calf, and several farmers still milk their cows and process milk into typical cheeses. Selection for improving beef characteristics has been regular and intense, and has taken advantage of artificial insemination, which is widespread in this breed. The first herd book was opened in 1887, and improvement campaign and standard of merit have led to many years of genetic selection to eliminate detrimental aspects generally associated with DM. In Italy, the Piedmontese have been (and many still are today) utilized as a dual-purpose animal having very rich milk used for speciality cheese production and beef marketed as a premium product (Purdy et al., 2008).

3.3.7 Simmental

The Simmental is one of the oldest and most widely distributed of all breeds of cattle in the world. Its history dates back to the Middle Ages and is believed to be the result of a cross between large German cattle and a small Swiss indigenous breed (Bonadonna, 1959). Although the first herd book was established in the Swiss Canton of Berne in 1806, there is evidence of large, productive red and white cattle found much earlier in ecclesiastical and secular property records of western Switzerland. These red and white animals were highly sought because of their rapid growth development; outstanding production of milk, butter, and cheese; and for their use as draught animals. Since its origin in Switzerland, the breed has spread to all six continents. Total numbers are estimated between 40 and 60 million Simmental cattle world-wide. More than half of these are in Europe. The spread was gradual until the late 1960s. Records show that a few animals were exported to Italy as early as the 1400s. During the 19th century, Simmentals were distributed through most of Eastern Europe, the Balkans, and Russia, ultimately reaching South Africa in 1895. Guatemala imported the first Simmental into the southern hemisphere in 1897, with Brazil following suite in 1918 and Argentina in 1922. The breed is known by a variety of names, including "Fleckvieh" in Germany, Austria and Switzerland as well as many other European countries, "Pie Rouge"; "Montbeliard", and "Abondance" in France and "Pezzata Rossa" in Italy. The Simmental name is derived from their original location, the Simme Valley of Switzerland. In German language, Thal or Tal means valley, thus the name literally means "Simme Valley" (Felius, 1995; [Canadian Simmental Association](#)).

3.3.8 Fleckvieh cattle

Fleckvieh cattle belongs to the group of European Simmental dualpurpose spotted cattle breeds (Czernekova et al., 2006) started in 1830 when original Simmental Cattle from Switzerland were imported to Bavaria and to Austria to improve the local dual-purpose breeds. At these times, the Simmental cattle were famous for their milk production and draught capacity but were late maturing with little depth and coarse bones. In 1920 the herd book in Southern Germany was closed and the Fleckvieh was developed as an independent dual-purpose breed in Southern Germany, Austria, later also in parts of Italy and France. The breeding aims now are focused on a „middle of the road type“ – dual purpose animal with excellent muscling, good milk production, in past also with good draught performance. Therefore an excellent performance testing system and a strict breeding programme exist. Thanks to the systematic improvement of the production traits Fleckvieh presents a modern, high productive dual purpose breed that fits the actual economical needs.

In the last decade of the last century and first decades of this century the breed was utilised for milk, beef and draught purposes in Europe and for extensive beef production in Namibia and South Africa. In the late sixties and during the decade between 1970 and 1980 the breed which had been changed to a dual purpose breed (milk and beef), established itself on all the continents. The basic colours of the original Simmental-Fleckvieh breed are light to dark yellow and red to dark red with white spots or patches in any pattern way be irregularly spread over the body. The muzzle is cream to pink and may have small grey brown pigmented spots. The breed is still bred as a milk-beef dual purpose breed in central European countries; however, a few countries have started to give more consideration to milk production.

3 Literature Review

During the first half of the 19th century, a great number of cattle breeds and cattle strains were bred in the what is now the Czech Republic. The original domestic cattle, typical for central Europe is called Cervinka. They were bred on the gentry and church estates. The Cervinka cattle were upgraded by crossing with imported animals, which came mainly from area of Austrian regions. Since the second half of the 19th century, Simmental cattle were imported in increasing numbers from Switzerland. At the close of the 19th century, the Fleckvieh cattle were present in many areas of the Czech Republic. In 1920, the records of the breeding of pied cattle were consolidated. Some of the top European bulls were bred with cows to initiate new lines of animals in Bohemia (World Simmental Fleckvieh Federation).

In the post-war period (since 1945), breeding system oriented at production of triple performance offspring was characteristic for Fleckvieh. More than 1/3 of animals was used for draught purposes. This influenced considerably the type of animals; these animals were not able to conform to the desired traits and higher body density. Because of this, the population was upgraded by using Ayshires, and later, Red Holstein cattle with the aim of maintaining the dual performance (beef-milk) characteristics. Regarding to Ayshires milk components were increased significantly to very suitable level, in combination with Red Holstein influence, good milk yield was also reached. The percentage of upgraded animals was from 25% to 37%. Then Hereford bulls were largely used to improve beef production potential of Czech Fleckvieh breed producing crossbreds, so it is evident that breed is suitable for production high quality milk on good production level as well as farmers can sell young animals or feeded bulls for beef. Beef oriented part of production brought very good reproduction performance as well as easy calving in crossbreds. Nowadays, Czech Fleckvieh appears to be one of economical sustainable breed under EU conditions, which represent compromise between high milk performance needings and treatment and extensively kept beef animals (Czech Fleckvieh Breeders Association).

3.4 Population genetic measures

The existence of genetic polymorphism or diversity in a population is the basis for genetic improvement by selection and needs to be accurately estimated (Tautz, 1993). In the past decades, animal genetic diversity has been assessed according to various criteria including phenotypic, biochemical, and molecular parameters. With the development of molecular biology techniques, nucleotide variations in DNA sequences can be detected directly such as microsatellite markers identified in all eukaryotic species investigated so far (Bradley et al., 1998; Saitou and Nei, 1987). Population genetic diversity and structure can be evaluated by numerous methods of inter- and within- populations and subpopulations parameters (Weir, 1996) (e.g. the number of alleles per locus, the average number of alleles for all loci, heterozygosity, and PIC value, F-statistics, PCA method, cluster methods, genetic distances etc.) which usage and state-of-art are described in following text.

3.4.1 Basic genetic diversity measures

Basic genetic diversity for subpopulation or population genetic diversity description are namely:

- number of observations,
- number of alleles per locus, per all watched loci,
- alleles and genotypes frequencies.

These measures give a basic information about the population genetic diversity and their calculation is necessary for decisions about connected results and their confidency (number of observations, number of alleles per loci). The alleles and genotype frequencies are also basic inputs into the estimation of numerous genetic diversity parameters (e.g. genetic distances) and they are necessary for frequency based calculations and algorithms (Weir, 1996; Liu and Muse, 2005; Hedrick, 2011).

3.4.2 Heterozygosity

Heterozygosity is an population-level parameter of genetic diversity. It gives an information about the proportion of loci expected to be heterozygous. Heterozygosity values range from 0 – no heterozygous genotypes to 1.0 – all genotypes heterozygous. Heterozygosity is the one of the basic parameter of the genetic structure of a population at time. Low heterozygosity value indicates effects of small basic population size (inbreeding, bottleneck or reduced of genetic variation). If heterozygosity value is high, we might consider an isolate-breaking effect in genetic structure of population (the mixing of two independent populations).

Typically, the observed (H_0) and expected (H_E) heterozygosities are calculated and compared. Differences between these two measures are connected closely with Hardy-Weinberg equilibrium and its testing. Expected heterozygosity refers about a heterozygosity in population under H-W equilibrium and observed heterozygosity reflects real state in population (Weir, 1996).

3.4.3 The Hardy–Weinberg equilibrium

The Hardy–Weinberg equilibrium (also principle or theorem) states that both allele and genotype frequencies in a population remain constant from generation to generation. Static allele frequencies in a population across generations assume: random mating, no mutation, no migration or emigration, infinitely large population size, and no selective pressure for or against any traits. The Hardy–Weinberg equilibrium is impossible in nature and is an ideal state that provides a baseline to measure genetic change against. In the simplest case of a single locus with one allelic pair, the allele frequencies are $p + q = 1$ and genotype frequencies are $p^2 + 2pq + q^2 = 1$, where p is the frequency of the dominant allele and q is the frequency of the recessive allele. Based on these equations, we can determine useful but difficult-to-measure facts about a population.

Disequilibrium coefficient D_{uv} for alleles A_u, A_v is calculated as

$$D_{uv} = \begin{cases} \tilde{P}_{uv} - p_u p_v, u = v \\ \tilde{p}_u \tilde{p}_v - \frac{1}{2} \tilde{P}_{uv}, u \neq v \end{cases}$$

The chi-square goodness of fit described in (Weir, 1996) is used for calculations of testing H-W equilibrium for multiallelic loci. Chi-square statistics for multiallelic loci is then given by

$$X_T^2 = \sum_u \frac{(n_{uv} - n\tilde{p}_u^2)^2}{n\tilde{p}_u^2} + \sum_u \sum_{u \neq v} \frac{(n_{uv} - 2n\tilde{p}_u \tilde{p}_v)^2}{2n\tilde{p}_u \tilde{p}_v}$$

with $k(k-1)/2$ degrees of freedom, where k is the number of alleles at the loci.

Hardy-Weinberg equilibrium refers to the expectation that genotype frequencies will tend to be stable and predictable as a simple function of individual allelic frequencies, unless there is some evolutionary force. Described model is useful only for two alleles loci. For multiallelic loci, it is useful to estimate Fisher's exact test or likelihood-ratio test for described task (Liu and Muse, 2005; Hedrick, 2011).

3.4.4 Linkage disequilibrium

Linkage disequilibrium (LD) has been defined as non-random association between two loci within a population (Weir, 1996). LD describes a situation in which some combinations of alleles or genetic markers occur more or less frequently in a population than would be expected from a random formation of haplotypes from alleles based on their frequencies. Non-random associations between polymorphisms at different loci are measured by the degree of LD. In natural populations, LD is affected by many factors such as genetic drift, population structure, migration, admixture, selection, mutation and recombination (Hedrick, 2011). LD can play an important role in identifying genes causing simple or complex disease in human populations (Wang et al., 2005). Recently, LD has been estimated in dairy cattle (Farnir et al., 2000; Tenesa et al., 2003; Tenesa et al., 2007) and beef cattle populations (Odani et al., 2006) using microsatellite markers. The extent of LD across genomic regions is a crucial parameter for defining the statistical power of association studies utilizing single nucleotide polymorphisms (SNPs) as surrogate genetic markers (Schork, 2002), and for guiding the selection and spacing of such polymorphisms to create marker maps useful in candidate gene, candidate region and wholegenome association studies (De la Vega et al., 2002). Linkage disequilibrium mapping methods have higher resolution than linkage mapping methods because they use information based on historical recombination with larger numbers of individuals (Pritchard and Przeworski, 2001).

3.4.5 Effective population size

When introduced, effective population size (N_e) was defined as „the number of breeding individuals in an idealized population that would show the same amount of dispersion of allele frequencies under random genetic drift or the same amount of inbreeding as the population under consideration“ (Wright, 1965). Estimation of effective population size is useful for understanding and modelling the genetic architecture of a population (Tenesa et al., 2007). Effective population size can be estimated using small DNA fragments (approximately 10 kb) based on the coalescent model. In cattle, recent effective population size has been estimated from inbreeding status in populations (Nomura et al., 2001; Sorensen et al., 2005). An alternative N_e estimation method applies the relationship between LD and recombination rate between closely linked markers (Hill, 1981). Hayes et al. (2009) estimated the past effective population size in dairy cattle using LD between microsatellite markers.

3.4.6 Genetic distances

Genetic distances are measures of similarities and dissimilarities between and among species and individuals. They seem to be good tools to construct genetic trees, dendograms and phylograms which are typical tasks in evolutionary and ecological studies. They were developed especially for describing wild animal species populations and their evolutionary processes. However domestic animal species have completely different type of evaluation, hardly influenced by geographic aspects, by selection and by breeding strategies (and business as well), genetic distances and connected methods are used as well for describing genetic influences like genetic pressure, bottleneck, population drift etc. which are evident in short term meaning thanks to mentioned influences. A lot of genetic distances types exist. We can divide them into the groups according to model on which based they are calculated.

Nei's 1972 standard distance has an expected value linearly related to the time since divergence. It is assuming that all loci have the same rate of neutral mutation, and that the genetic variation is maintained by the equilibrium between infinite-alleles mutation and genetic drift, with the effective population size of each population remaining constant (Nei, 1972).

None of the geometric distances described in (Nei, 1972; Nei, 1973; Nei et al., 1983) involve any evolutionary models. Assuming that there is no mutation, and that all gene frequency changes are done by genetic drift alone, the following two quantities are expected to rise linearly with the amount of genetic drift (Infinite Allele Model).

Genetic distances widely used in population studies are also Reynold's genetic distance for short term evolution (Reynolds et al., 1983), Cavalli-Sforza and Edwards (Cavalli-Sforza and Edwards, 1967) distance what gives the chord distance between two populations if we represent two populations on the surface of a multidimensional hypersphere, Roger's distance (Rogers, 1972). Roger's distance is based on geometric distances, which are not negative, symmetrical and which satisfy the triangle inequality.

Cavali-Sforza and Edwards' chord distance

Cavalli-Sforza and Edwards (1967) distance gives the chord distance between the two populations if we represent two populations on the surface of a multidimensional hypersphere using allele frequencies at the j th locus:

$$DC = \frac{2}{\pi m} \sum_{j=1}^m \sqrt{2 \left(1 - \sum_{i=1}^{a_j} \sqrt{p_{ij} q_{ij}} \right)}.$$

Rogers' distance

Geometric distances are not negative, symmetric and satisfy the triangle inequality. Rogers (1972) is a scaled Euclidian distance:

$$DR = \frac{1}{m} \sum_{j=1}^m \sqrt{\frac{1}{2} \sum_{i=1}^{a_j} (p_{ij} - q_{ij})^2}.$$

The most common distance is the Euclidean distance. The other distance based on geometric distances is Prevosti et al.'s distance (Prevosti et al., 1975).

The main difficulty posed by microsatellite loci for their use in the evaluation of genetic distance is their relatively high mutation rate. This makes it difficult to adopt any of the two main mutation models used in population genetics, the infinite alleles or the stepwise mutation model. There is still uncertainty as to whether allele sizes are unconstrained or whether there are certain limits to the number of repeats present (Estoup et al., 1995; Garza et al., 1995). Assuming a stepwise mutation model (Slatkin, 1995; Goldstein et al., 1995a), we have recently proposed distance measure for microsatellite alleles. The distance between two alleles is a simple transformation of the number of repeat units. The within population measure of distance is obtained as the average sum of squares of the differences in number of repeats between alleles (Liu and Muse, 2005).

Microsatellite variation appears to result from slippage in replication, which is most likely to add or delete a single repeat unit. As a result, alleles more similar in size will presumably be more closely related. This additional phylogenetic information can be used in assessing genetic differentiation or genetic distance. The stepwise mutation model (SMM) is an alternative to the infinite alleles model (IAM) as the basis for deriving measures of genetic differentiation.

Stepwise mutation index is defined as the maximal proportion of alleles that follows the stepwise mutation microsatellite data pattern. This statistic needs the length of the repeat unit to be specified. This information has to be included in the covariate table of the markers and the column property needs to be numeric. In the group of genetic distances calculated under SSM we include e.g. Goldstein et al. distance (Goldstein et al., 1995b; Goldstein et al., 1995a), Average Square Distance (ASD) (Slatkin, 1995).

Shriver investigated the correlation between observed and simulated values based on the SMM and estimated distance based on these correlations (Shriver et al., 1995). This study compared three parameters; the number of alleles, the range of allele sizes, and the number of modes in the distribution of alleles. Another commonly used distance, the shared allele distance D_{SA} was defined by (Chakraborty and Jin, 1993).

Also, other concepts of genetic models are considered for genetic distance calculations. Relative entropy is a very important concept in quantum information theory, as well as

statistical mechanics (Qian, 2001). On the basis of this concept, e.g. Kullback-Leibler distance was established.

Kullback-Leibler distance

A discrete distribution has probability function p_k , and let a second discrete distribution have probability function q_k . Then the relative entropy of p with respect to q , also called the Kullback-Leibler distance (Kullback, 1987), is defined by

$$DS_{KL} = \sum_k p_k \log_2 \left(\frac{p_k}{q_k} \right).$$

Although $DS_{KL}(p, q) \neq DS_{KL}(q, p)$, so relative entropy is therefore not a true metric, it satisfies many important mathematical properties. For example, it is a convex function of p_k , is always non-negative and equals zero only if $p_k = q_k$.

3.4.7 Wright's inbreeding coefficient and F-statistics

Level of inbreeding for individual can be defined by coefficient F_X , of the following statements:

1. The probability that both genes of a pair in an individual are identical by descent, i.e. homozygous.
2. The probable proportion of an individual's loci containing genes that are identical by descent.

An equation for estimation of individual inbreeding coefficient was formulated by Sewall Wright (Wright, 1922) as:

$$F_X = \sum \left(\left(\frac{1}{2} \right)^{n_1+n_2+1} (1 + F_A) \right),$$

where F_X is the inbreeding coefficient, F_A is the inbreeding coefficient of the common ancestor, n_1 is the number of generations from the sire to the common ancestor, and n_2 is the number of generations from the dam to the common ancestor.

A very useful measure of population subdivision is the F-statistics developed by (Wright, 1965). F-statistics can be thought of as a measure of the correlation of alleles within individuals. This correlation is influenced by several evolutionary processes, such as mutation, migration, inbreeding, natural selection, or the Wahlund effect, but it was originally designed to measure the amount of allelic fixation owing to genetic drift. F-statistics describe the amount inbreeding-like effects within subpopulations (F_{IS} or f) – the inbreeding coefficient of an individual (I) relative to the subpopulation (S) – inbreeding coefficient, among populations (F_{ST} or θ) – coefficient of subpopulations (S) compared to the total population (T) – fixation index, and within the entire population (F_{IT} or F) – the inbreeding coefficient of individual (I) relative to the total population (T) – overall fixation index. Let's define

- H_I = mean observed heterozygosity per individual within subpopulations,
- H_S = mean expected heterozygosity within random mating subpopulations,
- H_T = expected heterozygosity in random mating total population.

The inbreeding coefficient measures the reduction in individual heterozygosity due to deviations from random mating in the local populations. This inbreeding coefficient is represented by F_{IS} and is given by:

$$F_{IS} = \frac{\bar{H}_S - H_I}{H_S}.$$

The effects of population subdivision can be quantified by means of the fixation index F_{ST} . This index demonstrates the reduction in the heterozygosity in a subpopulation due to non-random mating with respect to the total population. F_{ST} is given by:

$$F_{ST} = \frac{H_T - \bar{H}_S}{H_T}.$$

An alternative interpretation of F_{ST} in its diallelic version is the ratio between the expected and observed variances of gene frequency considered among all subpopulations –

$$F_{ST} = \frac{\sigma_p^2}{p_0q_0}.$$

A coancestry matrix is formed by calculating θ for each pair of populations. It can be requested for the log transformation ($= -\ln(1 - \theta)$) to be performed, which leads to a measure of genetic distance under a drift model.

Overall fixation index is given by:

$$F_{IT} = \frac{H_T - \bar{H}_I}{H_T}.$$

The following relationship holds for F-statistics:

$$(1 - F_{IS})(1 - F_{ST}) = (1 - F_{IT}).$$

Population specific F-Statistics extends the classical F-statistics by allowing different levels of coancestry for different populations and by allowing non-zero coancestries between pairs of populations. The procedure of estimating population specific F-Statistics and between-population F-Statistics was formulated by (Weir and Hill, 2002).

3.4.8 Microsatellite panel effectiveness measures

Methods described below in this paragraph are necessary to calculate if we want to prove an usability of microsatellite set to routine genotyping and parentage testing. Usually, PIC (Botstein et al., 1980), $PE(1)$, $PE(2)$, $PE(3)$ and CEP (Jamieson and Taylor, 1997) measures are calculated across the loci in set.

Informativeness of polymorphic markers can be quantitatively measured by a statistic called the polymorphism information content, or PIC - the ability of microsatellite length polymorphism distinguish genotypes on small number of loci. This measure can be also used to identify and locate a hard-to-define marker locus.

The probability of exclusion non correct parent, when the genotypes of offspring and both parents are known is marked as $PE(1)$. One of the parent is verified based on sample of population allele frequencies in this case.

The probability of non correct parent exclusion, when one of genotype of parents is unknown is called $PE(2)$.

$PE(3)$ is the probability of exclusion non correct parents, when the genotype of offspring

3 Literature Review

and both parents are known. For given probabilities this relation – $PE(2) < PE(1) < PE(3)$ – is valid.

Combined Exclusion Probability is calculated for each type ($n = 1, 2, 3$) of paternity exclusion:

$$CEP(n) = 1 - (1 - PE(n)_1)(1 - PE(n)_2) \dots (1 - PE(n)_k),$$

where index 1,2,3...k indicates numbers of microsatellite loci.

3.5 Population genetic structure using machine learning methods

3.5.1 Machine learning – state of art

The convergence of computing and communication has produced a society that feeds on information. Anyway, most of the information is in its raw form: data. Data are characterized as recorded facts, then information is the set of patterns or expectations, that underlie the data. There is a huge amount of information locked – data include information that is potentially important but has not been discovered yet (Witten et al., 2011). Data mining is the extraction of implicit and previously unknown information from data (in comparison with statistical methods, where we prove a pre-formulated hypothesis). Concept of datamining was formulated in 90's. The basic idea is to built robust algorithms for seeking patterns (non banal or inunderstanding) in databases, generalize them to make an accurate predictions on future data.

Machine learning provides the technical basis of datamining. It is used to extract an important information from the raw data by using algorithms built on basic principles of artificial intelligence. As datamining is a process of data understanding, machine learning is one of the instruments to find and generalize patterns in data with usage of computers. Finding of patterns and their reliability are easy comparable across different algorithms thanks to established measures of different pattern quality (Berka, 2001).

Machine learning methods are in generally designed as mathematically fomulated simulations of one of basic human ability: learning. Basic scheme of machine learning system can be described as following schema:

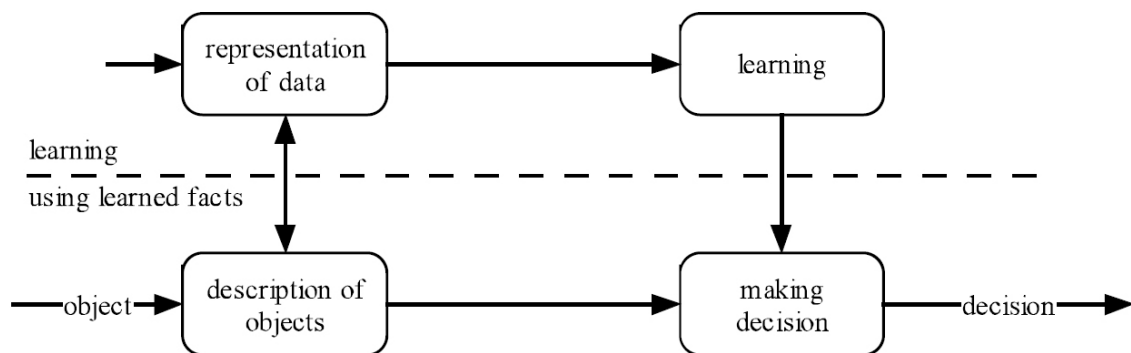


Figure 3.1 Schema of learning system.

Process of machine learning has basically three phases. For machine learning methods, good data preparation i.e. object description is necessary – first phase. On the other side, for most of machine learning methods, no predispositions (like distribution etc.) of data have to be satisfied as in classical mathematical statistics. Process of the model training (training data is used) creates a second phase. With usage of training data, model is prepared. The third phase is an usage of created model in one of typical task, or its continuous training (Bishop, 2006).

Typical usage of machine learning methods is in following problems areas (Berka, 2001; Witten et al., 1999):

- classification,
- prediction,
- clusterization,
- datamining – representation of obtained, primarily hidden information in data.

The basic concepts of probability are quite straight forward (Bishop, 2006) and directly connected with machine learning methods. First basic concept of probability can be illustrated e.g. by segregation in the mating of individuals with heterozygous genotypes. The possible genotypes of each progeny – AA, Aa, aa are realized with probabilities, $\frac{1}{4}$, $\frac{1}{2}$, $\frac{1}{4}$ (additional rule – according to OR in logic or + in probability of at least one of two independent events). In large dataset and under random mating, genotype frequencies will reflect these probabilities. Also multiplication rule – according to AND in logic or * in probability of two independent events happening together – is a second basic principle of probability. With regard on these paradigms all of machine learning methods are created.

We can divide machine learning methods according to type of data used for model training:

- methods based on learning with teacher – preclassified data for model building – e.g. decision trees,
- methods based on learning without teacher – dataset consists of non-classified data – hidden structure of data is exploring – e.g. clusterization methods,
- methods of learning based on imitation – continuous learning based on user reactions and behaviour – e.g. some web searching engines,
- methods of boosting – model is continuously trained and system of penalisation is defined for model – e.g. control systems.

Machine learning represents dynamic science. A complex system of methods, their modifications and theoretical background were created in past 20 years. Based on methods' theoretical background we divide them as follows (Bishop, 2006):

- statistical methods – Bayes Classifiers, Markov chains or Hidden Markov models (Andrieu et al., 2003), Instance Based Learning Methods (Aha et al., 1991), Cluster analysis, Regression methods (Quinlan, 1992),
- symbolic methods of artificial intelligence – Decision Trees (Quinlan, 1986), Decision Rules, Association rules, Inductive Logic Programming,
- sub-symbolic methods of artificial intelligence – Neural Networks, Support Vector Machines (Michie et al., 1994).

Two typical problems of machine learning (classification – class is known and clusterization – unknown class) and their applications in population genetics could be used widely thanks to routine genotyping under the condition of their validation. Typical usage of these methods represents a problem of a breed discrimination and problem of genetic population hidden structure. All of machine learning methods calculate with probability of assignment to the clusters or classes, therefore they can also be useful for the visualization of genetic structure. The precision of classification or other quality-of-fit measures on the population level can be also used as the measure of genetic distance.

These tools are still not widely applied for genetic analysis, except a few works (Beaumont and Rannala, 2004; Pearse and Crandall, 2004; Kruger et al., 2005). In following chapters, Following chapters describe basic usage of machine learning methods and their principles in tasks of population genetic structure and diversity.

3.5.2 Classification

Classification methods solve a problem described as to build a model which can classify instances-data rows to their known classes with the best possible parameters with which the model fits the training data. Also, classification problem is generalized problem of regression and vice versa.

Built models can be validated with a lot of methods which can describe models robustness (n-fold cross validation, training set validation, test set validation, leave-one-out, boot-strap) (Witten et al., 1999; Witten et al., 2011). By this validation methods, plenty of measures of classification quality can be obtained. Parameters of model performance calculated during model validation are typically (Berka, 2001):

- confusion matrix,
- overall accuracy and error,
- accuracy and error calculated for each class,
- precision and recall,
- F-measure,
- Kappa Statistics,
- Mean absolute error and Root mean squared error,
- Relative absolute error and Root relative squared error
- ROC curves and cost curves,
- etc.

In the context of classification task, we use some basic terms, which it is necessary to define at this place:

- attribute – parametr or variable used for description of objects' features, like income,
- temperature, genotype etc.
- dataspace – space given by values of attributes, each attribute represents an axis in multidimensional space,
- instance – data representation of described object, set of concrete attributes values, one row in dataset,
- dataset – collection of instances with the same attributes,
- class – one of attributes, the aim attribute, specify the attribute which is used for classification model – representation of dataset created by some method which represents a dataset usually in generalized form, it describes real state included in dataset,
- training of model – process of model creation based on dataset, the part of dataset used for is called training set,
- testing of model – process of testing model classification performance by using plenty of measures.

Methods of classification are useful when we have a data with typically a lot of attributes and we want to build model which can classify or discriminate data or predict one chosen attribute called a class. We can also analyze the structure and the relationships between and among classes.

Methods of classification of individuals into the breeds based on variable genotype data can be used as a practical result of a genetic diversity analysis (Masuda and Pella, 2004; Canon et al., 2000) for the purpose of breed discrimination (Burócziová and Říha, 2009). However, only if classification power of the model is decided to be usable. Models with high precision of classification based on genetic variability data can be used for breed or another

chosen level discrimination in traceability and safety of food resources e.g. (Guinand et al., 2002; Dawson and Belkhir, 2001; Manel et al., 2002; Masuda and Pella, 2004).

In case of genotype data, we can discuss the results as precision of classification calculated on many levels as the genetic distance given by chosen model. For groups of two breeds only we can use precision of classification using a chosen method as a genetic distance between the breeds (Kitada et al., 2000). Also we can analyze errors of classifications which lead to similarity of instances. When we want to discuss the relationships among breeds we can analyse a confusion matrix of classification models with high classification power (Masuda and Pella, 2004).

For the breed discrimination task, good results are achieved especially thanks to methods built for usage with frequency data (Burócziová and Říha, 2009). Therefore, Bayes Naive Classifier, Bayes Net and MCMC are typical representatives of this class of algorithms (Witten et al., 2011). Methods based on generalization of data are not applicable for this task usually (Guinand et al., 2002; Burócziová and Říha, 2009).

For discussion of variability, generalized algorithms (IB5, J48, JRip) are more useful than precise classifiers. If the results of classification match overall selected models, it is better to use generalized methods for discussion of genetic variability, i. e. based on confusion matrix, because we want to discuss breed characteristics, so the aim is different than in breed discrimination problem (Masuda and Pella, 2004).

As all selected methods work with probability, they are also useful for visualization of genetic variability data on e.g. plot diagrams of individuals and their predicted class or cluster classification (Pritchard et al., 2000; Glowatzki-Mullis et al., 2006).

3.5.2.1 ZeroR

ZeroR is a base classifier to control how different models of classification power are better or worse in comparison with the basic one. ZeroR always predicts the mean for numeric class or most frequent class for nominal attributes for each new instance, so it could be used as a good performance reference classification method which can show that model is overlearned especially in the case of non-equal numbers of instances in different classes.

3.5.2.2 Decision trees

Decision trees are effective method for representation of knowledge (e.g. keys to determination, guidelines etc.). Using decision trees we can resolve the problem of classification into the classes. So, when algorithms to their induction from large datasets appeared, they have become one of the key methods of machine learning and knowledge discovery. They do not function, as many methods of machine learning (e.g. Support Vector Machine) as blackbox, but they allow to represent derived knowledge about multinomial dataspace in comprehensible form.

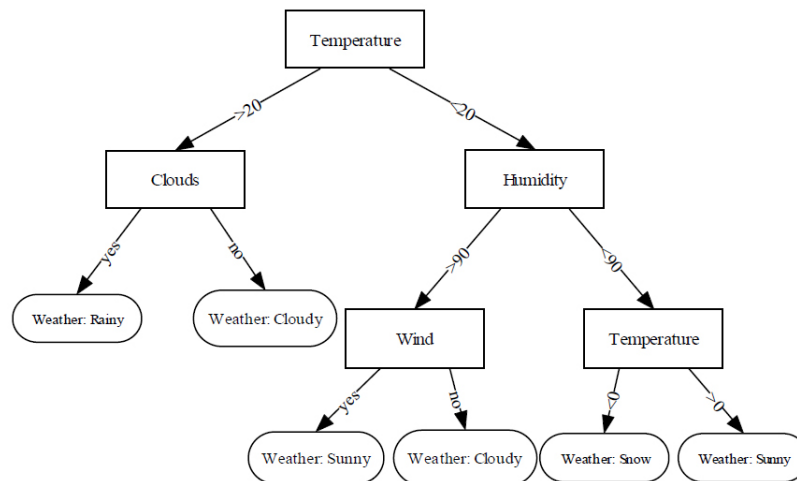


Figure 3.2 Example of decision tree.

Decision trees induction (Quinlan, 1986) uses „divide and conquer“ strategy. The dataspace is dividing into the subsets of selected parameter from dataspace during the induction of tree, so that subsets consisting mainly from instances of one of attribute classes. In other interpretation, decision trees divide the dataspace into the classes by using hyperplanes parallel with the axes (attributes).

The induction of decision tree is based on top-down specialization in hypothesis space. On the start of induction, all data create first node of the tree. During the process of induction, the tree is specialized and on the end the final tree has nodes with instances of one class for each attribute. Under assumption of representing sub-population selection for training dataset we can use decision tree for classification by walking through decision tree nodes according to values of attributes of classified instance. For this instance we make decisions according attributes values and the conditions in nodes. The inferred decision tree has the decision rule for selected attribute in each node which determine the subtree belongs for instances with this value of attribute. When a leaf node is reached, class of leaf node matches the class of instance.

TDIDT algorithm

1. Choose one attribute as the root of decision subtree.
 2. Divide data in this node according chosen attribute values into the subgroups and add nodes for each subgroup.
 3. If the node with instances of two or more classes exists, do 1 for this node, otherwise finish.
-

The selection of attribute in first step of algorithm can be done by plenty of measures:

- entropy

$$H = -\sum_{t=1}^T (p_t \log p_t)$$

where p_t is a probability of occurrence of the class t and T is the number of classes;

- information gain, $gain(A) = H(C) - H(A)$, $H(C) = -\sum_{t=1}^T \frac{n_t}{n} \log_2 \frac{n_t}{n}$,

where $H(C)$ is an entropy for full dataset, $H(A)$ in an entropy for attribute A ;

- Gini index, $Gini = 1 - \sum_{t=1}^T p_t^2$, where p_t is a probability that instance p has class t ;
- X^2 .

The presumption that node have to contain only instances of one class is often replaced due to generalization of trees (otherwise large trees could be inducted) by usage of „pruning“ methos e.g. Occam’s razor, so the node must have only prevalent number of instances of the dominant class because of when the noise is present in training data algorithm leads into very large trees.

We can also represent decission trees as the set of decission rules in the form:

IF conjunction of premises THEN deduction.

The set of rules that represents decission tree should be obtained by summarize of each way from the root of the tree to leafs. Conditions included in nodes make a conjunction in premise of implication, the leaf is deduction of decission rule. This principle is used also for decision trees post-pruning.

Pruning algorithm

1. convert a decision tree into the set of rules.
 2. generalize the rule by condition removal from premise if estimated improvement of classification is reached
 3. make an order of rules due estimated to improvement – in this order they will use for classification.
-

Pesimistic estimation of classification improvement

1. calculate accuracy of rule as portion of well classified instances and instances covered by rule
 2. calculate standard deviation of this accuracy
 3. take the lower estimatition of accuracy as characteristics of the rule
-

TDIDT algorithm can handle only categorical attributes. It is possible to handle categorical and continuous attributes in decission tree induction with a few modifications of the basic algorithm (e.g. binarization of numeric attributes). Most known modification of the basic induction algorithm is e.g. C4.5 algorithm (Quinlan, 1993) which is implemented in Weka-3-6-6 framework as J48 algorithm (Witten et al., 1999) and which is used in this work.

3.5.2.3 Decision rules

Decision rules are useful for the same task as decision trees – classification of instances into the classes. The construction of decision rules from decision trees in no only possibility to obtain decision rules based on the training dataset. In many cases more universal rules than implication with conjunction in premise are needed. We want to obtain decison rules in the form:

IF *Ant* THEN *Class*, i.e. $Ant \Rightarrow Class$

where *Ant* is universal combination of conditions based on attributes’ values connected with operators of basic logic (conjuncton, negation, disjunction, etc.). *Class* is a class of instance – category of the target attribute (Bishop, 2006).

The interpretation power of decision rules is the same as of decision trees. They also divide the space of attributes according to hyperplanes parallel with attributes' axes. In comparison, different type of algorithms is used for their induction – „separate and conquer“. Basic principle of these type of algorithms represents the algorithm of Sets covering.

Sets covering algorithm

1. Choose a class.
 2. Find rule that covers only positive examples (which belong to covering class, not negative ones) by generalization of positive example.
 3. Delete covered examples of training set.
 4. If any non covered examples remained go to 2, else done and continue for next class.
-

Also, in connection with decision rules we can mention the decision lists. Decision lists are decision rules in form:

$$\begin{aligned} &\text{IF } Ant_1 \text{ THEN } Class_i, \\ &\text{ELSE IF } Ant_2 \text{ THEN } Class_j, \\ &\dots, \\ &\text{ELSE IF } Ant_n \text{ THEN } Class_k. \end{aligned}$$

In this case, each rule (except the first one) in list contains exclusion of previous rules in list, so the rules in list are not independent. This method may be usable in cases when data contain general rules with small numbers of important exceptions which could not be captured by normal decision rule.

For this work purposes, the JRip algorithm implemented in Weka-3-6-6 software was selected. The JRip algorithm is the implementation of the RIPPER one (Cohen, 1995).

RIPPER algorithm

Initialize set of rules as $RS = \{\emptyset\}$, and for each class from the less prevalent one to the more frequent one, DO:

1. Building stage:

Repeat 1.1 and 1.2 until the description length (DL) of the rule set and examples is 64 bits greater than the smallest DL met so far, or there are no positive examples, or the error rate $\geq 50\%$.

- 1.1. Grow phase:

Grow one rule by greedily adding antecedents (or conditions) to the rule until the rule is perfect (i.e. 100% accurate). The procedure tries every possible value of each attribute and selects the condition with highest information gain: $p(\log(p/t) - \log(P/T))$.

- 1.2. Prune phase:

Incrementally prune each rule and allow the pruning of any final sequences of the antecedents; The pruning metric is $(p-n)/(p+n)$, respectively $2p/(p+n)-1$, so in this implementation we simply use $p/(p+n)$ (actually $(p+1)/(p+n+2)$, thus if $p+n$ is 0, it's 0.5).

2. Optimization stage:

After generating the initial ruleset $R_i \{ \}$, generate and prune two variants of each rule R_i from randomized data using procedure 1.1 and 1.2. But one variant is generated from an empty rule while the other is generated by greedily adding antecedents to the original rule. Moreover, the pruning metric used here is $(TP+ TN) / (P+ N)$. Then the smallest possible DL for each variant and the original rule is computed. The variant with the minimal DL is selected as the final representative of R_i in the ruleset. After all the rules in $R_i \{ \}$ have been examined and if there are still residual positives, more rules are generated based on the residual positives using Building Stage again.

3. Delete the rules from the ruleset that would increase the DL of the whole ruleset if it were in it. and add resultant ruleset to RS.

ENDDO

3.5.2.4 Naive Bayes classifier

Bayesian classification is based on applying Baye's theorem about the conditioned probabilities. Bayes theorem defines the probability of hypothesis H (class in this case) in conditon of hypothesis E (attributes) acceptance like

$$P(H | E) = \frac{P(E | H)P(H)}{P(E)}.$$

Apriori probability $P(H)$ corresponds to class distributions without regard to the other information. $P(E)$ represents probability of evidence (observation). Conditioned probability $P(H | E)$, called aposteriori probability, represents change of hypothesis probability when E occurs.

In the context of classification, we need to find the hypothesis with maximum probability for given evidence. This hypothesis is given by this formula:

$$H_{MAP} = H_J \Leftrightarrow P(H_J | E) = \max_t \frac{P(E | H_t)P(H_t)}{P(E)}.$$

This formula give most probable hypothesis in condition of one evidence. The advantage of Bayesian classification is also that they given probabilistic classification primarily by calculating of all aposteriori probabilities. Described classification is useful only for one evidence. For more evidences, we need to estimate aposteriori probability $P(H | E_1, \dots, E_k)$.

One of the method to do this under the assumption of independant evidences is Naive Bayes classifier. Naive Bayes classifier can estimate the probability of hypothesis which depend on the acceptance of independent evidences (variables, attributes) by this equation:

$$P(H | E_1, \dots, E_K) = \frac{P(E_1, \dots, E_K | H)P(H)}{P(E_1, \dots, E_K)} = \frac{P(H)}{P(E_1, \dots, E_K)} \prod_{k=1}^K P(E_K | H)$$

so we find hypothesis with maximum aposteriori probability as

$$H_{MAP} = H_J \Leftrightarrow P(H_J) \prod_{k=1}^K P(E_K | H_J) = \max_t \left(P(H_t) \prod_{k=1}^K P(E_K | H_t) \right).$$

Assume, that $E_k = A_j(v_k)$, where $A_j(v_k)$ means probability of occurence of attribute A_j with value v in class k ; $H_t = C(v_t)$ is probability of occurence of instance with value v in

class t . Afterthat, we can calculate probabilities $P(H_t)$ and $P(E_k | H_t)$ from training set as:

$$P(H_t) = P(C(v_t)) = \frac{n_t}{n},$$

$$P(E_k | H_t) = P(A_j(v_k) | C(v_t)) = \frac{n_t(A_j(v_k))}{n_t}.$$

Depending on the precise nature of the probabilities, Naive Bayes classifier can be trained very efficiently in described supervised learning scheme, especially for large datasets (because only a posteriori probabilities have to be calculated simply directly from dataset). In spite of naive design and assumptions of independent evidences, this classifier often work much better in many complex situations than one might expect (Kohavi, 1996). Recently, careful analysis of the Bayesian classification show us some theoretical reasons for the apparently unreasonable efficiency of Naive Bayes classifier. An advantage of the Naive Bayes classifier is that it requires, in case of representative dataset, not large training datasets to estimate the probabilities necessary for classification. Also, it is very useful, as it is evident from text above, for classification of data based on frequencies changes which the genotype datasets (Roeder et al., 1998; Dawson and Belkhir, 2001; Masuda and Pella, 2004; Beaumont and Rannala, 2004) exactly are.

3.5.2.5 BayesNet classifier

The main problem of Naive Bayes classifier – assumption that evidences are independent – is to compute all of combinations of evidences probability. As it is a complex problem, problem of independent evidences assumption or complex computations can be pass effectively by Bayes networks (Berka, 2001; Bishop, 2006).

Bayes networks can represent partial dependency evidences and use this representation to make a decisions. Evidences A and B are semi-independent when hypothesis H probability is given as

$$P(A, B | H) = P(A | H)P(B | H).$$

Bayes network is oriented as an acyclic graph where:

1. Edges are probabilities of dependency among independent variables (nodes).
2. Every nodes u has the probability of distributions $P(u | parents(u))$.

When we make an order of nodes and number them as parents of nodes have a lower number than children, it is valid, that every node is semi-independent at each node with the lower number with exception of its' parents and parents of parents.

On the base of this proposition, we can calculate probability distribution of all network as:

$$P(u_1, \dots, u_n) = \prod_{i=1}^n P(u_i | parents(u_i)).$$

Bayes networks combine two types of knowledge representations: probabilities of nodes (attributes) and structure of network (dependency of nodes). The stucture of network may be known but in many cases so we can design it (make known egdes between attributes), we

have to infer it which is also complex problem of searching in space of models. For this purpose, a lot of methods can be used like (Witten et al., 2011):

- genetic programming - works by having a population of Bayes network structures and allow them to mutate and apply cross over to get offspring. The best network structure found during the process is returned,
 - hill climber - uses a hill climbing algorithm adding, deleting and reversing arcs. The search is not restricted by an order on the variables (unlike K2),
 - K2 - uses a hill climbing algorithm restricted by an order on the variables,
 - EM algorithm etc.,
- which lead to appropriate solutions.

Bayes Network learning algorithm implemented in Weka-3-6-6 software using various search algorithms and quality measures for infer the network structure and probabilities of inferred nodes and facilities common to Bayes Network learning algorithms.

3.5.2.6 Instance based classifiers

Instance based classifiers are typical methods based on principle of analogy. The principle of analogy means, that in case of learning or making decision we do not use generalized examples for this but we use most similar example from training stage. On the base of this example, decision is processed and new example is classified (Bishop, 2006).

Key concept of methods based on analogy is the metrics of similarity. A lot of types of metrics (also called distance functions) are defined. Also, database for chosen instances storage have to be created for this type of learning methods.

Each instance based learning algorithm need following three methods implemented:

1. function to measure the instances distance – distance function – the similarity of instances can be expressed by this function
2. function for choosing of instances which are saved in the database,
3. function for classification of new instances.

Metrics is a function, which satisfy this definition:

$$\begin{aligned}
 f &: X \times X \rightarrow \mathbb{R}, \\
 \forall x_1, x_2 \in X &: f(x_1, x_2) \geq 0, \\
 f(x_1, x_2) = 0 &\Leftrightarrow x_1 = x_2, \\
 f(x_1, x_2) &= f(x_2, x_1), \\
 \forall x_1, x_2, x_3 \in X &: f(x_1, x_2) + f(x_2, x_3) \geq f(x_1, x_3).
 \end{aligned}$$

Typical distance functions which express distance between two instances $x_1 = [x_{11}, \dots, x_{1m}]$, $x_2 = [x_{21}, \dots, x_{2m}]$ are e.g.:

Euclidean distance

$$d_E(x_1, x_2) = \sqrt{\sum_{j=1}^m (x_{1j} - x_{2j})^2},$$

Manhattan distance

$$d_M(x_1, x_2) = \sum_{j=1}^m |x_{1j} - x_{2j}|,$$

Chebyshev distance

$$d_C(x_1, x_2) = \max_i |x_{1j} - x_{2j}|.$$

For storage of instances e.g. IBk methods (Aha et al., 1991) are widely used. IB1 method stored into the database all of training instances which is usable especially when data in-

clude noise. IB2 method try to classify new training instance first, if failed, instance is stored in database. IB3 uses sophisticated algorithms based on frequencies of corret and incorrect classification for choosing of instances to store.

Algorithm of k-nearest neighbours is often used for classification of new instances. It uses weighted voting of new instance class based on the distances of k nearest instances in the database. Centroids of classes are also often used instead of all of class instances (Aha et al., 1991).

k-NN algorithm

1. find k nearest neighbours of the classified instance .
2. classify the instance according to:

$$y_j = y_i \Leftrightarrow \sum_{k=1}^K \delta(y_j, y_k) = \max_i \sum_{k=1}^K \delta(y_j, y_k),$$

where $\delta(y_j, y_k) = 1$ for $y_i = y_k$ else $\delta(y_j, y_k) = 0$.

In other words, by the chosen metrics, k nearest instances are founded, afterthat, frequency of classes determines the class of new instance.

3.5.2.7 MCMC algorithm

Via a process called Markov chain Monte Carlo (MCMC) we can accurately reflect very complicated desired probability distributions. In recent years this allowed a wide range of posterior distributions to be simulated and their parameters found numerically in e.g. Bayes Networks.

Monte Carlo Markov Chain algorithm is based on intensive simulations generalization of Expectation Maximalization algorithm (Andrieu et al., 2003) and it is an approximation technique. Two phases of algorithm (burn-in period and after burn period with output) are performed to learn the distributions of propabilities of attributes and assignment of instances to the clusters. Every state of Markov Chain is depended only on the previous state and propabilities (frequencies) of it which are changed in the next step. After random sampling of dataspace an algorithm converges, what is mean that it found a local extreme. All algorithms presume that the caught local extreme is also the global one, so the model, inferred distributions and propabilities relates to given dataset.

Instead of primary usage for clusterization and distributions estimation, MCMC algorithm is also useable for classification (Andrieu et al., 2003) approach with primary class knowledge, also in population genetics (Pritchard et al., 2000). It results in the structure similar to Bayes network structure – Markov chain. In this case convergence is reached more quickly (because both of periods can be shorter, but the classifier can converge to local extreme) and classifiers built by this method are usually robust and with high level of precision of classification. This is result of primary usage of MCMC in intensive simulations so, the function is highly depend on the dataset (precision is very good also in small datasets, but they should not catch variability of real state of problem – problem of local extreme convergence).

3.5.2.8 Neural Networks

The principle of neural networks is derived from the simulation of human neurons (Michie et al., 1994). Neurons react on stimulations and pass on impulses (based on non-linear function) to connected neurons. One of the first concept of neuron simulation was Adaline (Figure 3.3) which may function as binary classifier. Input is represented by stimulation inputs a_1, a_2 (attributes in classification). Values of stimulations are weighted by weights w_1, w_2 .

When sum of weights is greater than threshold value w_0 , output from Adaline is -1. Else the value of sum is passed to the decision model, where computation of resulting value is performed. Typical function used for this purpose is sigmoidal function

$$f(SUM) = \frac{1}{1 + e^{-SUM}},$$

so the output from Adaline is in interval $[0,1]$.

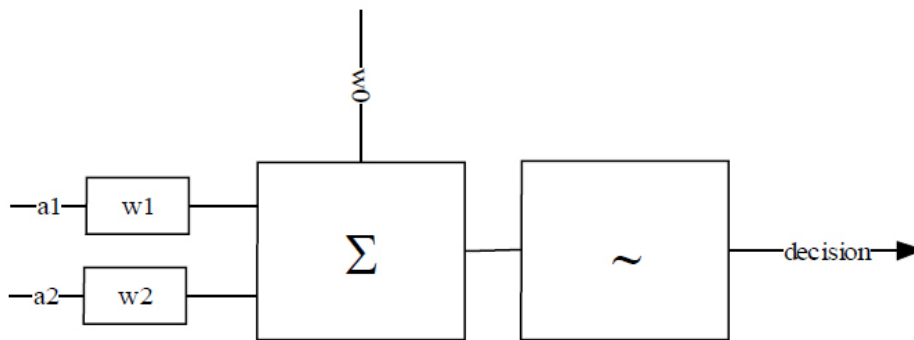


Figure 3.3 Adaline model.

Adaline, or the other model of neuron could be learned by modifications of weights from training set (e.g. with usage of Hebb law, gradient method of mean classification error minimization, stochastic approximation, back error propagation) (Michie et al., 1994).

Neural networks (layers of connected neurons) divide the space of instances by non-linear hyperplanes, so they can catch more complicated problems (e.g. in comparison with decision trees). The basic neural network is perceptron which was created according to human visual system. Perceptron consists from three layers of neurons (Figure 3.4).

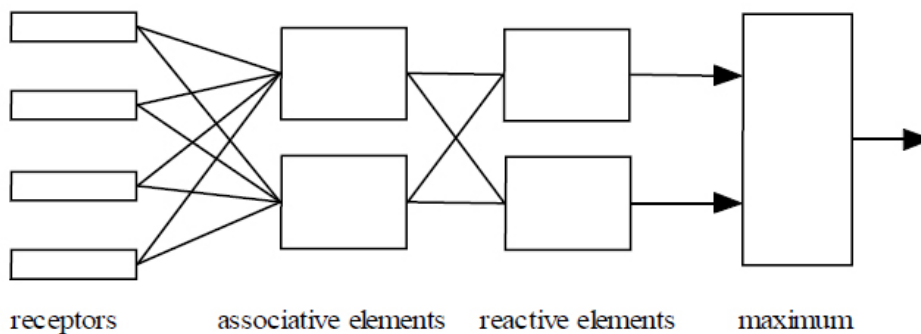


Figure 3.4 Model of perceptron.

The first layer – receptors – matches number of incoming stimulations (attributes values) with 0 and 1 outputs. Associative elements created the second layer of perceptron. They are connected with receptors by randomly selected linkages. After sum of inputs is calculated, associative elements output +1 or -1. The last layer is created by reactive elements (their number corresponds to number of classes), which realized weighted sum of inputs. Also, on this layer learning by corrections of weights is performed. In the last stage of classifica-

tion, element with maximum sum of weighted inputs is selected and it correspond to the class of instance. Modification of the basic perceptron is multilayer perceptron, where no connections between neurons in one layer exist, but every neuron from one layer is connected with every neuron in higher layer (Figure 3.4). Algorithms of neural network classification have generally a lot of properties to set. Working with neural networks is due to this reason often called as „alchemy“ in machine learning community – it is hard to make good neural network which reflect robustly the reality and is usable for classification. However, the ability of catch very complex problem (e.g. it is possible to solve any logical circuit by neural network) makes this method one of the most usage in machine learning. Multilayer perceptron algorithm with error backpropagation from Weka-3-6-6 (Witten et al., 1999; Witten et al., 2011) is used for this work purposes.

3.5.2.9 Support Vector Machines

Support Vector Machines have the similar predicative ability as neural networks (they divide dataspace using non-linear curves or hyperplanes respectively) but they actually do it by usage the different methods. The basic principle lie in finding of kernel function which transform non-linear separable data to the dataspace in which data are linear separable. I.e. kernel function transforms data into the dataspace with higher dimensionality, where the key goal is to find the linear hyperplane which divide them and which is used for classification afterthat (Figure 3.5).

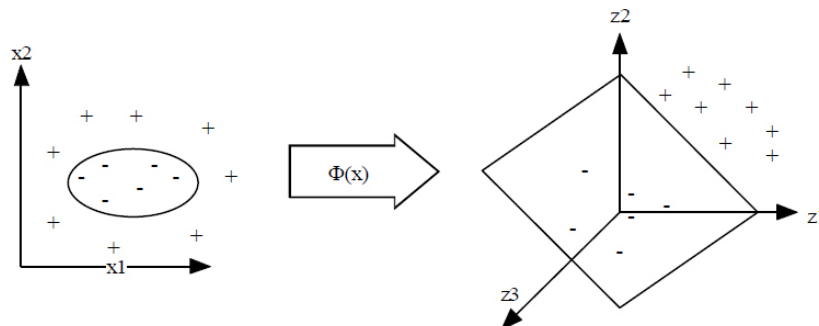


Figure 3.5 Principle of SVM method.

As the kernel functions, polynomial functions, gaussian function or popular RBF function is often used. New axes are created and instances are recalculated in this space using kernel function. The finding of hyperplane in higher dimension dataspace is realized by usage of quadratic programming. The most important for hyperplane finding are support vectors (Figure 3.6).

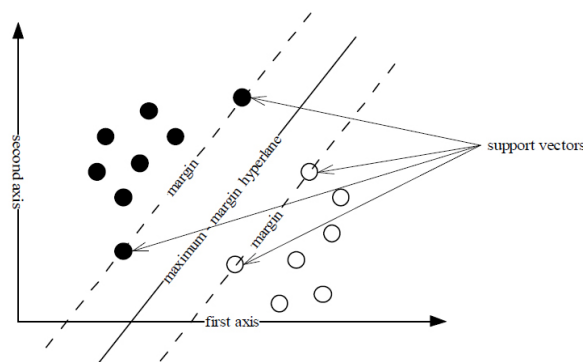


Figure 3.6 Classification by using SVM.

As we describe SVM, they are usable only for binary classification. When we want to classify into the more than two classes, one-against-all method is used. Firstly, one of the classes is selected and SVM for classification to this class and the others class is created. After that, next class is selected. So, at the end, tree of classifiers is created. SVM method is very computationally fast, because of only functional transformations of data are performed. I used SMO algorithm implemented in Weka-3-6-6 to process the genotype data. This algorithm replaces all missing values and transforms nominal attributes into binary ones. It also normalizes all attributes by default. Multi-class problems are solved using one-against-all classification.

3.5.2.10 Metalearning algorithms

In this paragraph I want to describe methods usually called as metalearning methods. They enable us to combine classification power of multiple classifiers into one classification method or they are intended for „power“ learning, when a lot of iterations of basic method are performed to improve the classification. Each of used method is implemented also in Weka-3-6-6 software (Witten et al., 1999).

Boosting

Algorithm of boosting (Michie et al., 1994).

1. Assign the initial weight to each instance.

$$w_x^1 = \frac{1}{N}.$$

2. For $t = 1, 2, \dots, T$ construct a classifier C^t as follows
 - a. calculate error of classifier as

$$\varepsilon^t = \sum w_x^t, w_x^t \in \{w \mid \text{incorrectly classified}\} \text{ is an error of classifier } t.$$

- b. if $\varepsilon^t > 0.5 \Rightarrow T = t - 1$, stop.
 - c. if $\varepsilon^t = 0$, stop.
 - d. else

$$(i) w_x^{t+1} = w_x^t \frac{\varepsilon^t}{1 - \varepsilon^t},$$

$$(ii) \sum_x w_x^{t+1} = 1, w_x^t \in \{w \mid \text{incorrectly classified}\}.$$

3. weighted voting between classifiers, each classifier C^t has weight $\log\left(\frac{1 - \varepsilon^t}{\varepsilon^t}\right)$.
-

Typical example of boosting methods is Adaboost M1 method (Witten et al., 2011). Only nominal class problems can be tackled. This method often dramatically improves performance, but sometimes overfits. Experiments on generated datasets proved that boosted classifiers have ~15 % lower error than non-boosted ones (Freund and Schapire, 1996).

3.5.2.11 Bagging

Bagging is used for a any basic classifier to reduce variance in datasets.

Bagging algorithm

1. For $t = 1, 2, \dots, T$ trials with randomly generated training datasets with N instances with repeats.
 2. For each $t = 1, 2, \dots, T$ make a classifier C^t .
 3. For new classified instance perform a voting on classifiers - C^* - the most frequently class – is voted.
-

3.5.2.12 Voting

Using different methods of voting we are able to combine number of classifiers. Each classifier can be usable for different task and different dataset, but many real problems generated datasets which contains few types of problems and also character of data may be different during the sampling e.g. according to noise or changed type of measuring. Many combinations of probability estimates for classification by a voted classifier combined from different type of basic classifiers, i.e. to perform voted method on classifiers are available:

- Average of probabilities,
- Product of probabilities,
- Majority voting,
- Minimum probability,
- Maximum probability,
- Median.

3.5.2.13 Evaluation of classification models

The main classification task is to build a model which can separate instances to their classes most accurately. Training of classification models can be evaluated with a plenty of methods (10-fold cross validation, training set validation, test set validation, bootstrapping). The goal of evaluation methods is to show how model created using training data will classify a new instances – i.e. how robust model is. Basic evaluation method is training set validation – we build a model using training data and perform a classification on the same dataset. However, no information about robustness is given by this method. We can only evaluate, how good is the model fit, how good model describe training set. It is also useful, when we need to derive representation of large dataset in comprehensible form (e.g. decision rules or trees).

We can also divide dataset into two groups – training examples and testing examples. The portion between their sizes is not strictly given. After model was inferred, we use an training dataset and perform a classification on them. But, derived model parameters are highly depend on algorithm of both dataset sampling.

Most usable for derivation robust metrics about model classification power is n-fold cross validation method (Witten et al., 1999; Berka, 2001). During the cross validation, dataset is divided into n parts. n-1 parts are given for model built (learning phase) and remaining one is used for model testing. n iterations are performed and in each iteration another

part of dataset is used as testing one. Afterthat, results of performed n classifications are averaged and output for model inferred for the whole dataset. Special case (but unusable due to computer time consumption in many machine learning methos) of n -fold validation is leave-one-out method where n is set to dataset instances count. In this work, every machine learning method is validated using 10-fold cross validation which is considered as robust enough (Witten et al., 2011).

In case of genetic (MS) data, we can discuss the results as precission of classification calculated on many levels as the genetic distance given by chosen model. Also we can discuss an errors of classifications which leads to similarity of instances. Models with high precission of classification based on genetic variability data can be used for breed or another chosen level discrimination.

Most used methods for calculation of model usability metrics using any chosen of validation method are mainly:

- confusion matrix,

| | | Predicted class | |
|--------------|---|-----------------|----------------|
| | | + | - |
| Actual class | + | True Positive | False negative |
| | - | False Positive | True negative |

- overall accuracy and error,

$$Acc = \frac{TP + TN}{TP + TN + FP + FN},$$

$$Err = \frac{FP + FN}{TP + TN + FP + FN}.$$

- accuracy and error calculated for each class,
- TP and FP rates,

$$tp = \frac{TP}{TP + FN} \times 100,$$

$$fp = \frac{FP}{FP + TN} \times 100,$$

- precission and recall,

$$recall = tp,$$

$$precission = \frac{TP}{TP + FP} \times 100,$$

- F-measure,

$$F - measure = \frac{2TP}{2TP + FP + FN}$$

- Kappa Statistics which is an index which compares the agreement against that which might be expected by chance. Kappa can be thought of as the chance-corrected proportional agreement, and possible values range from +1 (perfect agreement) via 0 (no agreement above that expected by chance) to -1 (complete disagreement) (Eugenio and Glass, 2004).

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)},$$

$$P(E)_{S\&C} = \sum_j \left(\frac{\sum_i n_{ij}}{Nk} \right)^2$$

- Mean absolute error

$$AD = \sum_{i=1}^N |E_i - O_i| / (N - 1).$$

- Root mean squared error

$$LSD = \sum_{I=1}^N (E_i - O_i)^2 / (N - 1).$$

- Root relative squared error

$$RSE = \sum_{i=1}^N [(E_i - O_i) / E_i]^2 / (N - 1).$$

- Relative absolute error

$$RAD = \sum_{i=1}^N |E_i - O_i| / E_i / (N - 1).$$

Where

N - number of observations or sum of weights,

E_i - predicted value of case,

O_i - observed value of case.

3.5.3 Clusterization

Clusterization methods allow us to derive the hidden structure of the data. They are used to infer clusters of instances with similar attributes values. These clusters being derived, we can discuss, in our case, inferred structure in connection with some traits or other attributes, number of inferred clusters, clusters assignments of individuals (Pritchard et al., 2000; Pritchard and Przeworski, 2001). Many methods e.g. k-means clusterization, MCMC simulations, hierarchical clustering (unknown number of clusters), EM algorithm can be performed to infer population structure. It depends on the method of clusterization if number of clusters is inferred too or if we have to know how many clusters we want to infer. Methods of clusterization are useful when we have a data, with typically a lot of attributes, and we do not know anything about their structure and relationships between structure elements (clusters). In our case, attributes are represented by high polymorphic genotype data, so during the clusterization genetically similar groups of individuals are clustered together. We also know the breed membership of this individual, but this attribute is not primarily used in analysis. We use it only for comparison of analysis results with the real state of problem and for dealing with unexpected results.

We can perform analyses for a user-defined number of clusters to find out how clusterization converges and to uncover population structure. On the basis of this type of results we can discuss if our expectations about the real state of the problem correlated with the clusters (Pritchard et al., 2000).

According to the previous paragraph, when the results are unexpected, we can continue with the experiment for another assumed number for clusters. We can reach explainable results for another assumed number of clusters when they are clearly defined – clusterization is not random.

This method is useful for the identification of the relationship among genetically different groups. In connection with the information about the breed membership of individuals, we

obtained results about variability and migration proportions in selected populations. As the results of clustering (EM, MCMC algorithms) show, the results of this type of analysis based on MS genotype data also correspond to empirical facts of breeding programmes of each breed.

Same modifications of the common methods also compute under assumptions of population genetic postulates, so we have to use datasets with numeric values (Pritchard and Przeworski, 2001) what means e.g. the length of microsatellite repetitions. Results of clusterization are usually:

- number of inferred clusters (in case they are unknown),
- the probabilities of membership of the individual in cluster,
- the probabilities of membership calculated on the population (or any aggregation) level,
- main aggregation attribute in inferred clusters,
- calculated distances of clusters. Usage of cluster methods to describe genetic structure of sampled subpopulations.

Definition of the problem

Problem is to explore clusters in genotype dataset of predefined cattle breed to describe their genetic structure, i.e. find an algorithm of clusterization which produce uniform clusters based on genotype data or produce clusters most similar to predefined breeds.

In context of machine learning, cluster analysis is used to divide the samples/datarows into the clusters with similar attributes, i.e. clusters have describable properties based on attributes. These methods are example of methods based on learning without teacher. All of algorithms works with probability that instance belongs to the cluster, but the known classification of instance have not influence to clustering. We can only use preclassified instances to prove the clusterization. We can divide these algorithms into two groups – methods in first of them can estimate number of clusters, for methods in second group, number of clusters has to be specified (Berka, 2001). Basic method of clusterization is hierachical structuring.

3.5.3.1 Hierarchical clustering

Initialization

1. calculate distances between all of instances
2. all of instances are in separate cluster

Main procedure

1. until more than one cluster exists
2. find most similar clusters and join them
3. calculate new distances between all of clusters

A lot of problems are connected with hierachical clusterization. When the number of clusters reflects best the real state? How to choose optimal stop step in clusterization? The order of instances influence results. Another principles of clusterization offer the solution

based on statistical clusterization. However, the accuracy of the clusters assignments when genotype data are analyzed depends on a number of loci, the amount of admixture, and the extent of allele- frequency differences among populations.

3.5.3.2 K-means algorithm

K-means algorithm is one of the basic clusterization algorithms. It needs to specify apriori number of clusters. In comparison of hierarchical clusterization, the results are not too dependent on order of instances, but we can not inspect process of clusterization. So, for optimal clusterization when we do not know this number, we need to perform clusterization several times for optimal results. E.g. algorithm implemented in Weka-3-6-6 (Witten et al., 1999) seems to be usable to cluster genotype datasets.

k-means clustering algorithm

1. Randomly separate instances into the k clusters
 2. Estimate centroids for each cluster in actual separation
 3. Estimate distance from each centroid for each instance
 4. Remove instance in the cluster which centroid minimalize the distance
 5. If a remove happens go to 2, otherwise exit
-

3.5.3.3 EM algorithm

The EM algorithm assigns a probability distribution to each instance which indicates the probability of it belonging to each of the clusters. The EM (Expectation Maximalization) algorithm described bellow is implemented also in Weka-3-6-6 software (Witten et al., 1999). Algorithm can predict optimal number of clusters using a cross validation, so we do not need to specify it. After clusters number estimation, generalization of k-means clusterization is used under mixture model to estimate optimal probabilities of instances belonging to clusters. Two steps are used – expectation step, when probabilities of instances are calculated and maximatization step, when new average probabilities of clusters are estimated. These two steps iterations converge to optimal solution of clusterization measured by overall likelihood (Witten et al., 2011).

Basic EM algorithm

Initialization

1. Randomly choose distribution parameters – probability that instance belongs to the cluster

Iteration

1. Calculate values of hidden elements on parameters of distribution – cluster average probabilities and instance probabilities
 2. On actualized data calculate maximal likelihood estimate of distribution parameters
 3. If distribution parameters were changed go to 1 else end
-

The cross validation performed to determine the number of clusters is done in the following steps.

Cross validation

1. the number of clusters is set to 1
 2. the training set is split randomly into 10 folds.
 3. EM is performed 10 times using the 10 folds the usual CV way.
 4. the loglikelihood is averaged over all 10 results.
 5. if loglikelihood has increased the number of clusters is increased by 1 and the program continues at step 2.
-

3.5.3.4 Markov chains

A Markov chain is a sequence of random variables X_1, X_2, X_3, \dots with the Markov property, namely that, given the present state, the future and past states are independent. Formally, $\Pr(X_{n+1} = x \mid X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = \Pr(X_{n+1} \mid X_n = x_n)$.

The probability of being state x_i at the „time“ $n + 1$ is conditioned only by state n and not by the previous states. Strictly speaking this is a first order Markov chain. Figure 3.7 shows an example of a Markov chain. Markov chains are often described by a directed graph, where the edges are labeled by the probabilities of going from one state to the other states.

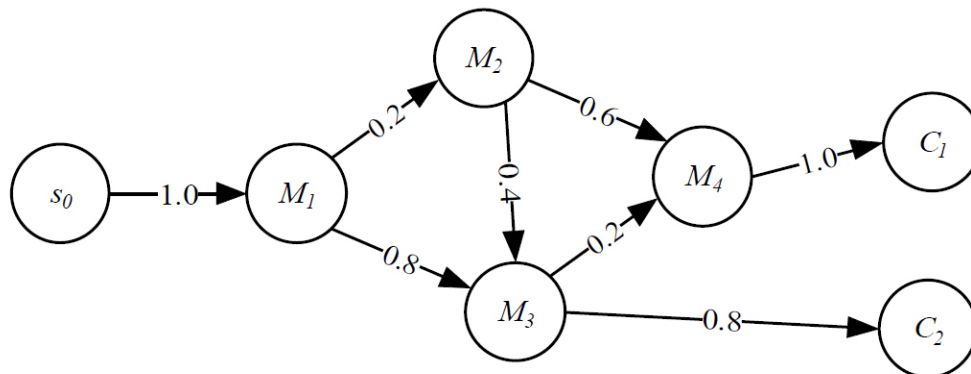


Figure 3.7 Example of Markov Chain to clustering/classification based on genotype data. Figure 3.7 shows the example of Markov Chain which can be used for clusterization of genotype data. States of Markov chain labeled M_1, \dots, M_4 represent genotypes of four loci, s_0 represents a start state of Markov chain and states labeled C_1, C_2 represent two clusters of individuals inferred using genotype data. When classes of clusters are specified, described Markov chain will classify individuals according to their class (e.g. breed). This Markov chain shows probabilities of transitions from each state based on genotype data. The figure shows the state when Markov Chain is „trained“, so the probabilities of transitions from each state are determined. How do we obtain this solution?

The precise solving of this problem represents a hard complexity problem (exponential number of combinations should be computed to calculate accurate probabilities which reflect original dataset). A lot of approximative algorithms were created to avoid the complexity of exact solution. They found an optimal solution of probabilities computation, so they find a local minimum of the problem.

One of these approximative algorithms is called Monte Carlo, so the Markov chain

created by this method is called Monte Carlo Markov chain. Monte Carlo approximation is based on intensive simulations based on random walking through the dataspace. First phase consist of probabilistic determination of transitions. Algorithm selected random seed what represents randomly selected instance of data. Afterthat, random walking through dataspace is performed. Neighbouring instances are selected and they are processed in initial Markov chain (where all of states can be joint) to asses transitions between states (which represent attributes of data). This phase is called „burn-in period“. The second phase (called „after burn“) serve to estimate of probabilities of transitions between the states and it is also performed by random walking in dataspace. Throught the random walking, each instance in dataset can be choose more than one times, so we do not need the large dataset to built Monte Carlo Markov Chain.

But, there are a lot of problems connected with this approximative algorithm. The result is highly depent on the chosen random seed and we can obtain different result during the repetition of processing. The other problem is to estimate length of burn-in and after burn periods on which the result is also highly depent. These problems are large discussed, but no appropriate resolution was given nowadays. The general recommendations say, that both of two phases should be as long as it is possible due to computational time and to process more than one computation to prove the results.

Method recommended for clusterization task of genotype data was developed by (Pritchard et al., 2000). Described method is based on MCMC clusterization to the defined number of clusters. Also, software implementation of this method exists – STRUCTURE 2.0 software. As the other clusterization methods, this method is also useful for indetifying populations and assigning individuals with a little information about population structure for indetifying populations and assigning individuals with a little information about population structure as (Pritchard et al., 2000) wrote.

3.6 Microsatellites in Cattle Studies

Microsatellites are abundant highly polymorphic markers well dispersed over the genome. They have been described as length variations within tandem arrays of short nucleotide motifs and are unequivocally defined by specific sequences of primers in PCR. They have been shown to be useful for a variety of purposes, such as genome mapping, parentage determination, legal medicine, disease research, cancer research, and determination of genetic variation (Hancock et al., 1999). Several studies have shown that microsatellites can be used to identify the population of the origin of an individual e.g. (Paetkau et al., 1998; Rannala and Mountain, 1997; Cornuet et al., 1999) and for the estimation of genetic diversity and relationships among livestock breeds (Buchanan et al., 1994; Saitbekova et al., 1999; Schmid et al., 1999). The high variability of microsatellites and their distribution give them advantages over other markers. (Arranz et al., 1996) showed that microsatellite loci were much more useful than protein markers in determining heterozygosity and genetic distances between Brown Swiss and three other breeds of Spanish cattle.

Recently, microsatellites represent the commonest markers used for population genetic studies of cattle. Applications of this technique was successful in characterization of cattle breeds throughout the world, it is elucidation of origin, migration and admixture of cattle breeds during domestication, assessment of intrapopulation diversity, genetic differentiation and relationship of modern cattle populations (Kantanen et al., 2000; Canon et al., 2001; Hanotte et al., 2002; Beja-Pereira et al., 2003; Troy et al., 2001; Freeman et al., 2004; Freeman et al., 2006). These studies have progressively used common sets of microsatellite markers thus facilitating comparative surveys of diversity and relationship and the consolidation and analysis of large datasets for multiple breeding, evolutionary and conservation applications. Using an array of microsatellite markers, it is thus possible to use individual genotype information to determine the source population with a high rate of confidence (Bjornstad and Roed, 2002; Fan et al., 2002; Koskinen, 2003).

Up to now most studies have focused on a small set of microsatellite loci, typically the ones suggested by the FAO. The microsatellite markers used according to FAO recommendations are *BM1824*, *CSSM08*, *CSSM33*, *CSSM60*, *CSSM66*, *ETH3*, *ETH10*, *ETH225*, *HAUT27*, *HEL01*, *HEL5*, *HEL09*, *ILSTS005*, *ILSTS006*, *ILSTS011*, *ILSTS033*, *ILSTS034*, *INRA05*, *INRA63*, *INRA35*, *MM8*, *MM12*, *TGLA53*, *TGLA122* and *TGLA227*. Also set recommended by ISAG is widely used in population genetic studies, thanks to routine genotyping of this set in numerous countries and consecutive easy comparison of obtained results. A set of the ISAG recommended microsatellite loci: *BM2113*, *BM1824*, *ETH3*, *ETH10*, *ETH225*, *INRA23*, *SPS115*, *TGLA53*, *TGLA122*, *TGLA126* and *TGLA227*.

3.7 Diversity in Studied Cattle Breeds

Cattle holds a unique position among domestic livestock species thanks to its key agriculture, economic, cultural and even religious roles in historical and current societies (Bradley et al., 1998). Most of the productive dairy and beef cattle breeds kept worldwide are of European origin (Lenstra, 2006). Along with the evolutionary forces (gene flow, genetic drift and natural selection), artificial breeding practices (e.g. artificial insemination and embryo transfer) have been the main process influencing the genetic diversity of domestic animal populations (Hall, 2004; Li et al., 2007). Highly selected and intensively managed breeds such as the dairy Holstein Friesian cattle have now grown in numbers at the expense of local cattle breeds, which have become endangered or extinct. At the same time, the intensive selection of top sires resulted in more uniform populations and in decrease of the effective population sizes. These developments lead to the loss of genetic variation and adaptations to local conditions or extensive management (Barker, 1999). Loss of genetic variability and inbreeding in small populations is the main focus of attention in conservation genetics (Hall, 2004). Description of biodiversity in cattle is important as an aid to conservation of animal genetic resources and national heritage.

Genetic diversity can be evaluated on the basis of the number of alleles per locus, heterozygosity and polymorphic information content (Czernekova et al., 2006). Machugh et al. (1997; 1998) used 20 microsatellites for clarifying the genetic relationships between cattle populations from Africa, Europe and Asia and provided support for a separate origin of domestication for *Bos taurus* and *Bos indicus* cattle. Kumar et al. (2003) used 20 microsatellite markers for studying extent of genetic differentiation among breeds of cattle from India, Europe and the Near East. When considering European breeds, similar values of genetic diversity have been obtained using microsatellite data: 11.2% of similarity for 7 European breeds (MacHugh et al., 1998), 10.7% for 20 northern European breeds (Kantanen et al., 2000), and 6.8% for 18 southwestern European cattle breeds (Beja-Pereira et al., 2003), 9.9% for European breeds (Gautier et al., 2007).

3.7.1 Hereford

Studies on genetic diversity of Hereford breed are scarce. Blott et al. (1998a) subsampled Herefords from Canada and the United Kingdom and found significant genetic differences between countries. Furthermore, all the Hereford populations were significantly different from the six other breeds tested (Aberdeen Angus, Chianina, Limousin, Shorthorn, Simmental and Sussex, $P < 0.00001$). According to their findings, Canadian Herefords were more homozygous than cattle in other countries, and displaced almost completely British Hereford genetics in a significant proportion of the British Hereford population. The expected average heterozygosities ranged between 0.19 and 0.26. The lowest heterozygosities were observed in the Canadian polled and in the Canadian horned groups (1960s). As for the British groups, the lowest heterozygosity was found in „traditional“ Hereford, the highest ones were observed in the „hybrid“ animals (1970s), and in Swedish Herefords. Other studied breeds had generally higher heterozygosities than the Hereford groups, with the exception of the Sussex breed. Number of alleles observed in the Hereford groups ranged from 48 in the „traditional“ group to 60 in Swedish, and 61 in Irish Herefords. There are several earlier studies focused on inbreeding in the Hereford population. Willhalm (1937)

sampled population lines from the entire breed in 1930 and calculated a mean inbreeding coefficient of 8.1%. Stonaker (1951) reported 30.7% inbreeding for a closed herd in 1947, and Russell et al. (1984) reported 37.0% inbreeding for another closed line in 1984.

3.7.2 Holstein

The Holstein breed is known worldwide as one of the top yielding dairy breeds. Breeding strategies to improve milk production, based on the import of purebred Holstein heifers and semen have been implemented by many developed and developing countries over the last 40 years. The strong selection for milk production in conjunction with the extensive usage of artificial insemination has reduced the genetic diversity within this breed as apparent in the data of (Mc Kay et al., 2008). No significant divergence is evident between geographically separated populations of Holstein cattle probably due to mentioned facts, historic occurrences of gene flow between populations and selection for similar traits. Maudet et al. (2002) observed low number of alleles per locus (5.83) and the heterozygote deficit; mean observed and expected heterozygosities were 0.669 and 0.686, respectively, in the sample of Holstein bulls. Furthermore, the exact test for Hardy-Weinberg disequilibrium within breeds showed a significant deviation in the French Holstein breed ($P < 0.0001$). The average heterozygosity in Holstein population for the 17 loci in the study of (Del Bol et al., 2001) was 0.68. Total of 110 alleles were found in the Holstein breed. Czernekova et al. (2006) reported the lowest variability in microsatellite loci in Holstein cattle in comparison to other breed studies in their work. According to (Hanslik et al., 2000) the low number of alleles in Holstein cattle is resulting from an intensive selection. Furthermore, in the study of (Czernekova et al., 2006) or (Del Bol et al., 2001) Holstein was found as the most divergent breed. Similarly, Hanslik et al. (2000) showed that the Holstein breed is highly structured between the Old World and the New World populations. The possible explanation for this findings is that the Holstein breed was created relatively recently in comparison to other breeds. On the other hand, Del Bol et al. (2001) found that Italian breed Burlina (among all the Italian breeds) was the closest to Holstein (0.272) probably because of the common origin of these two breeds. Machado et al. (2003) who analysed the genetic diversity within and among four cattle breeds found that Holstein breed was the most distinct from the other breeds: 1.15 in relation to Gyr, 1.12 in relation to Nellore and 0.94 in relation to Guzerat. In their study, 64 alleles were detected in all four breeds using 9 microsatellite primers. The average heterozygosity detected for the nine loci was 35 % and the expected value for Hardy-Weinberg equilibrium was 53 %. Further, each breed showed 53% of the total number of alleles. The average number of alleles per locus was 7.11 +/- 3.21.

3.7.3 Piedmontese

Studies on genetic diversity of Piedmontese breed are scarce. Ciampolini et al. (1995) used 17 bovine microsatellite system to study four different Italian breeds (Chianina, Marchigiana, Romagnola and Piedmontese). They observed 181 alleles in total, considering these four breeds as a whole. The average number of alleles per microsatellite was 10.59. Furthermore, they discovered some Piedmontese-specific alleles, sometimes with relatively high frequencies (36% for INRA5, allele 4, 27% for INRA5, allele 5 and 20% for INRA16,

allele 10). The sum of the frequencies of typical alleles in the Piedmontese breed shows an average of 3.4 specific alleles which are never found in other breeds. Genetic distances calculated for these four breeds demonstrated that the Piedmontese is the most distinct of all four breeds. Recently, Moioli et al. (2004) studied three native Italian cattle breeds, Piedmontese, Maremmana, and Podolica using 21 microsatellites located on 13 chromosomes. As for the Piedmontese, mean number of alleles per locus was 7.3. Average gene diversity over all loci in the Piedmontese breed was 0.738. The values of heterozygosity, both observed and expected was 0.148 (+0.083) and 0.163 (+0.083), respectively. Inbreeding rate was 0.102. Genetic distances were as follows: 0.069 (Piedmontese versus Maremmana) and 0.050 (Piedmontese versus Podolica). According to findings of (Moioli et al., 2004), 82 % Piedmontese animals analyzed in this study fit in the appropriate cluster with a probability higher than 90 %. The mentioned fact may result from the long-term selection activity made in the framework of one performance station and using AI in large extent.

3.7.4 Simmental

Del Bol et al. (2001) used 17 microsatellites to determine the genetic structure of seven Italian and six Swiss and German cattle breeds. In Simmental cattle, they found a total allele number of 95, with average heterozygosity being 0.62. Furthermore, Simmental showed a certain differentiation as compared to Brown Swiss (0.164), in spite of their common geographical origin, probably due to limited admixture. The genetic distances between Simmental and the other alpine breeds were roughly similar (0.204 ± 0.337).

3.7.5 Fleckvieh

Czernekova et al. (2006) found a very close similarity between Fleckvieh (Czech Pied cattle) and Slovak Pied breeds that is a result of an analogous breeding programme in the former Czechoslovakia. The mentioned study and the study of (Čítek and Řehout, 2001) demonstrated the highest values of heterozygosity, polymorphic information content, and effective population size in Czech Pied Cattle.

3.7.6 Limousin, Charolais, Aberdeen Angus

For the Limousin breed, Maudet et al. (2002) found the mean number of alleles per locus of 5.78. The values of heterozygosity, both observed and expected, was 0.674 and 0.675, respectively. In the same study, mean number alleles per locus in Charolais cattle was 6.00. Mean observed heterozygosity was 0.640 and mean expected heterozygosity was 0.661. Maudet et al. (2002) and Moazamigoudarzi et al. (1994) determined that the Limousin and Charolais breeds clustered together and were clearly differentiated from the dairy breeds, suggesting a possible common origin or recent gene flow between these two breeds. Russell et al. (2000) used Aberdeen Angus, Hereford, Charolais and Simmental as comparisons to be er define Criollo cattle. They found that the Charolais breed is the most similar or closely related to Criollo cattle, and the Angus breed is the most different. According to (Russell et al., 2000), geographic location could be possible explanation for relatedness of the Criollo and

3 Literature Review

the Charolais. The Criollo originated from Spanish animals, and the Charolais originated in France. The Angus originated in Scotlandland, therefore, was geographically distant from the Criollo. Russell et al. (2000) used the methods described in (Nei, 1972) for determining genetic distances between Simmental and Charolais (0.405), Simmental and Aberdeen Angus (1.569), Simmental and Hereford (0.944), Charolais and Aberdeen Angus (0.750), Charolais and Hereford (1.651) and finally Angus and Hereford (1.895).

4 Material and Methods

4.1 Microsatellite loci

Ten microsatellite loci (*BM1824*, *BM2113*, *ETH3*, *ETH10*, *ETH225*, *INRA023*, *SPS115*, *TGLA122*, *TGLA126*, *TGLA227*) considered by ISAG (International Society for Animal Genetics) for individual identification (genetic type) and parentage verification are used for routine genotyping in the Czech Republic.

Individuals included in datasets were genotyped in Lamgen laboratory (registered laboratory no. 1030.3 accredited by ČIA according ČSN EN ISO/IEC 17025) on the Department of Animal Morphology, Physiology and Genetics, Faculty of Agronomy, Mendel University, Brno. Genotyping of the 10 microsatellite markers was performed by usage multiplex PCR using StockMarks® Genotyping kit and for detection of PCR products DNA sequencing machine ABI PRISM™ 3100 Genetic Analyzer (Applied Biosystems, Foster City, CA, USA) was used.

4.2 Datasets

4.2.1 General Dataset - Purebred Individuals

For the purpose of this work, 3300 from 7776 total (genotyped from 2002 to July 2009) purebred genotyped individuals with breed attribute declared were selected using the software support described in following text. The breed counts included in this basic dataset are illustrated in table 4.1 with abbreviations used below in following chapters.

| Breed | Breed abbreviation | Number of individuals |
|--------------------|--------------------|-----------------------|
| Czech Simmental* | SM100 | 730 |
| Charolais | T100 | 705 |
| Aberdeen Angus | G100 | 700 |
| Czech Fleckvieh** | C100 | 363 |
| Holstein | H100 | 243 |
| Limousin | Y100 | 188 |
| Hereford | U100 | 137 |
| Piedmontese | P100 | 125 |
| Blonde d'Aquitaine | Q100 | 73 |
| Galloway | W100 | 66 |

Table 4.1 Numbers of individuals of selected breeds in general dataset. *Czech Simmental is here defined as beef breed of Simmental cattle kept in Czech Republic. **Czech Fleckvieh is here defined as dual purpose cattle breed of Fleckvieh type, known as České červenostrakaté.

Individual level of identification attribute was set to “laboratory number” – unique number within laboratory operations. This attribute can be used as unique identification key of individual without regard on declared data. It is used it for the control purposes – if some insane results appear, it is possible to find an errors in dataset by return. Individual is represented in the general dataset as:

$$individual_{general\ DS} = [laboratory\ number, breed, m_1, \dots, m_{10}].$$

laboratory number: string – unique identification key,

breed: string $\in B = \{SM100, T100, G100, C100, H100, Y100, U100, P100, Q100, W100\}$,

$M = \{BM1824, BM2113, ETH3, ETH10, ETH225, INRA023, SPS115, TGLA122, TGLA126, TGLA227\}$,

m_1, \dots, m_{10} : integer+string+integer = $m_i[a_1/a_2] \in M$, where

$m_i[a_1 / a_2]$ is genotype in locus (marker) i of M , $a_{1,2}$ represent the length of repetitions in alleles,

$a_1 < a_2$, "?" is missing allele sign with length ∞ , "/" is allele separator sign.

So, the general dataset is defined as $generalDS = \{individuals \mid individuals : individual_{generalDS}\}$.

4.2.2 Crossbred dataset

For purposes of machine learning methods used for breed discrimination, dataset of 380 crossbred individuals with various portions of breed attribute declared were selected. In all of individuals Czech Fleckvieh breed is included in breed declaration, as the Czech Fleckvieh is mostly used for producing crossbreds under the conditions in the Czech Republic. This dataset *crossDS* is used as control dataset in machine learning processing and the results of classification task can be inspected afterthat to explore if they are usable also for breed mixture prediction. In all tasks, where this dataset is used, it is valid that none of instance from *crossDS* is not used in training set. Also, $crossDS \cap generalDS = \emptyset$, type of *crossDS* = type of *generalDS* with exception:

$$breed_{crossDS} : string \notin B_{generalDS} = \{SM100, T100, G100, C100, H100, Y100, U100, P100, Q100, W100\},$$

$$breed_{crossDS} : string \notin B_{crossDS} = \{breed\ portion \mid breed\ portion : string\}.$$

Union of crossbred and general datasets are used to show and describe genetic variability.

4.2.3 Machine learning datasets

In case of unknown gametic phase, we need to define representation the genotype data of microsatellite loci. So, we do not know which allele in genotype representation $m_i = [a_1/a_2]$ is on the first position in fact. The genotyping method does not allow us to derive this knowledge a priori. Because of machine learning methods are created for categorical data which represent in one attribute complex knowledge of this attribute (problem 1), we need to define form of genotype data attribute which contains knowledge about two alleles which position in genotype representation is unknown (problem 2). In other words, the position of one allele has to be standardized because it is the key aspect of machine learning presumptions. So, three types of genotype data representation are decided for the purpose of usage machine learning methods to breed discrimination task and they are evaluated for this task in this work.

4.2.3.1 Genotype dataset

The first type of the dataset used for machine learning is the genotype dataset described above as *generalDS*. Two alleles a_1, a_2 in genotype representation $m_i = [a_1/a_2]$ are ordered according to their length. Imagine two genotypes $[195 / 210], [178 / 195]$ of two different individuals in the same loci. They will be considered as different by any machine learning method but they contain half similar information about genotypes in fact. However, this representation does not reflect the real state and can distort real relationships in described populations if it is used in machine learning methods, the results show, that is is usable in breed discrimination (Dawson and Belkhir, 2001; Manel et al., 2002; Masuda and Pella, 2004; Burócziová and Říha, 2009).

4.2.3.2 Allele-length dataset

In this representation of genotype data, all loci are represented by two attributes, the first expressed the shorter allele, the second attribute the longer one. The *allelelengthDS* is defined as *generalDS* with exception of:

$$\underset{allelelengthDS}{individual} = [laboratory\ number, breed, m_1, \dots, m_{20}].$$

m_1, \dots, m_{20} : integer = $m_{ij}[a_j]$, where

$m_{ij}[a_j]$ is genotype of allele j in locus i , $a_j \in N + ?$, a_j represents the length of allele repetition,

$m_{ij}, m_{(i+1)k} \in M, a_j < a_k$ where i is odd.

So, now the genotype information is not reduced (problem 1 is resolved), and machine learning methods can use this information in classifications. But the problem 2 is still present, because of independent attributes expressing genotype information in one loci which are dependent in fact.

4.2.3.3 Allele-frequency dataset

The method which can avoid us partly the simple order according to allele lengths described above and reflect the real genetic relationships in populations better is to order alleles according to their frequencies in population sample. So, the third dataset *allelefrequencyDS* is defined as *allelelengthDS* with exception of:

$$\underset{allelefrequencyDS}{individual} = [laboratory\ number, breed, m_1, \dots, m_{20}].$$

m_1, \dots, m_{20} : integer = $m_{ij}[a_{ij}] \in M$, where

$m_{ij}[a_j]$ is genotype of allele j in locus i , $a_j \in N + ?$, a_j represents the length of allele repetition,

$m_{ij}, m_{(i+1)k} \in M, frequency(a_j) < frequency(a_k), i$ is odd.

Function $frequency(a_i)$ returns frequency of allele a_i computed across the whole dataset *generalDS*.

4.3 Used methods

4.3.1 Description of the genetic diversity and characterisation of the selected cattle breeds in the Czech Republic

Definition of the problem

Problem is to describe effectively genetic diversity based on microsatellite markers of the selected cattle breeds in the Czech Republic and to create comparable results of their genetic characterization.

Selected methods

All selected methods to describe the genetic diversity are implemented in PowerMarker V3.25 software (Liu, 2006). Most of methods implemented in this software are designed according (Weir, 1996). These methods are namely:

- number of observations and availability,
- major allele frequencies,
- number of alleles and genotypes,
- within-population inbreeding coefficient,
- allele and genotype frequencies,
- observed and expected heterozygosities,
- selected genetic distances (Euclidean, Shared Allele, Golsdstein, Nei 72, Shriver, Slatkin),
- vizualization of genetic distances (UPGMA and neighbour-joining genetic trees).

For all of these summary statistics, nonparametric bootstrapping (1000 repetitions) across loci is used for variances and confidence intervals estimation. The bootstrap is based on the statistical procedure of sampling with replacement. The idea is to built datasets randomly with replacement, perform calculations on these datasets and average the results. More real-like distributions and variances can be obtain by this method. It is also good solution for situation when nonequal numbers of individuals are included in dataset groups (e.g. nonequal number of individuals of each breed) (Witten et al., 1999; Witten et al., 2011). Also, for genetic distances visualization, PhyloDraw 1.2.2 (Choi et al., 2000), ATV 4.00 Alpha 5 (Zmasek and Eddy, 2001) and Archeopteryx 0.96 (Han and Zmasek, 2009) software implementations were used.

4.3.1.1 Number of observation

The number of observation for a microsatellite loci is formulated as the number of nonmissing alleles (for haploid data) or nonmissing genotypes (for diploid data) observed in the dataset. A genotype is already considered as missing if one of its two alleles is missing (Liu and Muse, 2005).

4.3.1.2 Availability

Availability is defined as $1 - \text{Observed} / n$, where *Observed* is the number of alleles observed (succesfully genotyped) and n is the number of individuals sampled (Liu and Muse, 2005).

4.3.1.3 Within-population inbreeding coefficient

An EM algorithm described in (Weir, 1996) is used fo calculation of this measure. The parameter is estimated using method of moments and this method may converge for negative values of inbreeding coefficient.

4.3.1.4 Allele and genotype frequencies

Allele and genotype frequencies are calculated for different input datasets subdivisions in this work (on breed level, on entire dataset level etc.). These two measures are necessary to calculate as input to other methods (e.g. genetic distances calculations).

The sample allele frequencies are calculated as $\tilde{p}_u = n_u / 2n$ (Weir, 1996), with the variance estimated as

$$\text{var}(\tilde{p}_u) \triangleq \frac{1}{2n} (\tilde{p}_u + \tilde{P}_{uv} - 2\tilde{p}_u^2).$$

Where \triangleq means “estimated by”. The sample genotype frequencies \tilde{P}_{uv} are calculated as n_{uv} / n , n is the number of individuals sampled. Both the $\tilde{p}_u, \tilde{P}_{uv}$ are unbiased maximum likelihood estimates (MLEs) of the population frequencies. Confidence intervals for allele and genotype frequencies are formed by resampling individuals from the dataset (Liu and Muse, 2005).

4.3.1.5 Observed heterozygosity (H_0)

Observed heterozygosity is the observed proportion of heterozygous individuals in the population. At single locus it is estimated (Weir, 1996) as

$$\hat{H}_l = 1 - \sum_{u=1}^k \tilde{P}_{l_{uu}}$$

where, $\tilde{P}_{l_{uu}}$ is population frequency of genotype $A_u A_u$ in the l th locus of k loci.

4.3.1.6 Expected heterozygosity (H_e)

Expected heterozygosity known also as gene diversity is defined as 1.0 minus the sum of the squared gene frequencies. The values range from zero (no heterozygosity) to nearly 1.0 (for a large number of equally frequent alleles).

Expected heterozygosity is defined as the probability that two randomly chosen alleles from the population are different. It can be calculated as the common biased estimator of the gene diversity in locus l , \tilde{p}_{lu}^2 is an allele frequency of A_u in population, n is number of individuals, k is number of alleles in loci l ,

$$\hat{D}_l = \left(1 - \sum_{u=1}^k \tilde{p}_{lu}^2 \right).$$

If we want to calculate gene diversity across several loci, we need double summation and subscripting as follows:

$$1 - \frac{1}{m} \sum_{l=1}^m \sum_{i=1}^k p_i^2$$

where the first summation is for the l th of m loci.

An unbiased estimator of gene diversity based at method of moments (Weir, 1996) at the l th locus is

$$\hat{D}_l = \left(1 - \sum_{u=1}^k \tilde{p}_{lu}^2 \right) / \left(1 - \frac{1+f}{n} \right),$$

where f is an inbreeding coefficient.

4.3.1.7 Genetic distances

Genetic distances are measures of similarity between and among species, populations sub-populations or individuals. They are suitable to construct genetic trees – phylograms.

For this chapter purposes, let p_{ij} and q_{ij} be the frequencies of i th allele at the j th locus in the populations X and Y respectively, while a_j is the number of alleles at the j th locus and m is the number of loci examined.

4.3.1.8 Euclidean distance

Euclidean distance, as the most common geometric based distance is defined by:

$$D_{EU} = \frac{1}{m} \sum_{j=1}^m \sqrt{\sum_{i=1}^{a_j} (p_{ij} - q_{ij})^2}.$$

4.3.1.9 Nei's standard genetic distance

Nei (1972) standard distance has an expected value linearly related to the time since divergence, assuming that all loci have the same rate of neutral mutation, and that the genetic variation is maintained by the equilibrium between infinite-alleles mutation and genetic drift, with the effective population size of each population remaining constant.

The quantity is defined as:

$$DS = -\ln \left(J_{XY} / \sqrt{J_X J_Y} \right),$$

$$\text{where } J_X = \sum_{j=1}^m \sum_{i=1}^{a_j} p_{ij}^2 / m, J_Y = \sum_{j=1}^m \sum_{i=1}^{a_j} q_{ij}^2 / m, J_{XY} = \sum_{j=1}^m \sum_{i=1}^{a_j} p_{ij} q_{ij} / m.$$

As well, Nei's distance modification comes from (Nei et al., 1983) is widely used for calculation of geometric based distances.

4.3.1.10 Goldstein distance

A modification of the average sum of square distance method has recently been made by (Goldstein et al., 1995a). With the stepwise mutation model (SMM) assumption, Goldstein et al. (1995a) proposed that the following distance be used for microsatellite loci:

$$(\delta\mu)^2 = \frac{1}{m} \sum_{j=1}^m (\mu_{X_j} - \mu_{Y_j}),$$

where $\mu_{X_j} = \left(\sum_k k p_{kj} \right)$ and $\mu_{Y_j} = \left(\sum_k k q_{kj} \right)$ are average number of repeats found, and p_{kj} , q_{kj} are frequencies of the allele with k repeats at the j th locus in population X and population Y . The measures can be useful for estimation genetic distance closely related populations. An extension of this method, incorporating the analysis of microsatellite data into an AMOVA framework, has been recently proposed by (Michalakis and Excoffier, 1996).

4.3.1.11 Slatkin ASD distance

A distance measure closely related to $(\delta\mu)^2$ is the average square distance (ASD) defined by (Slatkin, 1995), which is given by formula:

$$ASD = \frac{1}{m} \sum_{j=1}^m \sum_{u,v} (u - v)^2 p_{uj} q_{vj}.$$

4.3.1.12 Shriver DSW distance

Shriver investigated the correlation between observed and simulation values based on the SMM. This study compared three parameters; the number of alleles, the range of allele sizes, and the number of modes in the distribution of alleles. Shriver et al. (1995) distance is defined as:

$$D_{SW} = W_{XY} - (W_X + W_Y) / 2, \text{ where}$$

$$W_X = \frac{1}{m} \sum_{j=1}^m \sum_{u,v} |u - v| p_{uj} q_{vj}, W_Y = \frac{1}{m} \sum_{j=1}^m \sum_{u,v} |u - v| p_{uj} q_{vj}, W_{XY} = \frac{1}{m} \sum_{j=1}^m \sum_{u,v} |u - v| p_{uj} q_{vj}.$$

4.3.1.13 Shared Allele Distance

Another commonly used distance, the shared allele distance DSA (Chakraborty and Jin, 1993), is defined as:

$$D_{SA} = \frac{1}{m} \sum_{j=1}^m \sum_{i=1}^{a_j} \min(p_{ij}, q_{ij}).$$

The measure $D_{LS} = -\ln(1 - D_{SA})$ usually known as Log Shared Distance has also been proposed.

4.3.2 Estimation and validation of paternity testing by microsatellite loci in selected cattle breeds

Definition of the problem

Is the panel of 10 selected microsatellites useful for paternity testing for overall and each breed?

Selected methods

Calculating of polymorphish information content (PIC) is implemented in PowerMarker V3.25 software (Liu and Muse, 2005; Liu, 2006). Another measures described above (in section) are also implemented in the software for data operations.

4.3.2.1 Polymorphism information content

Informativeness of polymorphic markers can be quantitatively measured by a statistic called the polymorphism information content, or PIC (Botstein et al., 1980). Metrics shows ability of microsatellite lenght polymorphism distinguish genotypes on small number of loci. It is also used to identify and locate a hard-to-define marker locus.

$$PIC = 1 - \sum_{i=1}^k x_i^2 - \left(\sum_{i=1}^k x_i^2 \right)^2 + \sum_{i=1}^k x_i^4,$$

where x_i – the frequency of i th allele; k – the number of alleles.

4.3.2.2 Paternity exclusion (PE1)

The probability of exclusion non correct parent, when the genotypes of offspring and both parents are known os given by formula created by (Jamieson and Taylor, 1997). One of the parent is verified based on sample of population allele frequencies.

$$PE(1) = 1 - 2 \sum_{i=1}^k x_i^2 + \sum_{i=1}^k x_i^3 + 2 \sum_{i=1}^k x_i^4 + 3 \sum_{i=1}^k x_i^5 - 2 \left(\sum_{i=1}^k x_i^2 \right)^2 + 3 \sum_{i=1}^k x_i^2 \sum_{i=1}^k x_i^3,$$

where x_i – the frequency of i th allele; k – the number of alleles.

4.3.2.3 Paternity exclusion - one parental genotype unavailable (PE2)

The probability of exclusion non correct parent, when one of genotype of parents is unknown (Jamieson and Taylor, 1997).

$$PE(2) = 1 - 4 \sum_{i=1}^k x_i^2 + 2 \left(\sum_{i=1}^k x_i^2 \right)^2 + 4 \sum_{i=1}^k x_i^3 - 3 \sum_{i=1}^k x_i^4,$$

where x_i – the frequency of i th allele; k – the number of alleles.

4.3.2.4 Parentage exclusion (PE3)

The probability of exclusion non correct parents, when the genotype of offspring and both parents are known (Jamieson and Taylor, 1997).

$$PE(3) = 1 + 4 \sum_{i=1}^k x_i^4 + 4 \sum_{i=1}^k x_i^5 - 3 \sum_{i=1}^k x_i^6 - 8 \left(\sum_{i=1}^k x_i^2 \right)^2 + 8 \sum_{i=1}^k x_i^2 \sum_{i=1}^k x_i^3 + 2 \left(\sum_{i=1}^k x_i^3 \right)^2,$$

where x_i – the frequency of i th allele; k – the number of alleles.

For paternity exclusion measures, following relationship is valid:

$$PE(2) < PE(1) < PE(3).$$

4.3.2.5 Combined Exclusion Probability

Combined Exclusion Probability is calculated for each Paternity exclusion ($PE(n)$, $n = 1, 2, 3$) type as $CEP(1)$, $CEP(2)$, $CEP(3)$ (Jamieson and Taylor, 1997).

$$CEP(n) = 1 - (1 - PE(n)_1)(1 - PE(n)_2) \dots (1 - PE(n)_k)$$

where index $1, 2, 3, \dots, k$ indicates numbers of microsatellite loci.

4.3.3 Creation of the software support for routine genotyping of microsatellite loci under the reference laboratory conditions

Definition of the problem

Problem is to create a software support for the reference laboratory Lamgen (MENDELÚ Brno) which can handle large cattle genotype datasets, can operate with them on local networks effectively and securely and allows:

- automatization of genotype and individual data inserting,
 - making effective individual and group selections based on identification data,
 - creation of individual genotyping reports,
 - performing a parentage testing,
 - calculations of MS panel usability,
 - secure operations of multiple users.
-

Selected methods

As genotype and identification data storage framework for described purposes, freeware SQL database solution Firebird 2.0 (Firebird Database Project, 2008) was selected. The main user interface is created in Borland Delphi 2005 Architect Edition development environment (Borland Inc., 2007). As described, reporting functions are needed, so freeware component (Fast Reports Inc., 2007) was used for this purpose. FreeReport is reporting tool component. It consists of report, designer and preview engines. It is fully written in Object Pascal

programming language (Delphi programming language). Capabilities of FreeReports 2.33 component used during the software creation are mainly:

- Band-oriented report generator.
- Build-in powerful designer, also available in run-time.
- Preview like in MS Word.
- Compact code - w/o designer smaller than QR1.
- Unlimited number of pages in prepared report.
- Multi-page reports; composite reports; subreports; groups;
- Multi-column reports; master-detail-detail reports;
- Cross-tab reports; two-pass reports.
- Full control over printing process; support all paper sizes.
- TXT, RTF, CSV, HTML export.
- Text search in prepared report.

For fast calculations of MS panel effectivity measures, implementation of container datatypes (hash tables, lists and vectors) is needed. Programming with these datatypes can reduce computational time radical (thanks to non linear or non exponential searching in large data structures). The component package DIContainers 3.0 (The Delphi Inspiration, 2008) implementation was used for handling described datatypes.

4.3.4 Proving of usability of machine learning methods in cattle breed discrimination task

Definition of the problem

Problem is defined as to find an machine learning algorithm and its parameters which is most appropriate to classify individuals into their breed class declared using cattle genotype data according to measures of classification power and to find most suitable data representation for usage with machine learning algorithms for the same issue - breed discrimination in cattle based on genotype data.

Subproblem: To find an appropriate form of genotype data set for breed discrimination task.

Selected methods

For comparison of results obtained in previous work in horses (genotype data of 17 MS markers) (Burócziová and Říha, 2009), methods described in mentioned paper were used:

- J48 algorithm (decision trees),
- JRip algorithm for decision rules induction,
- Naive Bayes classifier,
- Bayes Net probability classifier,
- IB1, IB5 instance based classifier;

and these additional methods not yet analyzed for this task were added:

- ZeroR as the base for result comparison,
- Support Vector Machines (SMO implementation),

- modification of IB1 classifier (distance function created for genotype data),
- algorithms for combining of classifiers (Vote classifier algorithms).

Most of described methods are implemented in Weka-3-6-6 software (Witten et al., 1999) framework, modification of IB1 classifier with new distance function was self-implemented in Borland Delphi 2005 development environment (Borland Inc., 2007). Three types of datasets *generalDS*, *allelelengthDS*, *allelefrequencyDS* described in 4.2 section were used for each method (except modified IB1 classifier where only *generalDS* dataset and designed metrics were used). Then all of datasets were processed in Weka 3-6-6 Explorer and Experimenter environment for all of proposed algorithms to find most suitable parameters for each algorithm. Only best reached results are then reported for each method.

4.3.4.1 G-metric classifier

In case of special character of genotype data described in section 4.2, special distance metric was determined. This measure reflects genetic distance of two individuals based on genotype data obtained in unknown stage of gametic phase. Metric for measuring distance of two genotypes *Gdis* is defined as follows.

Let

$$\begin{aligned}
 (i) \quad & \text{diff}_{MAX} = \max_i (|l(a_j) - l(a_k)|_i \mid \forall a_j, a_k \in A : j \neq k), \\
 (ii) \quad & A_1 = [a_{11}, a_{12}], A_2 = [a_{21}, a_{22}], \\
 (iii) \quad & |l(a_{1j}) - l(a_{2k})| = \min_i (|l(a_{1l}) - l(a_{2m})|_i \mid \forall a_{1l} \in A_1, a_{2m} \in A_2), \\
 (iv) \quad & (|l(a_{1n}) - l(a_{2o})| \mid n \neq j, o \neq k) = (|l(a_{1n}) - l(a_{2o})| \mid \forall a_{1n} \in A_1, a_{2o} \in A_2). \\
 (v) \quad & Gdis_{pair}(a_j, a_k) = \frac{1}{2} \left(\frac{|l(a_j) - l(a_k)|}{\text{diff}_{MAX}} \mid a_j, a_k \in A, Gdis_{pair}(a_j, a_k) = 0.5 \mid a_j, a_k \notin A \right), \\
 & Gdis_{pair1}(a_{1j}, a_{2k}), Gdis_{pair2}(a_{1n}, a_{2o}) \Rightarrow Gdis(A_1, A_2) = \left(Gdis_{pair1} + Gdis_{pair2} \right).
 \end{aligned}$$

Then, distance between two individuals I_1, I_2 based on genotypes of n markers is given by:

$$Gmeasure(I_1, I_2) = \frac{\sum_{i=1}^n Gdis(A_{1i}, A_{2i})}{n}, \text{ where } A_{1i} \in I_1, A_{2i} \in I_2.$$

This metrics was implemented with IB1 algorithm and Slope classifications in Borland Delphi (Borland Inc., 2007) to prove a classification power of method used *Gmeasure* on genotype dataset i.e. *generalDS*.

G-metric

- compute max diff in length
 - chose two most similar alleles compute distance
 - distance=((100-(abs(length1-length2)/max diff*100)):2)/100 compute distance for second pair
 - whole distance: 1-sum distances for pairs
-

4.3.4.2 Evaluation of classification models

For this thesis purpose, we used following measures of model quality as they are defined in 3.5.2.13:

- confusion matrix,
- overall accuracy and error,
- accuracy and error calculated for each class,
- precision and recall,
- F-measure,
- Kappa Statistics,
- graph of probability predictions for individuals on training dataset.

5 Results and Discussion

5.1 Description of the genetic diversity and characterisation of selected cattle breeds in the Czech Republic.

5.1.1 Summary results of genetic variability for microsatellite data by breeds

5.1.1.1 General dataset

| Marker | Major Allele Freq. | No. of Genotypes | No. of Obs. | No. of Alleles | Availability | H_e | H_0 | PIC | F_{IS} |
|---------|--------------------|------------------|-------------|----------------|--------------|-------|-------|-------|----------|
| BM1824 | 0.313 | 26 | 3329 | 10 | 1.000 | 0.745 | 0.712 | 0.699 | 0.045 |
| BM2113 | 0.215 | 52 | 3277 | 12 | 0.984 | 0.854 | 0.793 | 0.837 | 0.071 |
| ETH3 | 0.466 | 36 | 3284 | 10 | 0.986 | 0.700 | 0.686 | 0.661 | 0.020 |
| ETH10 | 0.566 | 31 | 3269 | 8 | 0.982 | 0.614 | 0.526 | 0.570 | 0.144 |
| ETH225 | 0.342 | 37 | 3299 | 11 | 0.991 | 0.763 | 0.718 | 0.726 | 0.059 |
| INRA023 | 0.321 | 70 | 3245 | 14 | 0.974 | 0.799 | 0.741 | 0.774 | 0.073 |
| SPS115 | 0.535 | 33 | 3053 | 11 | 0.917 | 0.661 | 0.622 | 0.629 | 0.059 |
| TGLA122 | 0.390 | 121 | 3259 | 24 | 0.979 | 0.765 | 0.689 | 0.737 | 0.100 |
| TGLA126 | 0.476 | 27 | 3250 | 8 | 0.976 | 0.667 | 0.631 | 0.616 | 0.054 |
| TGLA227 | 0.251 | 80 | 3269 | 16 | 0.982 | 0.848 | 0.818 | 0.832 | 0.036 |
| Mean | 0.388 | 51.3 | 3253.4 | 12.4 | 0.977 | 0.742 | 0.694 | 0.708 | 0.065 |

Table 5.1 Results of basic genetic variability for general dataset (n=3300).

Summary values obtained by analysis of the whole general dataset for 3300 purebred individuals are summarized in table 5.1. Under availability equals 0.977, across all ten loci 0.388 as major allele frequency, 51.3 genotypes per loci 12.4 alleles per loci were reached.

5 Results and Discussion

Expected heterozygosity was calculated as 0.742 and observed heterozygosity as 0.694. PIC was calculated as 0.708 and inbreeding coefficient F_{is} as 0.065 what represent near random mating across all loci and the whole sampled subpopulatin. Most different alleles were detected in *TGLA122* (24) which then create 121 different genotypes. As the most divergent loci was evaluated *BM2113* what create only with 12 different alleles 52 different genotypes ($H_e=0.854$, $H_0=0.793$, $PIC=0.837$) and *TGLA227* with 16 different alleles create 80 genotypes ($H_e=0.848$, $H_0=0.818$, $PIC=0.832$).

Both *ETH3* and *TGLA126* contain only 8 different detected alleles and have 31 and 27 different genotypes, respectively. They seem to be the most less divergent loci in panel for general dataset.

All of loci show positive values of F_{is} , however ~ 0.000 , except *ETH10* (0.144) and *TGLA122* (0.100). As *ETH10* is used as genetic marker as well (DeAtley et al., 2008; DeAtley et al., 2011; Meirelles et al., 2011a), we can see influence of breeding selection in this case.

5.1.1.2 Crossbred dataset

| Marker | Major Allele Freq. | No. of Genotypes | No. of Obs. | No. of Alleles | Availability | H_e | H_0 | PIC | F_{is} |
|---------|--------------------|------------------|-------------|----------------|--------------|-------|-------|-------|----------|
| BM1824 | 0.267 | 17 | 380 | 7 | 1.000 | 0.757 | 0.718 | 0.714 | 0.053 |
| BM2113 | 0.230 | 35 | 376 | 11 | 0.989 | 0.844 | 0.856 | 0.825 | -0.014 |
| ETH3 | 0.446 | 25 | 379 | 8 | 0.997 | 0.703 | 0.694 | 0.660 | 0.015 |
| ETH10 | 0.444 | 22 | 375 | 8 | 0.987 | 0.705 | 0.656 | 0.663 | 0.071 |
| ETH225 | 0.320 | 25 | 378 | 9 | 0.995 | 0.764 | 0.767 | 0.726 | -0.003 |
| INRA023 | 0.316 | 41 | 375 | 13 | 0.987 | 0.785 | 0.771 | 0.754 | 0.019 |
| SPS115 | 0.507 | 19 | 353 | 7 | 0.929 | 0.683 | 0.637 | 0.649 | 0.068 |
| TGLA122 | 0.387 | 54 | 377 | 17 | 0.992 | 0.775 | 0.682 | 0.750 | 0.122 |
| TGLA126 | 0.468 | 21 | 374 | 8 | 0.984 | 0.667 | 0.642 | 0.613 | 0.039 |
| TGLA227 | 0.270 | 61 | 368 | 13 | 0.968 | 0.854 | 0.851 | 0.840 | 0.006 |
| Mean | 0.366 | 32 | 373.5 | 10.1 | 0.983 | 0.754 | 0.727 | 0.719 | 0.036 |

Table 5.2 Results of genetic variability for crossbred dataset (n=380).

Table 5.2 summarizes results calculated for crossbred dataset in the meaning of genetic variability across all of loci. With availability across all loci 0.983, major allele frequency for dataset was calculated as 0.366, average number of genotypes as 32 (ranges from 19 in *SPS115* to 61 in *TGLA227*), average number of distinct alleles as 10.1 (from 7 in *BM1824* and *SPS115* to 17 in *TGLA122*). Regarding the most and the less variable markers, results are fully comparable to general dataset ones (see chapter 5.1.1.1). Anyway, in crossbred dataset, little bit higher results were obtained for observed heterozygosity (0.856 in *BM2113* and 0.851 in *TGLA227*), what is in accordance with minimal presence of selection pressure on crossbreds represented mainly by beef breeds and dual purpose ones. Also, values less than zero and smaller values in comparison with purebreds were obtained for inbreeding coefficient in *BM2113* and *ETH225* as signs of bulls usage for production crossbred animals selected carefully by farmers of beef cattle.

5.1.1.3 Czech Fleckvieh

| Marker | Major Allele Freq. | No. of Genotypes | No. of Obs. | No. of Alleles | Avail-ability | H _e | H ₀ | PIC | F _{IS} |
|---------|--------------------|------------------|-------------|----------------|---------------|----------------|----------------|-------|-----------------|
| BM1824 | 0.288 | 14 | 363 | 7 | 1.000 | 0.749 | 0.736 | 0.702 | 0.019 |
| BM2113 | 0.296 | 30 | 358 | 10 | 0.986 | 0.799 | 0.799 | 0.771 | 0.001 |
| ETH3 | 0.350 | 18 | 363 | 7 | 1.000 | 0.736 | 0.771 | 0.691 | -0.047 |
| ETH10 | 0.464 | 21 | 362 | 7 | 0.997 | 0.704 | 0.693 | 0.666 | 0.017 |
| ETH225 | 0.409 | 23 | 358 | 9 | 0.986 | 0.734 | 0.743 | 0.695 | -0.011 |
| INRA023 | 0.294 | 41 | 357 | 11 | 0.983 | 0.793 | 0.784 | 0.764 | 0.012 |
| SPS115 | 0.511 | 18 | 348 | 7 | 0.959 | 0.674 | 0.615 | 0.637 | 0.088 |
| TGLA122 | 0.355 | 41 | 361 | 14 | 0.994 | 0.783 | 0.723 | 0.755 | 0.078 |
| TGLA126 | 0.499 | 21 | 360 | 8 | 0.992 | 0.649 | 0.672 | 0.596 | -0.034 |
| TGLA227 | 0.291 | 43 | 358 | 13 | 0.986 | 0.833 | 0.869 | 0.814 | -0.041 |
| Mean | 0.376 | 27 | 358.8 | 9.3 | 0.988 | 0.745 | 0.741 | 0.709 | 0.008 |

Table 5.3 Results of genetic variability for Czech Fleckvieh breed (n=363).

The highest value of different genotypes and alleles were detected in Czech Fleckvieh dataset in loci *TGLA227* (43, 13) and *TGLA122* (41, 14). As the most divergent loci, *TGLA227* with $H_e=0.833$, $H_0=0.869$, $PIC=0.814$ was evaluated. Negative values of inbreeding coefficient were obtained for *ETH3*, *ETH225*, *TGLA126* and *TGLA227*. All of values of F_{IS} point Czech Fleckvieh as population near to random mating what corresponds to large breeding animals pool and carefully selected bulls used in breeding strategies in fact.

Very high average results for H_e , H_0 and PIC were reached for Czech Fleckvieh as well as in study of (Čítek and Řehout, 2001) or (Putnova et al., 2011; Radko, 2010). Similar results were also obtained in (D'Andrea et al., 2011) study for Serbian local breed called Podolica. Otherwise, it is local kept breed as Czech Fleckvieh is but not as widely influenced by other breeds, results are highly comparable - $H_e=0.73$, $H_0=0.71$, $PIC=0.70$, $F_{IS}=0.05$ in Podolian cattle in comparison with $H_e=0.745$, $H_0=0.741$, $PIC=0.709$, $F_{IS}=0.008$ in Czech Fleckvieh. As well, similar results were reached in study by (Stevanovic et al., 2010) where YU Simmental population were examined with following average results - $H_e=0.750$, $H_0=0.651$, $PIC=0.720$, average number of alleles per loci equaled to 8.273 and major allele frequency with value 0.379. In this study, also *TGLA227* was evaluated as the most divergent loci with values as follows - $H_e=0.851$, $H_0=0.733$, $PIC=0.820$.

5 Results and Discussion

5.1.1.4 Aberdeen Angus

| Marker | Major Allele Freq. | No. of Genotypes | No. of Obs. | No. of Alleles | Avail-ability | H _e | H ₀ | PIC | F _{IS} |
|---------|--------------------|------------------|-------------|----------------|---------------|----------------|----------------|-------|-----------------|
| BM1824 | 0.384 | 10 | 700 | 4 | 1.000 | 0.714 | 0.739 | 0.662 | -0.034 |
| BM2113 | 0.252 | 34 | 691 | 8 | 0.987 | 0.819 | 0.816 | 0.795 | 0.005 |
| ETH3 | 0.361 | 14 | 685 | 5 | 0.979 | 0.717 | 0.745 | 0.666 | -0.037 |
| ETH10 | 0.500 | 17 | 685 | 6 | 0.979 | 0.649 | 0.650 | 0.594 | 0.000 |
| ETH225 | 0.284 | 20 | 695 | 8 | 0.993 | 0.767 | 0.758 | 0.727 | 0.012 |
| INRA023 | 0.517 | 23 | 684 | 8 | 0.977 | 0.668 | 0.684 | 0.630 | -0.024 |
| SPS115 | 0.458 | 14 | 631 | 6 | 0.901 | 0.665 | 0.667 | 0.608 | -0.002 |
| TGLA122 | 0.620 | 30 | 685 | 12 | 0.979 | 0.575 | 0.587 | 0.543 | -0.020 |
| TGLA126 | 0.399 | 18 | 692 | 7 | 0.989 | 0.681 | 0.698 | 0.624 | -0.025 |
| TGLA227 | 0.283 | 30 | 693 | 10 | 0.990 | 0.789 | 0.766 | 0.757 | 0.030 |
| Mean | 0.406 | 21 | 684.1 | 7.4 | 0.977 | 0.704 | 0.711 | 0.661 | -0.009 |

Table 5.4 Results of genetic variability for Aberdeen Angus breed (n=700).

The similar situation to other beef breeds in the meaning of genetic variability appeared for Aberdeen Angus beef breed. As one of the most interesting result, we can reported only 4 different alleles detected in locus *BM1824*, when 700 individuals were genotyped in dataset. Anyway, with so small variability, negative value of inbreeding coefficient was calculated, what shows, as the other results, that Aberdeen Angus subpopulation in Czech is well controlled in the meaning of inbreeding avoiding and preservation of genetic variability. It can be caused by fact, that bulls of Aberdeen Angus used in Czech came mainly from very different (geographical, genetical) imported sources in early 90's and their offsprings are still used in breeding by farmers who are avoiding to use inbred animals carefully, otherwise, bulls are usually used to produce couple of generations on farms when they are selected and bought by breeder.

5.1.1.5 Holstein

| Marker | Major Allele Freq. | No. of Genotypes | No. of Obs. | No. of Alleles | Availability | H_e | H_0 | PIC | F_{IS} |
|---------|--------------------|------------------|-------------|----------------|--------------|-------|-------|-------|----------|
| BM1824 | 0.284 | 15 | 243 | 6 | 1.000 | 0.758 | 0.790 | 0.715 | -0.040 |
| BM2113 | 0.331 | 29 | 239 | 8 | 0.984 | 0.785 | 0.787 | 0.755 | 0.000 |
| ETH3 | 0.456 | 16 | 238 | 6 | 0.979 | 0.707 | 0.693 | 0.667 | 0.022 |
| ETH10 | 0.523 | 20 | 239 | 8 | 0.984 | 0.667 | 0.611 | 0.633 | 0.087 |
| ETH225 | 0.381 | 21 | 239 | 7 | 0.984 | 0.719 | 0.707 | 0.673 | 0.019 |
| INRA023 | 0.300 | 27 | 237 | 9 | 0.975 | 0.792 | 0.797 | 0.762 | -0.004 |
| SPS115 | 0.573 | 17 | 232 | 7 | 0.955 | 0.608 | 0.621 | 0.564 | -0.019 |
| TGLA122 | 0.279 | 41 | 231 | 16 | 0.951 | 0.816 | 0.784 | 0.793 | 0.042 |
| TGLA126 | 0.457 | 16 | 232 | 6 | 0.955 | 0.662 | 0.651 | 0.605 | 0.020 |
| TGLA227 | 0.302 | 45 | 235 | 12 | 0.967 | 0.833 | 0.843 | 0.814 | -0.010 |
| Mean | 0.389 | 24.7 | 236.5 | 8.5 | 0.973 | 0.735 | 0.728 | 0.698 | 0.011 |

Table 5.5 Results of genetic variability for Holstein breed (n=243).

When results in table 5.5 are inspected, we can identify sampled subpopulation (243 individuals) of Holstein breed in Czech Republic a stable and well controlled population in the meaning of its parameters of genetic diversity. Extremely carefully selected animals are recommended and used in population as it is normal all over the world in this breed. *TGLA122* with 16 alleles and 41 different genotypes ($H_e=0.951$, $H_0=0.816$, $PIC=0.793$) and *TGLA227* with 12 alleles and 45 genotypes ($H_e=0.967$, $H_0=0.833$, $PIC=0.814$) represent the most divergent markers included in this dataset. In the contrary, comparison of heterozygosity expected and observed values in these two loci shows, that large breeding effort reduced heterozygosity as well.

We recognize higher number of alleles per loci (8.5) than (Maudet et al., 2002) reported (5.83) as well as higher expected (0.735) and observed (0.728) heterozygosity in comparison with Maudet's study (0.669, 0.686) and than Del Bol et al. (2001) who showed average heterozygosity equals to 0.68. Results are then more comparable to (Machado et al., 2003) (e.g. average number of alleles per loci 7.11).

In opposite of what was reported by (Mc Kay et al., 2008), we can not say that there is evident reduction of genetic variability in subsampled population selected in Czech. As breeding associations and private companies have good control over the breeding strategies, it can be said that Holstein has comparable genetic variability with breeds which are not so sophisticatedly and hard-pressured breed.

Better results in the meaning of higher genetic variability were reached also in comparison with (D'Andrea et al., 2011). Numbers reached for Serbian Holstein population (n=34) were: $H_e=0.64$, $H_0=0.62$, $PIC=0.59$, $F_{IS}=0.05$ in comparison of average values of subpopulation equals to $H_e=0.735$, $H_0=0.728$, $PIC=0.698$, $F_{IS}=0.011$.

5.1.1.6 Piedmontese

| Marker | Major Allele Freq. | No. of Genotypes | No. of Obs. | No. of Alleles | Availability | H _e | H _o | PIC | F _{IS} |
|---------|--------------------|------------------|-------------|----------------|--------------|----------------|----------------|-------|-----------------|
| BM1824 | 0.432 | 18 | 125 | 8 | 1.000 | 0.724 | 0.672 | 0.686 | 0.076 |
| BM2113 | 0.238 | 25 | 124 | 8 | 0.992 | 0.819 | 0.871 | 0.794 | -0.059 |
| ETH3 | 0.512 | 13 | 124 | 6 | 0.992 | 0.656 | 0.702 | 0.610 | -0.065 |
| ETH10 | 0.416 | 16 | 119 | 6 | 0.952 | 0.704 | 0.697 | 0.654 | 0.013 |
| ETH225 | 0.347 | 20 | 124 | 7 | 0.992 | 0.768 | 0.758 | 0.734 | 0.017 |
| INRA023 | 0.268 | 37 | 123 | 11 | 0.984 | 0.822 | 0.854 | 0.800 | -0.035 |
| SPS115 | 0.483 | 14 | 115 | 5 | 0.92 | 0.698 | 0.722 | 0.663 | -0.029 |
| TGLA122 | 0.335 | 40 | 124 | 15 | 0.992 | 0.812 | 0.847 | 0.791 | -0.039 |
| TGLA126 | 0.365 | 13 | 122 | 5 | 0.976 | 0.733 | 0.713 | 0.688 | 0.032 |
| TGLA227 | 0.216 | 40 | 125 | 10 | 1.000 | 0.870 | 0.880 | 0.857 | -0.007 |
| Mean | 0.361 | 23.6 | 122.5 | 8.1 | 0.980 | 0.761 | 0.772 | 0.728 | -0.010 |

Table 5.6 Results of genetic variability for Piedmontese breed (n=125).

Results obtained for Piedmontese breed subpopulation sampled only within 125 individuals show similar results in comparison with another beef breed examined. As there are just small population of Piedmontese in the Czech Republic, we can expect results summarized in table 5.6 - results show population with selected bulls used, genetically near to random mating one. Results reached for *INRA023* can be mentioned in comparison with the other breeds examined - with 11 different allele and 37 different genotypes, there is difference in comparison to the other breeds, where *INRA023* is not so divergent.

At all when results are compared to the ones previous reported, (Ciampolini et al., 1995) shows higher number of alleles per loci in Italian population and 17 MS analyzed (10.59 in comparison with 8.1 in our study), so ours are more similar to (Moioli et al., 2004) who reported 7.3 alleles per locus for 21 MS loci, gene diversity (0.738) and inbreeding coefficient 0.102.

In comparison with results reached by (D'Andrea et al., 2011) in Italian Piedmontese (n=48), expected heterozygosity was found as 0.72 (in comparison with 0.761 in our observation), observed heterozygosity equaled to 0.71 (0.772), PIC=0.68 (0.728) and positive inbreeding coefficient was reached as 0.03 (-0.01).

5.1.1.7 Blonde d'Aquitaine

| Marker | Major Allele Freq. | No. of Genotypes | No. of Obs. | No. of Alleles | Availability | H _e | H ₀ | PIC | F _{IS} |
|---------|--------------------|------------------|-------------|----------------|--------------|----------------|----------------|-------|-----------------|
| BM1824 | 0.425 | 10 | 73 | 4 | 1.000 | 0.698 | 0.726 | 0.646 | -0.034 |
| BM2113 | 0.347 | 23 | 72 | 9 | 0.986 | 0.797 | 0.819 | 0.772 | -0.021 |
| ETH3 | 0.389 | 12 | 72 | 6 | 0.986 | 0.728 | 0.653 | 0.683 | 0.110 |
| ETH10 | 0.430 | 12 | 71 | 6 | 0.973 | 0.704 | 0.789 | 0.658 | -0.113 |
| ETH225 | 0.390 | 13 | 73 | 6 | 1.000 | 0.749 | 0.822 | 0.713 | -0.090 |
| INRA023 | 0.433 | 16 | 67 | 6 | 0.918 | 0.739 | 0.657 | 0.708 | 0.119 |
| SPS115 | 0.583 | 9 | 60 | 5 | 0.822 | 0.599 | 0.650 | 0.556 | -0.076 |
| TGLA122 | 0.281 | 23 | 73 | 10 | 1.000 | 0.812 | 0.849 | 0.788 | -0.039 |
| TGLA126 | 0.326 | 13 | 72 | 5 | 0.986 | 0.746 | 0.736 | 0.702 | 0.020 |
| TGLA227 | 0.342 | 30 | 73 | 11 | 1.000 | 0.809 | 0.795 | 0.789 | 0.025 |
| Mean | 0.395 | 16.1 | 70.6 | 6.8 | 0.967 | 0.738 | 0.750 | 0.702 | -0.008 |

Table 5.7 Results of genetic variability for Blonde d'Aquitaine breed (n=73).

In the subpopulation of Blonde d'Aquitaine, we can identify (table 5.7) quite a small number of alleles detected in comparison with the other breeds - *BM1824* (4), *SPS115* (5), *TGLA126* (5) - what could be caused by really small sample of individuals (73) in dataset. Positive values of inbreeding coefficient in *ETH3* and *INRA023*, if we assume that results obtained for this subpopulation are valid, can show that there is real genetic relationship between these two loci and beef yield parameters (Choroszy et al., 2006; Ciampolini et al., 2002) as Blonde d'Aquitaine is widely, longterm and hardly bred in France as well as Western Europe.

5.1.1.8 Czech Simmental

| Marker | Major Allele Freq. | No. of Genotypes | No. of Obs. | No. of Alleles | Avail-ability | H _e | H ₀ | PIC | F _{IS} |
|---------|--------------------|------------------|-------------|----------------|---------------|----------------|----------------|-------|-----------------|
| BM1824 | 0.328 | 15 | 730 | 5 | 1.000 | 0.734 | 0.725 | 0.686 | 0.013 |
| BM2113 | 0.480 | 35 | 712 | 11 | 0.975 | 0.716 | 0.719 | 0.690 | -0.003 |
| ETH3 | 0.558 | 23 | 717 | 8 | 0.982 | 0.629 | 0.660 | 0.589 | -0.049 |
| ETH10 | 0.692 | 15 | 718 | 7 | 0.984 | 0.489 | 0.500 | 0.456 | -0.023 |
| ETH225 | 0.368 | 22 | 719 | 9 | 0.985 | 0.715 | 0.694 | 0.667 | 0.030 |
| INRA023 | 0.345 | 36 | 711 | 11 | 0.974 | 0.763 | 0.765 | 0.726 | -0.002 |
| SPS115 | 0.663 | 17 | 673 | 6 | 0.922 | 0.518 | 0.525 | 0.481 | -0.011 |
| TGLA122 | 0.380 | 40 | 708 | 13 | 0.970 | 0.733 | 0.614 | 0.690 | 0.163 |
| TGLA126 | 0.479 | 19 | 700 | 8 | 0.959 | 0.615 | 0.594 | 0.541 | 0.035 |
| TGLA227 | 0.351 | 48 | 707 | 12 | 0.968 | 0.794 | 0.795 | 0.770 | 0.000 |
| Mean | 0.465 | 27 | 709.5 | 9 | 0.972 | 0.671 | 0.659 | 0.630 | 0.018 |

Table 5.8 Results of genetic variability for Simmental breed (n=730).

From results summarizing genetic variability in all observed loci for Czech Simmental, we can mention results reached for SPS115 locus. With 6 different alleles and 17 different genotypes, it has major allele frequency 0.663, so related parameters as ($H_e=0.518$, $H_0=0.525$, $PIC=0.481$) are very low as well in comparison with results obtained for other breeds. Anyway, other results are fully comparable to the other beef breed kept in Czech and observed in this work. Results are comparable to (Del Bol et al., 2001) who reported e.g. average heterozygosity equals to 0.62. Similar results were also reached in comparison with (Radko, 2010).

Stevanovic et al. (2010) reported when they examined population of Simmental cattle in Serbia, average number of alleles per loci as 8.364 in comparison with 9.0 reached in our dataset, average PIC was investigated as 0.73 in comparison with 0.630, most frequency allele (0.372) in comparison with 0.692 in *ETH10*. They reported *TGLA227* as one of the most polymorphic loci with results like 9 alleles per loci, $PIC=0.840$, frequency of major allele equaled to 0.274 as well as *INRA023* (11 alleles per loci, $PIC=0.860$, most frequent allele frequency=0.167). We can compare results obtained for our dataset - *TGLA227* has 12 different alleles with major one's frequency equals to 0.351 and $PIC=0.770$, *INRA023* has 11 different alleles with major one's frequency equals to 0.345 and $PIC=0.726$. Then, we can describe Czech Simmental as more uniform breed in comparison with Serbian population and results shown. D'Andrea et al. (2011) analyzed similar parameters in 13 of European cattle breeds. In Italian Simmental, they observed average expected heterozygosity equaled to 0.59 (0.671), expected one equaled to 0.59 (0.659), $PIC=0.55$ (0.630) and similar values of inbreeding coefficient - 0.02 (0.018).

5.1.1.9 Charolais

| Marker | Major Allele Freq. | No. of Genotypes | No. of Obs. | No. of Alleles | Availability | H _e | H ₀ | PIC | F _{IS} |
|---------|--------------------|------------------|-------------|----------------|--------------|----------------|----------------|-------|-----------------|
| BM1824 | 0.448 | 14 | 704 | 6 | 0.999 | 0.699 | 0.670 | 0.652 | 0.041 |
| BM2113 | 0.256 | 28 | 695 | 7 | 0.986 | 0.833 | 0.833 | 0.812 | 0.001 |
| ETH3 | 0.526 | 19 | 698 | 8 | 0.990 | 0.616 | 0.635 | 0.555 | -0.029 |
| ETH10 | 0.902 | 8 | 697 | 5 | 0.989 | 0.181 | 0.188 | 0.171 | -0.039 |
| ETH225 | 0.417 | 23 | 702 | 7 | 0.996 | 0.675 | 0.682 | 0.616 | -0.011 |
| INRA023 | 0.372 | 60 | 687 | 14 | 0.974 | 0.803 | 0.803 | 0.784 | 0.000 |
| SPS115 | 0.570 | 21 | 651 | 8 | 0.923 | 0.625 | 0.634 | 0.591 | -0.014 |
| TGLA122 | 0.376 | 52 | 696 | 14 | 0.987 | 0.743 | 0.751 | 0.707 | -0.011 |
| TGLA126 | 0.638 | 15 | 694 | 6 | 0.984 | 0.541 | 0.556 | 0.497 | -0.028 |
| TGLA227 | 0.253 | 48 | 697 | 10 | 0.989 | 0.834 | 0.845 | 0.814 | -0.013 |
| Mean | 0.476 | 28.8 | 692.1 | 8.5 | 0.982 | 0.655 | 0.660 | 0.620 | -0.007 |

Table 5.9 Results of genetic variability for Charolais breed (n=705).

Table 5.9 summarizes results obtained regarding genetic diversity parameters for all of loci used in subpopulation of 705 individuals of Charolais kept in Czech. It can be pointed, that results reached for *ETH10* are interesting from the point of view of genetic variability. Five different alleles were investigated in *ETH10* which create only 8 different genotypes with major allele frequency 0.902. Extremely low values for $H_e=0.181$, $H_0=0.188$, $PIC=0.171$ are then calculated for this loci. Again, *ETH10* seems to be connected with beef traits which are used in breeding strategy in Charolais widely (Choroszy et al., 2006; DeAtley et al., 2008; Moore and Hansen, 2003; Kuhn et al., 2005; Meirelles et al., 2011b; DeAtley et al., 2011; Hall et al., 2009).

The highest value of heterozygosity was reached for *TGLA227* locus what is similar to (Putnová et al., 2011). Regarding the other results, what was written in previous and following chapters about beef breeds and their genetic variability result in Czech is valid for Charolais as well (imports of selected bulls from different locations, carefully selected bulls, population near random mating one). Anyway, negative inbreeding coefficients were obtained for *ETH3*, *ETH10*, *ETH225*, *SPS115*, *TGLA122*, *TGLA126* and *TGLA227*. This points to quite large usage of AI and embryo transfer in the past which allow to select really different genetic material used for breeding.

We have found higher number of alleles per loci than (Maudet et al., 2002) as well as very comparable values of average expected (0.655) and observed (0.640) heterozygosities.

5.1.1.10 Hereford

| Marker | Major Allele Freq. | No. of Genotypes | No. of Obs. | No. of Alleles | Availability | H_e | H_0 | PIC | F_{IS} |
|---------|--------------------|------------------|-------------|----------------|--------------|-------|-------|-------|----------|
| BM1824 | 0.547 | 9 | 137 | 4 | 1.000 | 0.620 | 0.591 | 0.568 | 0.050 |
| BM2113 | 0.328 | 22 | 134 | 8 | 0.978 | 0.790 | 0.791 | 0.761 | 0.002 |
| ETH3 | 0.519 | 7 | 135 | 5 | 0.985 | 0.542 | 0.578 | 0.440 | -0.062 |
| ETH10 | 0.467 | 12 | 135 | 6 | 0.985 | 0.662 | 0.630 | 0.605 | 0.052 |
| ETH225 | 0.350 | 16 | 137 | 7 | 1.000 | 0.767 | 0.803 | 0.732 | -0.043 |
| INRA023 | 0.780 | 9 | 134 | 5 | 0.978 | 0.367 | 0.373 | 0.337 | -0.013 |
| SPS115 | 0.434 | 16 | 129 | 6 | 0.942 | 0.714 | 0.705 | 0.671 | 0.016 |
| TGLA122 | 0.601 | 18 | 134 | 10 | 0.978 | 0.601 | 0.604 | 0.572 | -0.002 |
| TGLA126 | 0.596 | 9 | 130 | 6 | 0.949 | 0.559 | 0.615 | 0.498 | -0.096 |
| TGLA227 | 0.379 | 23 | 136 | 10 | 0.993 | 0.745 | 0.772 | 0.710 | -0.032 |
| Mean | 0.500 | 14.1 | 134.1 | 6.7 | 0.979 | 0.637 | 0.646 | 0.589 | -0.011 |

Table 5.10 Results of genetic variability for Hereford breed (n=137).

Similar results as for other beef breeds were obtained at all for Hereford one and are comparable to study of (Blott et al., 1998b) regarding number of alleles observed and heterozygotes in traditional Hereford groups within breed. In Hereford, we can see reduced genetic variability in *INRA023* (compare with Charolais in section 5.1.1.9). Five different alleles resulted only into 9 different genotypes with a major allele frequency 0.780 and $H_e=0.367$, $H_0=0.373$, $PIC=0.337$. Reasons for these results are comparable with Charolais as *INRA023* could be connected with beef traits as well (Choroszy et al., 2006).

Not such inbreeding was observed in Czech subpopulation sampled as (Stonaker, 1951) or (Russell et al., 2000) reported. It can be result of general breeding strategy in Hereford as well as very carefully (or completely randomly) selected bulls or animals imported and used in Czech Republic.

5.1.1.11 Galloway

| Marker | Major Allele Freq. | No. of Genotypes | No. of Obs. | No. of Alleles | Availability | H _e | H ₀ | PIC | F _{is} |
|---------|--------------------|------------------|-------------|----------------|--------------|----------------|----------------|-------|-----------------|
| BM1824 | 0.553 | 9 | 66 | 4 | 1.000 | 0.622 | 0.621 | 0.576 | 0.010 |
| BM2113 | 0.652 | 7 | 66 | 5 | 1.000 | 0.521 | 0.561 | 0.474 | -0.068 |
| ETH3 | 0.625 | 7 | 64 | 4 | 0.970 | 0.549 | 0.594 | 0.500 | -0.074 |
| ETH10 | 0.523 | 5 | 65 | 4 | 0.985 | 0.527 | 0.462 | 0.417 | 0.132 |
| ETH225 | 0.280 | 18 | 66 | 6 | 1.000 | 0.795 | 0.773 | 0.764 | 0.036 |
| INRA023 | 0.364 | 12 | 66 | 6 | 1.000 | 0.723 | 0.667 | 0.672 | 0.085 |
| SPS115 | 0.398 | 12 | 49 | 5 | 0.742 | 0.702 | 0.551 | 0.650 | 0.225 |
| TGLA122 | 0.508 | 16 | 65 | 8 | 0.985 | 0.693 | 0.723 | 0.665 | -0.036 |
| TGLA126 | 0.589 | 6 | 62 | 3 | 0.939 | 0.557 | 0.532 | 0.488 | 0.053 |
| TGLA227 | 0.208 | 26 | 65 | 10 | 0.985 | 0.854 | 0.831 | 0.837 | 0.035 |
| Mean | 0.470 | 11.8 | 63.4 | 5.5 | 0.961 | 0.654 | 0.631 | 0.604 | 0.043 |

Table 5.11 Results of genetic variability for Galloway breed (n=66).

Only 66 individuals of Galloway were sampled in general dataset. Anyway, similar results as for the others beef breeds were obtained. Similar to Charolais, *ETH10* locus reported results should be considered as ones indicate low or reduced genetic variability with H_e=0.527, H₀=0.462, PIC=0.417. In the opposite of Charolais results, in *ETH10* there was calculated positive (0.132) inbreeding coefficient in Galloway. The quite high positive value of F_{is} (0.225) was reached as well in locus *SPS115*. This can be caused by just very genetically related animal analyzed in small subpopulation coming from common ancestors when just couple of embryos, semen dosages were imported in Czech republic instead of live animals in past. Plus, still nowadays, Galloway genetics is geographically isolated in comparison with whole world spread breeds.

5.1.1.12 Limousin

| Marker | Major Allele Freq. | No. of Genotypes | No. of Obs. | No. of Alleles | Avail-ability | H _e | H _o | PIC | F _{is} |
|---------|--------------------|------------------|-------------|----------------|---------------|----------------|----------------|-------|-----------------|
| BM1824 | 0.402 | 10 | 188 | 4 | 1.000 | 0.689 | 0.707 | 0.631 | -0.024 |
| BM2113 | 0.234 | 29 | 186 | 10 | 0.989 | 0.830 | 0.866 | 0.807 | -0.041 |
| ETH3 | 0.473 | 14 | 188 | 6 | 1.000 | 0.676 | 0.707 | 0.625 | -0.045 |
| ETH10 | 0.368 | 21 | 178 | 7 | 0.947 | 0.751 | 0.753 | 0.713 | 0.000 |
| ETH225 | 0.395 | 14 | 186 | 5 | 0.989 | 0.689 | 0.613 | 0.634 | 0.113 |
| INRA023 | 0.344 | 28 | 179 | 10 | 0.952 | 0.769 | 0.721 | 0.735 | 0.066 |
| SPS115 | 0.503 | 11 | 165 | 5 | 0.878 | 0.647 | 0.697 | 0.593 | -0.075 |
| TGLA122 | 0.341 | 41 | 182 | 12 | 0.968 | 0.810 | 0.813 | 0.789 | -0.001 |
| TGLA126 | 0.551 | 15 | 186 | 6 | 0.989 | 0.639 | 0.651 | 0.602 | -0.016 |
| TGLA227 | 0.297 | 35 | 180 | 10 | 0.957 | 0.823 | 0.867 | 0.802 | -0.050 |
| Mean | 0.391 | 21.8 | 181.8 | 7.5 | 0.967 | 0.732 | 0.739 | 0.693 | -0.007 |

Table 5.12 Results of genetic variability for Limousin breed (n=188).

When we inspect results obtained for Limousin subpopulation (188 individuals), it can be seen that similar results in comparison with the others beef breed were obtained. Only higher value of F_{is} in *ETH225* should be mentioned (0.113). As the other values of inbreeding coefficient are rather ~0 or negative, it pointed to situation similar as in Galloway and its results obtained for loci *ETH10*.

In comparison, we have found higher number of alleles per loci (7.5) than (Maudet et al., 2002), as well as higher expected and observed heterozygosity (0.732, 0.739) as Maudet has found (0.675, 0.674).

When we compare results with (D'Andrea et al., 2011), following comparison is shown: expected heterozygosity among Limousine kept in Serbia was 0.67 (0.732), observed one was 0.65 (0.739), PIC=0.62 (0.693) and inbreeding coefficient was 0.03 (-0.007). As in all of other same breeds, Serbian cattle populations show lower genetic variability what can be caused by later start of imports, so results are more comparable to previous ones obtained in Czech in past (Czernekova et al., 2006).

5.1.1.13 Summary results of genetic variability for breeds

| Breed | Major Allele Freq. | No. of Genotypes | Sample Size | No. of obs. | No. of Alleles | Availability | H _e | H ₀ | PIC | F _{IS} |
|-------|--------------------|------------------|-------------|-------------|----------------|--------------|----------------|----------------|-------|-----------------|
| C100 | 0.376 | 27 | 363 | 358.800 | 9.300 | 0.988 | 0.745 | 0.741 | 0.709 | 0.008 |
| G100 | 0.406 | 21 | 700 | 684.100 | 7.400 | 0.977 | 0.704 | 0.711 | 0.661 | -0.009 |
| H100 | 0.389 | 24.7 | 243 | 236.500 | 8.500 | 0.973 | 0.735 | 0.728 | 0.698 | 0.011 |
| P100 | 0.361 | 23.6 | 125 | 122.500 | 8.100 | 0.980 | 0.761 | 0.772 | 0.728 | -0.010 |
| Q100 | 0.395 | 16.1 | 73 | 70.600 | 6.800 | 0.967 | 0.738 | 0.750 | 0.702 | -0.008 |
| SM100 | 0.465 | 27 | 730 | 709.500 | 9.000 | 0.972 | 0.671 | 0.659 | 0.630 | 0.018 |
| T100 | 0.476 | 28.8 | 705 | 692.100 | 8.500 | 0.982 | 0.655 | 0.660 | 0.620 | -0.007 |
| U100 | 0.500 | 14.1 | 137 | 134.100 | 6.700 | 0.979 | 0.637 | 0.646 | 0.589 | -0.011 |
| W100 | 0.470 | 11.8 | 66 | 63.400 | 5.500 | 0.961 | 0.654 | 0.631 | 0.604 | 0.043 |
| Y100 | 0.391 | 21.8 | 188 | 181.800 | 7.500 | 0.967 | 0.732 | 0.739 | 0.693 | -0.007 |
| X | 0.366 | 32 | 380 | 373.500 | 10.100 | 0.983 | 0.754 | 0.727 | 0.719 | 0.036 |

Table 5.13 Summary results of genetic variability for 10 selected breeds and crossbred dataset (n=3680). For breeds abbreviations see table 4.1.

When we see results of genetic variability calculated across all loci examined for whole breeds and crossbred dataset (table 5.13), in the meaning of major alleles frequencies, most uniform breed is Hereford (0.500), the lowest value was detected for Piedmontese breed (0.361) and crossbred dataset (0.366). The highest average number of genotypes across all loci was detected at all in Charolais (28.8), the lowest value in Galloway (11.8). In average, 9.3 different alleles were detected in all of loci in Czech Fleckvieh, only 5.5 different alleles in Galloway. As Czech Fleckvieh is relatively young breed established after WW II, with influence of couple of breeds and as a dual purpose one, these results are in accordance with breeding strategy and selection processes in the Czech Fleckvieh. In Galloway, as we mentioned above, only 66 individuals were analyzed and only very limited imports happened in past what is probably reason of results mentioned. Very similar results in mentioned parameters as in Czech Fleckvieh were obtained for Czech Simmental as well.

As well, there is nothing surprising on fact, that crossbred dataset resulted in 10 different alleles with 32 different genotypes which are highest values in the whole dataset ordered by breeds.

Regarding results of expected and observed heterozygotes, it can be noted that largely kept breeds in Czech or world-wide (like Czech Fleckvieh, Holstein, Limousin, Blonde d'Aquitaine), genetically and evolutionary different kept in Czech (Piedmontese) and crossbreds are more heterozygous in comparison with minor kept ones however genetically different (Galloway) and more uniform beef populations like Hereford, Czech Simmental, Charolais and Abredeen Angus are. These results can be near generalized according to breeding strategies for beef and dairy or dual purpose breeds with exceptions mentioned.

In results of all methods for computing genetic distances, Hereford breed was observed as the most different one, what is result highly comparable to previous studies done (Blott et al., 1998b).

5.1.2 Genetic distances

5.1.2.1 Euclidean and Nei 1972 genetic distance

| | C100 | G100 | H100 | P100 | Q100 | SM100 | T100 | U100 | W100 | X | Y100 |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| C100 | | 0.281 | 0.276 | 0.223 | 0.239 | 0.167 | 0.264 | 0.423 | 0.373 | 0.090 | 0.244 |
| G100 | 0.173 | | 0.334 | 0.299 | 0.347 | 0.339 | 0.325 | 0.438 | 0.337 | 0.241 | 0.334 |
| H100 | 0.185 | 0.254 | | 0.249 | 0.306 | 0.318 | 0.345 | 0.421 | 0.429 | 0.239 | 0.353 |
| P100 | 0.115 | 0.184 | 0.141 | | 0.256 | 0.267 | 0.316 | 0.412 | 0.370 | 0.177 | 0.257 |
| Q100 | 0.140 | 0.277 | 0.213 | 0.152 | | 0.289 | 0.353 | 0.422 | 0.446 | 0.228 | 0.277 |
| SM100 | 0.047 | 0.228 | 0.232 | 0.150 | 0.165 | | 0.251 | 0.459 | 0.403 | 0.168 | 0.293 |
| T100 | 0.129 | 0.185 | 0.292 | 0.224 | 0.241 | 0.109 | | 0.472 | 0.380 | 0.234 | 0.309 |
| U100 | 0.342 | 0.426 | 0.360 | 0.331 | 0.341 | 0.382 | 0.458 | | 0.503 | 0.402 | 0.449 |
| W100 | 0.304 | 0.212 | 0.390 | 0.292 | 0.421 | 0.332 | 0.275 | 0.515 | | 0.339 | 0.432 |
| X | 0.018 | 0.122 | 0.139 | 0.075 | 0.128 | 0.048 | 0.106 | 0.306 | 0.233 | | 0.225 |
| Y100 | 0.131 | 0.231 | 0.283 | 0.151 | 0.180 | 0.187 | 0.243 | 0.431 | 0.387 | 0.114 | |

Table 5.14 Euclidean (above) and Nei 1972 (bellow diagonal) genetic distance.

Geometric distance results represented by Euclidean distance calculated across the whole general and crossbred datasets show, that in the meaning of simple geometric distance, seems to be the most divergent from each other Hereford breed with average distance of each other equals 0.4401. The highest result was reached for Hereford x Galloway distance (0.503), the lowest one for Czech Fleckvieh x Czech Simmental (0.167) and Czech Fleckvieh x crossbreds which is in accordance with state when the most crossbreds are produced within Czech Fleckvieh cows and Czech Fleckvieh is largely influenced by Simmental as well.

Nei 1972 genetic distance reported similar results, as one which is under the assumption of biological model which assume drift and mutations as power of genetic evolution (Nei, 1972). Anyway, under this assumptions allele frequencies which are considered to remain the same in time as mutation is assumed as a random process. The Hereford seems to be the most distanced breed with average value of Nei 1972 distance equals 0.3892. As well as in Euclidean distance, pairs represented by Czech Fleckvieh x crossbreds, Czech Fleckvieh, Czech Simmental seem to be the most closest under assumptions of Nei 1972 distance.

In comparison with study done in (Russell et al., 2000), who reported Nei 1972 distance of Charolais and Simmental qualed to 0.405, we obtained just 0.109 between Charolais and Czech Simmental, 0.129 between Charolais and Czech Fleckvieh. Regarding Aberdeen Angus, we have found distance between this breed and Czech Simmental equals to 0.228 and between Aberdeen Angus and Czech Fleckvieh equals to 0.173 in contrary of 1.569 in (Russell et al., 2000). This can be caused by specificity of Czech Simmental as well as of Czech Fleckvieh which can significantly differ from breed kept like pure Simmental breeds abroad. As Russell et al. (2000), we have found highest value of distance between Aberdeen Angus and Hereford.

5.1.2.2 Goldstein and Shriver genetic distance

| | C100 | G100 | H100 | P100 | Q100 | SM100 | T100 | U100 | W100 | X | Y100 |
|-------|-------|-------|-------|-------|-------|-------|-------|--------|-------|-------|-------|
| C100 | | 1.324 | 3.175 | 1.185 | 3.708 | 0.979 | 1.377 | 7.316 | 3.566 | 0.102 | 2.313 |
| G100 | 0.231 | | 3.603 | 1.986 | 2.953 | 1.962 | 1.850 | 8.387 | 1.976 | 1.102 | 1.363 |
| H100 | 0.266 | 0.397 | | 4.039 | 7.491 | 4.769 | 6.607 | 11.812 | 4.392 | 3.375 | 6.572 |
| P100 | 0.164 | 0.269 | 0.299 | | 2.078 | 1.837 | 2.526 | 7.177 | 4.868 | 0.812 | 1.489 |
| Q100 | 0.326 | 0.407 | 0.470 | 0.194 | | 3.004 | 2.680 | 4.617 | 7.056 | 3.041 | 2.953 |
| SM100 | 0.094 | 0.314 | 0.390 | 0.225 | 0.341 | | 1.088 | 5.719 | 3.767 | 0.738 | 2.942 |
| T100 | 0.171 | 0.251 | 0.477 | 0.293 | 0.377 | 0.184 | | 6.609 | 5.186 | 1.305 | 3.000 |
| U100 | 0.763 | 0.917 | 0.881 | 0.674 | 0.658 | 0.782 | 0.825 | | 9.449 | 6.880 | 8.309 |
| W100 | 0.421 | 0.307 | 0.534 | 0.500 | 0.725 | 0.436 | 0.480 | 0.896 | | 3.317 | 3.503 |
| X | 0.019 | 0.183 | 0.246 | 0.107 | 0.269 | 0.080 | 0.149 | 0.689 | 0.360 | | 1.767 |
| Y100 | 0.262 | 0.259 | 0.550 | 0.201 | 0.352 | 0.330 | 0.360 | 0.845 | 0.480 | 0.206 | |

Table 5.15 Goldstein (above) and Shriver (bellow diagonal) genetic distance.

Goldstein's genetic distance is assuming and calculating genetic distances under the SMM model and it represents one which is designed especially for microsatellite markers, however SMM can cause errors especially in small datasets when genetic drift is assumed as mutation (Goldstein et al., 1995a). Anyway, from most distanced from each other under assumptions mentioned seems to be Hereford (7.6275) followed by Holstein breed with average distance 5.58. The longest distance was obtained between Hereford and Holstein breeds (11.812), the lowest ones between crossbreds and Czech Fleckvieh, crossbreds and Simmental.

Shriver's genetic distance is another one created especially for high polymorphic loci which extends Nei minimum genetic distance and is based on frequency-weighted means of the absolute value of the difference in number of repeats over pairs of alleles, both within and between populations (Shriver et al., 1995). Under these assumptions, also Hereford population (with average distance from each other breed equals 0.793) is on of the most divergent. Also, relationship between Hereford and Aberdeen Angus has the highest distance - 0.917. In opposite, Czech Fleckvieh and crossbred dataset seems to be most closest with distance equals to 0.019.

5 Results and Discussion

5.1.2.3 Slatkin and SharedAllele genetic distance

| | C100 | G100 | H100 | P100 | Q100 | SM100 | T100 | U100 | W100 | X | Y100 |
|-------|-------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| C100 | | 46.480 | 59.915 | 55.065 | 53.958 | 46.027 | 47.016 | 49.053 | 49.254 | 51.784 | 51.484 |
| G100 | 0.301 | | 53.758 | 49.281 | 46.619 | 40.426 | 40.904 | 43.538 | 41.079 | 46.198 | 43.949 |
| H100 | 0.302 | 0.364 | | 62.919 | 62.740 | 54.817 | 57.246 | 58.548 | 55.079 | 60.056 | 60.743 |
| P100 | 0.241 | 0.334 | 0.295 | | 54.467 | 49.024 | 50.304 | 51.053 | 52.695 | 54.633 | 52.799 |
| Q100 | 0.271 | 0.374 | 0.342 | 0.297 | | 46.562 | 46.828 | 44.863 | 51.253 | 53.232 | 50.634 |
| SM100 | 0.173 | 0.351 | 0.307 | 0.274 | 0.300 | | 40.034 | 40.763 | 42.762 | 45.727 | 45.421 |
| T100 | 0.286 | 0.342 | 0.363 | 0.330 | 0.360 | 0.273 | | 42.244 | 44.771 | 46.884 | 46.069 |
| U100 | 0.423 | 0.407 | 0.413 | 0.401 | 0.406 | 0.449 | 0.435 | | 45.133 | 48.558 | 47.476 |
| W100 | 0.385 | 0.334 | 0.448 | 0.390 | 0.462 | 0.398 | 0.384 | 0.473 | | 48.945 | 46.622 |
| X | 0.100 | 0.262 | 0.259 | 0.200 | 0.259 | 0.167 | 0.251 | 0.394 | 0.353 | | 50.878 |
| Y100 | 0.267 | 0.349 | 0.377 | 0.280 | 0.297 | 0.300 | 0.301 | 0.438 | 0.440 | 0.242 | |

Table 5.16 Slatkin (above) and Shared Allele (bellow diagonal) genetic distance.

Slatkin's genetic distance (Slatkin, 1995) is an analogue of Wright's F_{ST} adapted to microsatellite loci by assuming a high-rate stepwise mutation model instead of a low-rate K- or infinite-allele mutation model. Under these conditions, Holstein with average distance 58.58 is the most distant of all other breeds. Smallest distance was observed between Czech Simmental and Charolais breeds (40.034) and between Czech Simmental and Aberdeen Angus (40.426).

As a SharedAllele genetic distance is one of the simplest estimations of genetic distance, it is based on the proportion of shared alleles and can be also used to evaluate microsatellite data, especially with usage of small datasets within well defined populations (Chakraborty and Jin, 1993). Highest values of SharedAllele distance can be observed for Hereford as well as for other distances evaluated. As well, highest distance was observed between Hereford and Galloway breeds (0.473). The lowest value was observed between Czech Fleckvieh and crossbred dataset.

5.1.3 Phylogenetic trees

Phylogenetic trees based on 6 genetic distances calculated on general and crossbred joint datasets are used to show compressed results given by these distances. Dendrograms constructed by UPGMA (Unweighted Pair Group Method with Arithmetic mean) and Neighbour Joining methods of clustering are used to show results for each genetic distance calculated. As trees are unrooted in time, there are displaying just relationship in the meaning of each genetic distance and clustering algorithm instead of evolutionary relationships themselves. Mainly, results are discussed and are in accordance at all with e.g. (Negrini et al., 2007) and (Feliuss et al., 2011).

5.1.3.1 Euclidean NJ and UPGMA trees

G:\Dropbox\disttree\EuclidianNJ.pvf
Number Of Species = 11

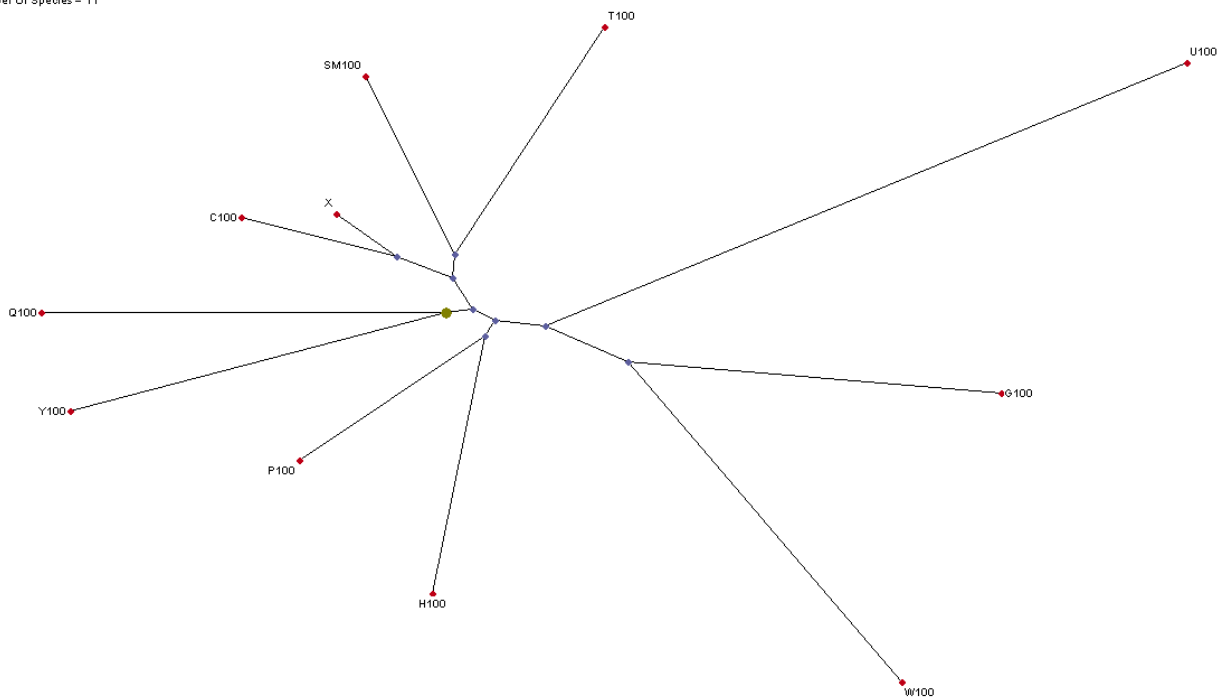


Figure 5.1 Euclidean NJ phylogenetic tree.

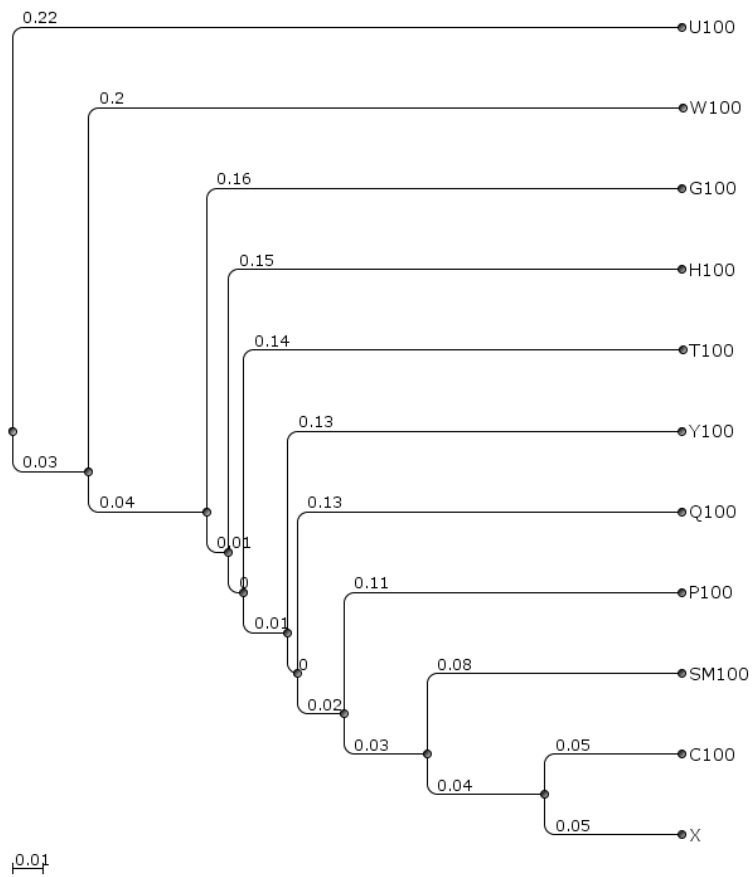


Figure 5.2 Euclidean UPGMA phylogenetic tree.

5 Results and Discussion

On the figures 5.1 and 5.2 NJ and UPGMA dendrogram of Euclidean distance can be seen. As Euclidean distance is geometric based measure without any assumptions based on biological basis, it can not be used to discuss similarities between breeds on biological basis as well. Anyway, as we used machine learning algorithms (like IB1 and IB5), they are usually data independent, so it can be useful to discuss geometric based distance in comparison with the others. On both of trees can be identified, that Euclidean distance reflects quite well real state of breeds - Czech Fleckvieh is clustered in the same branche as crossbreds, then Czech Simmental and Charolais are on the same subtree of NJ one. Aberdeen Angus and Galloway create another well defined subtree, however they are quite distant of each other. Blonde d'Aquitaine and Limousin, Piedmontese and Holstein then create two separate branches. The most distant of each other, not grouped with another breed is Hereford (both of clustering methods).

5.1.3.2 Nei 1972 NJ and UPGMA trees

G:\Dropbox\disttree\NeiNJ.pvf
Number Of Species = 11

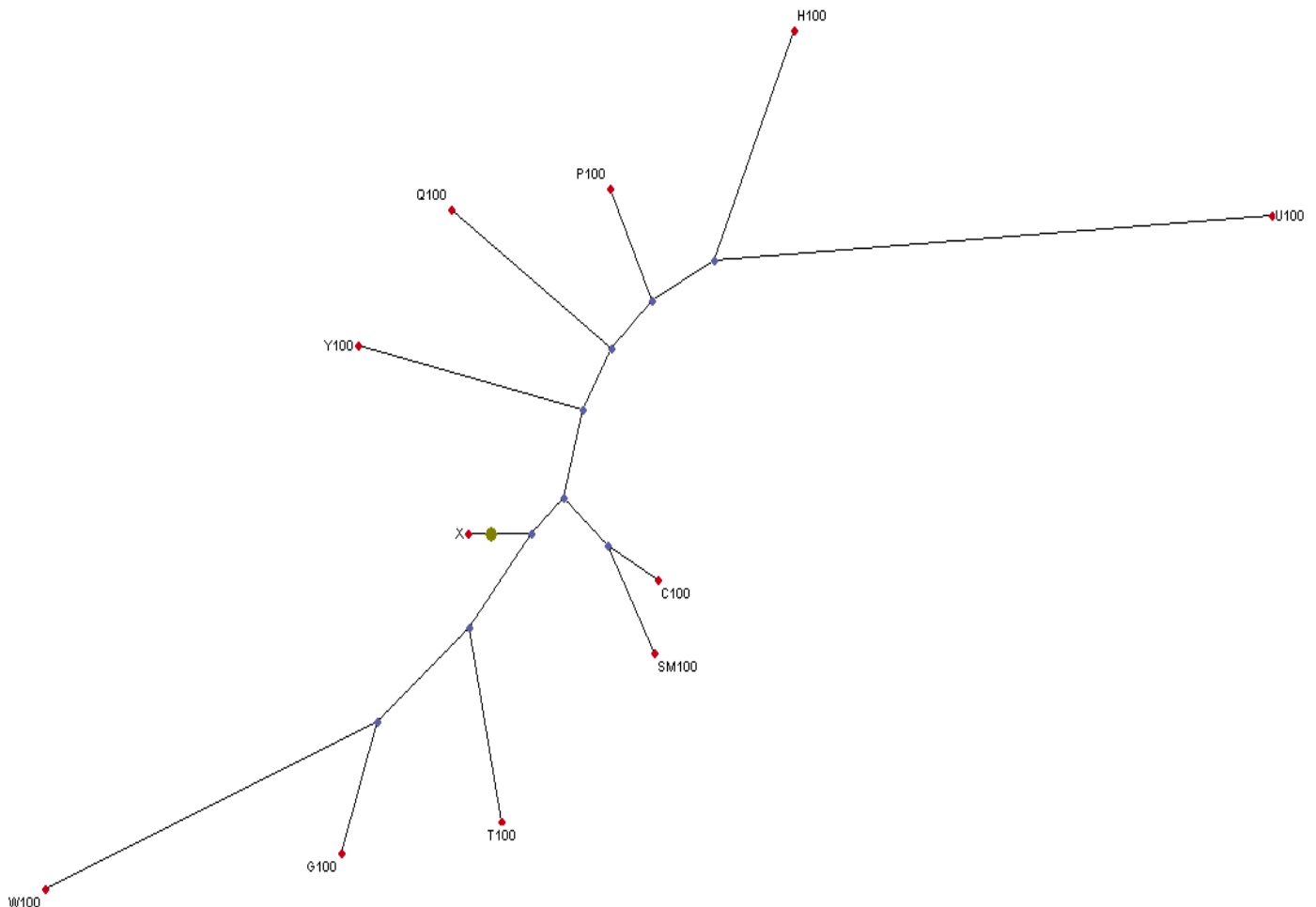


Figure 5.3 Nei 1972 NJ phylogenetic tree.

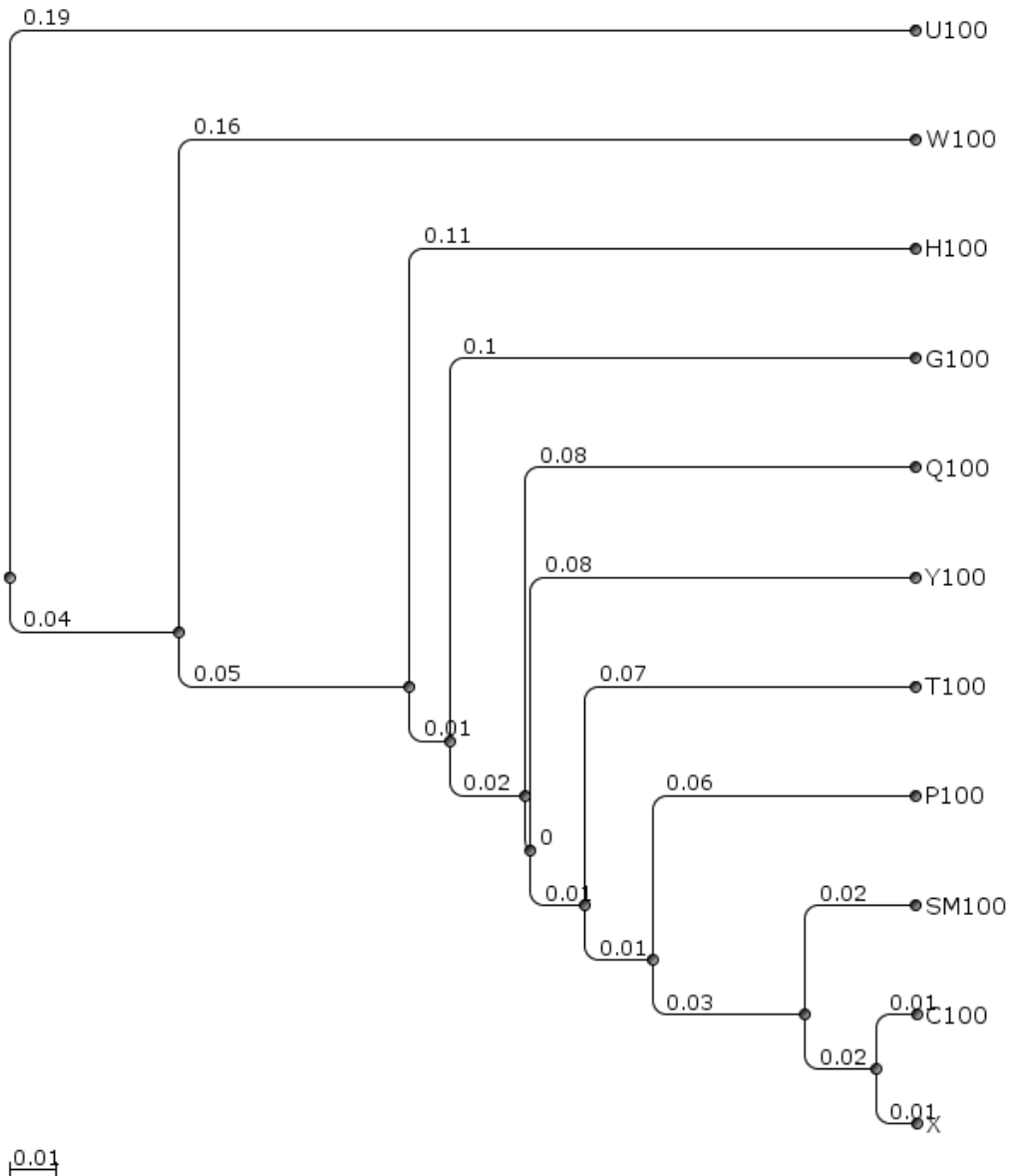


Figure 5.4 Nei 1972 UPGMA phylogenetic tree.

Dendrograms of Nei 1972 distance are displayed on figures 5.3 and 5.4. Most distant and separate is Hereford breed with closest connection to Holstein. Then Galloway represent the most distant breed from Hereford and Holstein. Limousin, Piemontese, Blonde d'Aquitaine with Holstein represent breeds closer connected to Hereford. On the other hand, Aberdeen Angus and Charolais represent are closer to Galloway branch on the other side of tree. Czech Simmental and Czech Fleckvieh represent one branch of tree between both of groups, closest connected to group of crossbreds. However, Nei's standard distance is one of basic one under biological assumptions, it reflects quite good real state of cattle breeds in Czech Republic.

5.1.3.3 Goldstein NJ and UPGMA trees

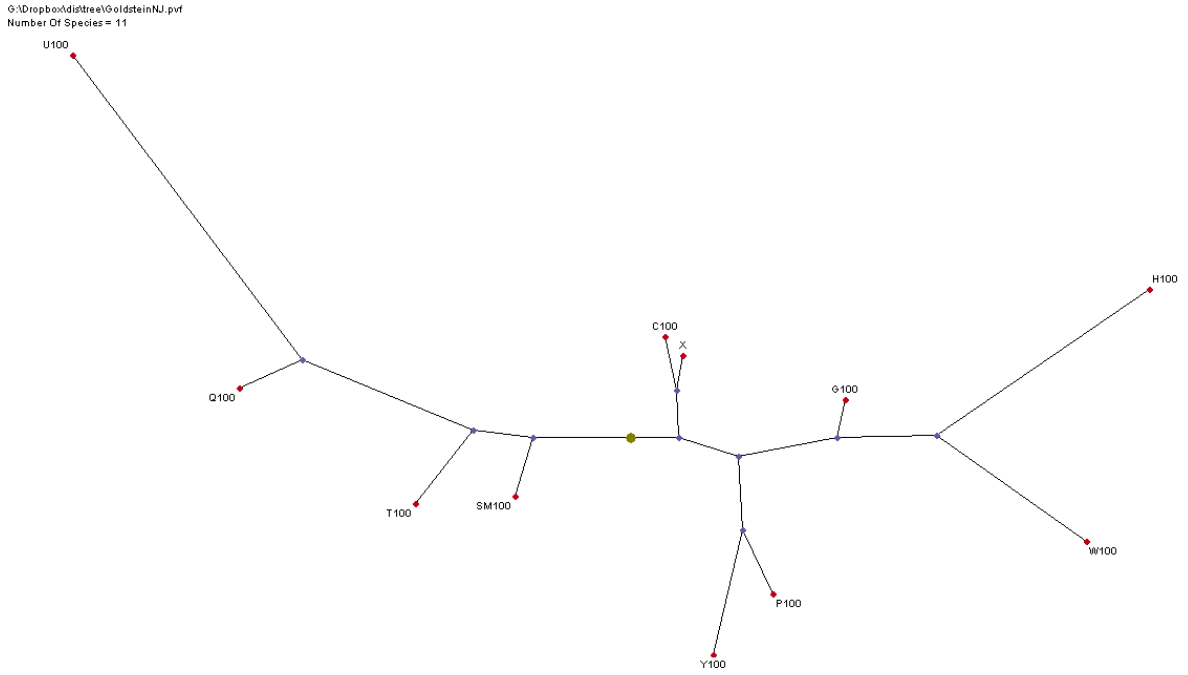


Figure 5.5 Goldstein NJ phylogenetic tree.

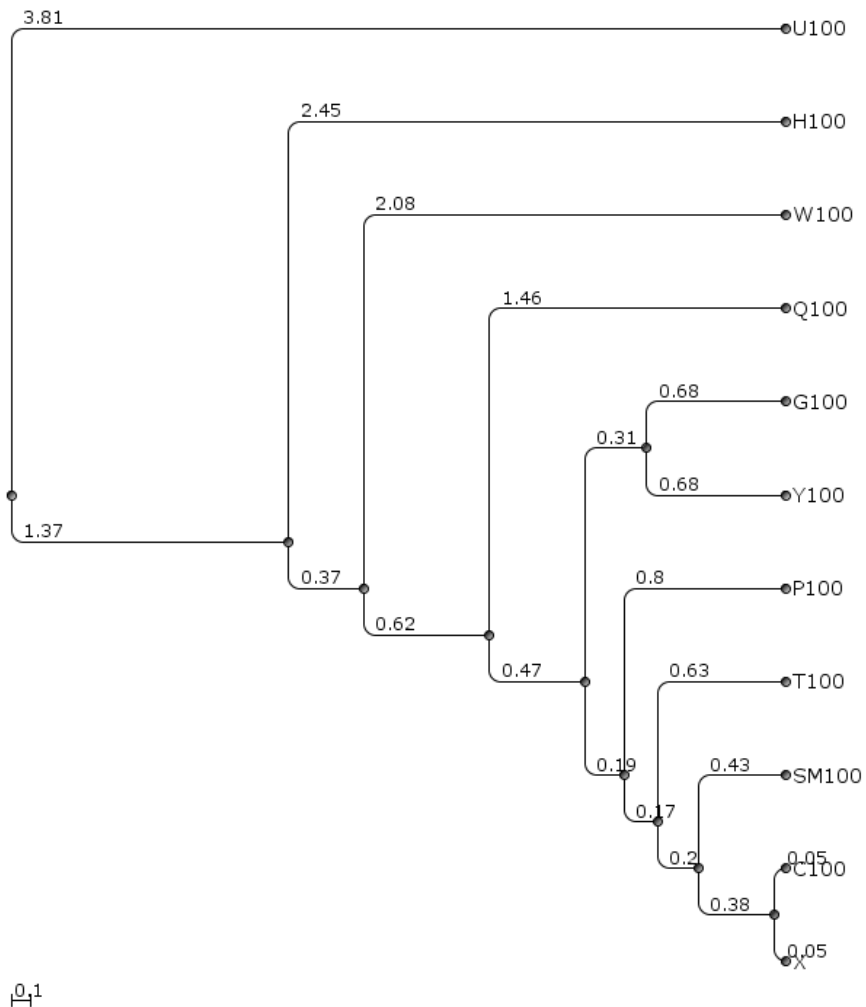


Figure 5.6 Goldstein UPGMA phylogenetic tree.

5 Results and Discussion

Under assumptions of SMM which is used for calculation of Goldstein's genetic distance, dendrograms on figures 5.5 and 5.6 were obtained. Most than real relationships of breeds it can be analyzed from point of view of similar breeding strategies, however it is highly influenced by small population sizes of some breeds. Also, if we known real state of cattle breeding, it is clear that similarity between Holstein and Galloway, Piedmontese and Limousin can not be explained anyway by breeding of populations, mating in them etc. However, it is interesting how populations resulted to similarities under complete different genetic pressure on them in point of view of SMM. Czech Fleckvieh and crossbreeds are clustered together in spite of facts discussed.

5.1.3.4 Shriver NJ and UPGMA trees

G:\Dropbox\dia\tree\ShriverNJ
Number Of Species = 11

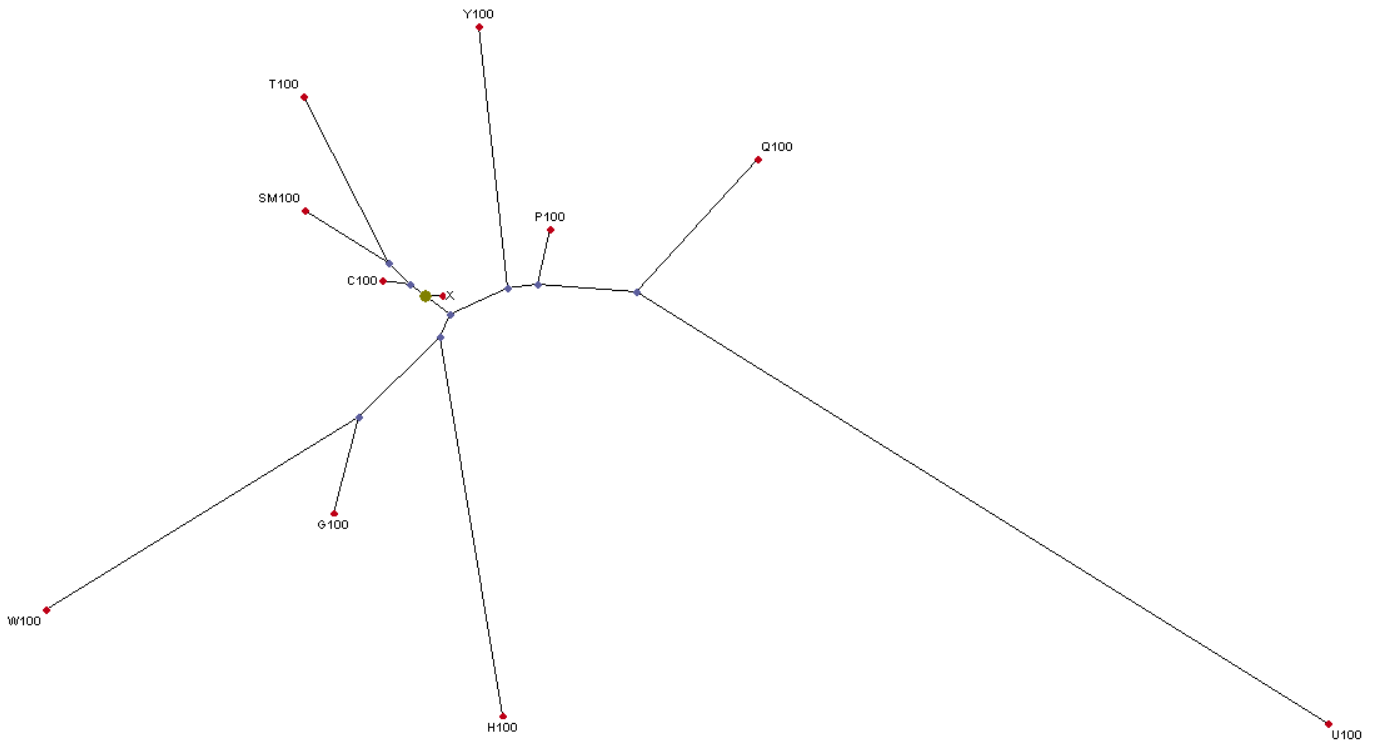


Figure 5.7 Shriver NJ phylogenetic tree.

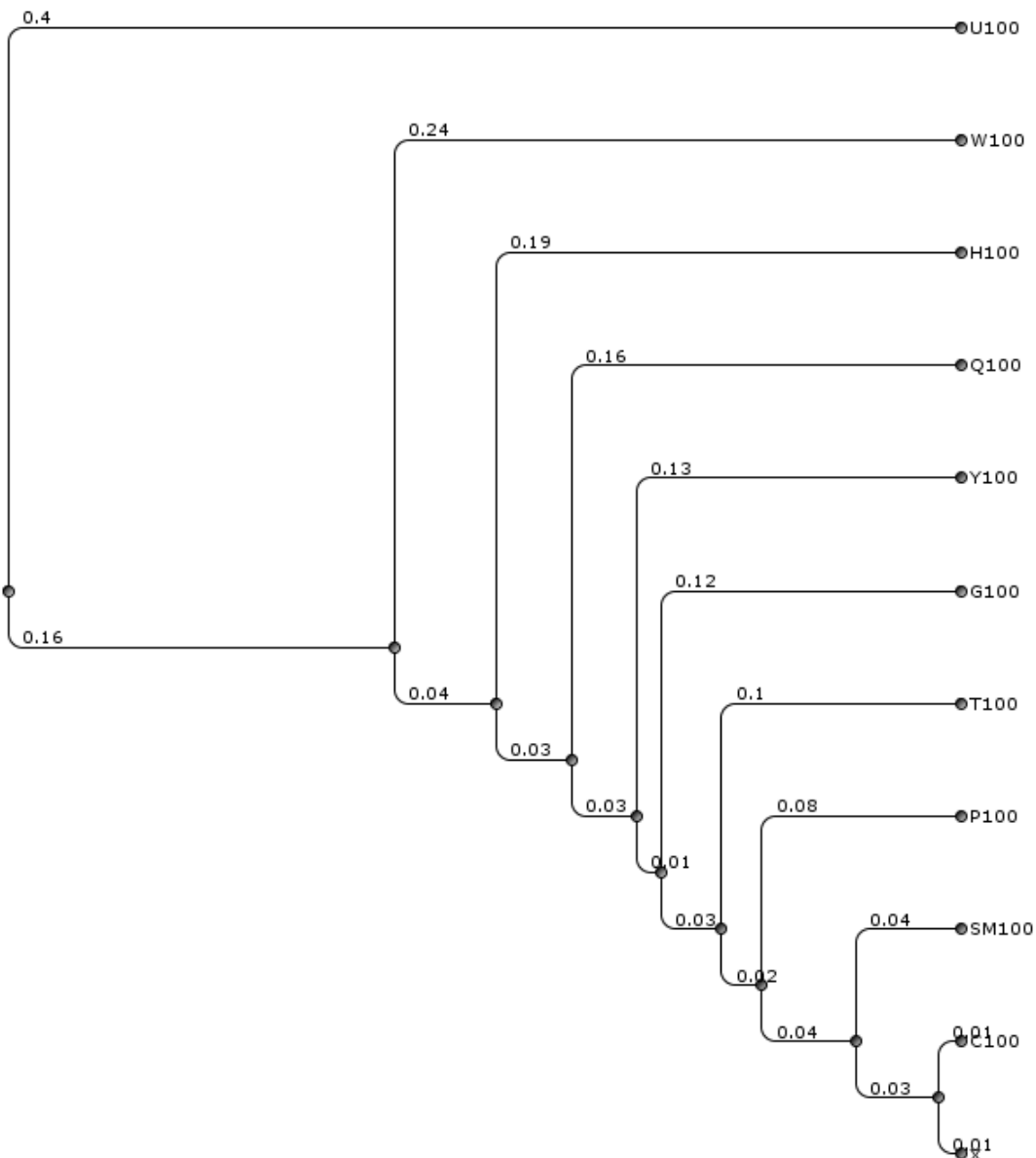


Figure 5.8 Shriver UPGMA phylogenetic tree.

Results obtained for Shriver distance which measures similarities between breeds under the same assumption of Nei's distance with regards to multiallelic loci show Hereford and Galloway the most distant as the other methods. Shriver distance can express well how Simmental like breeds are structured (Czech Fleckvieh, Czech Simmental and crossbreds). Charolais is then clustered with this group together what can be explained by very similar results of genetic variability of this breed as is mentioned in previous chapters. Then, Holstein, Limousin and Blonde d'Aquitaine plus Piedmontese and Aberdeen Angus represent very separated branches in the meaning of genetic drift and similarities of microsatellite loci.

5.1.3.5 Slatkin NJ and UPGMA trees

G:\Dropbox\distree\SlatkinNJ.prf
Number Of Species = 11

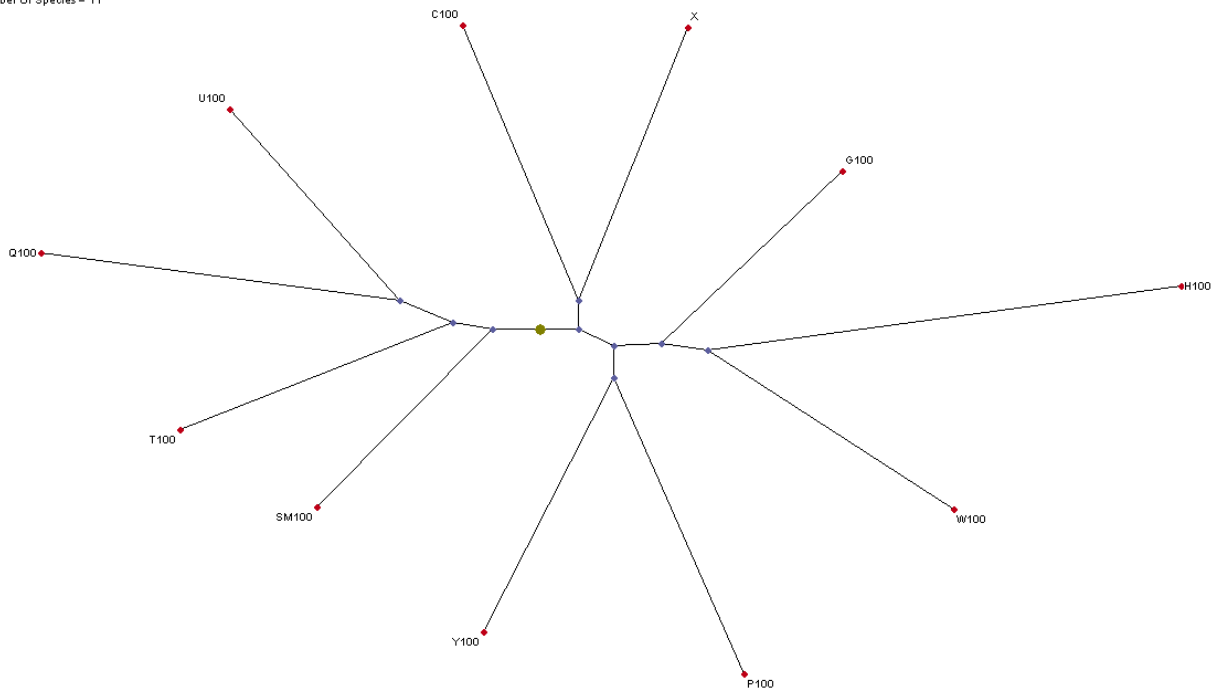


Figure 5.9 Slatkin NJ phylogenetic tree.

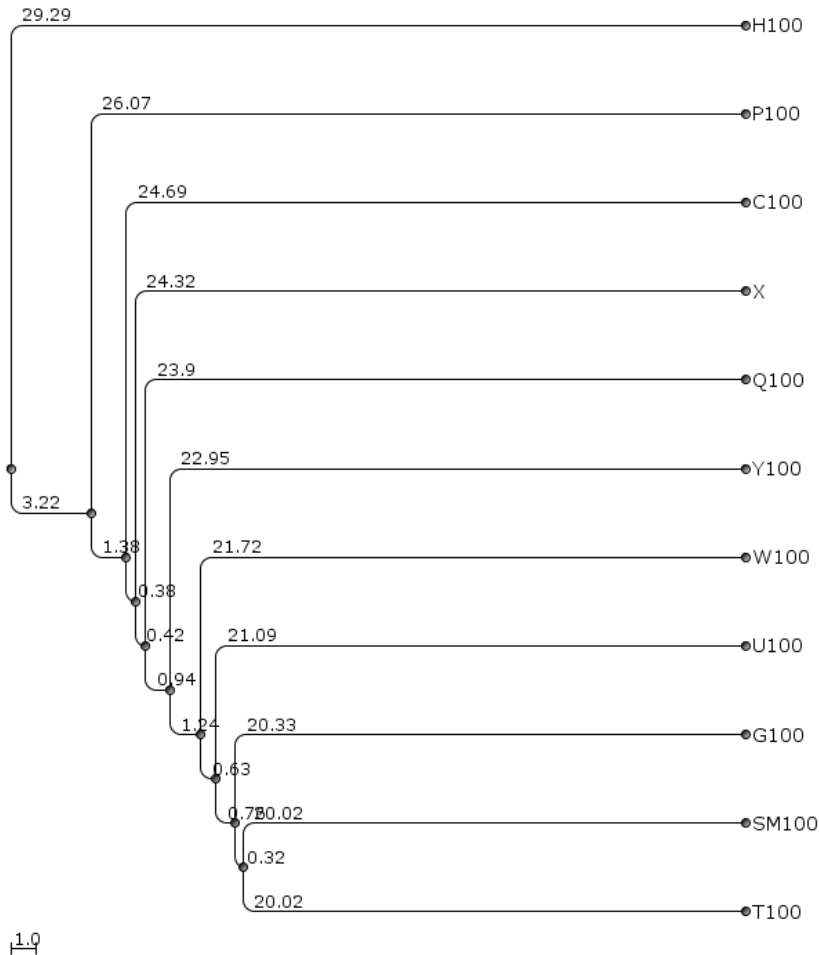


Figure 5.10 Slatkin UPGMA phylogenetic tree.

5 Results and Discussion

As analogous method in comparison with Wright F_{ST} Slatkin's distance represent method how to separate populations the most effectively based on microsatellite multiallelic loci. The results can be seen on figure 5.9 easily. All of breeds are well separated except Czech Fleckvieh which create branch with crossbreeds and Piedmontese and Limousine breeds. Best separated by this methods are Holstein and Blonde d'Aquitaine breeds what refers to very different genetic (based on microsatellites selected) basis of both. Similar results were obtained by (Yves Amigues, Simon Boitard) and showed also that Blonde d'Aquitaine and well-defined cattle populations (Salers) are genetically more similar to each other than to the Limousin.

5.1.3.6 Shared Allele NJ and UPGMA trees

G:\Dropbox\distree\SharedAllNJ.pvf
Number Of Species = 11

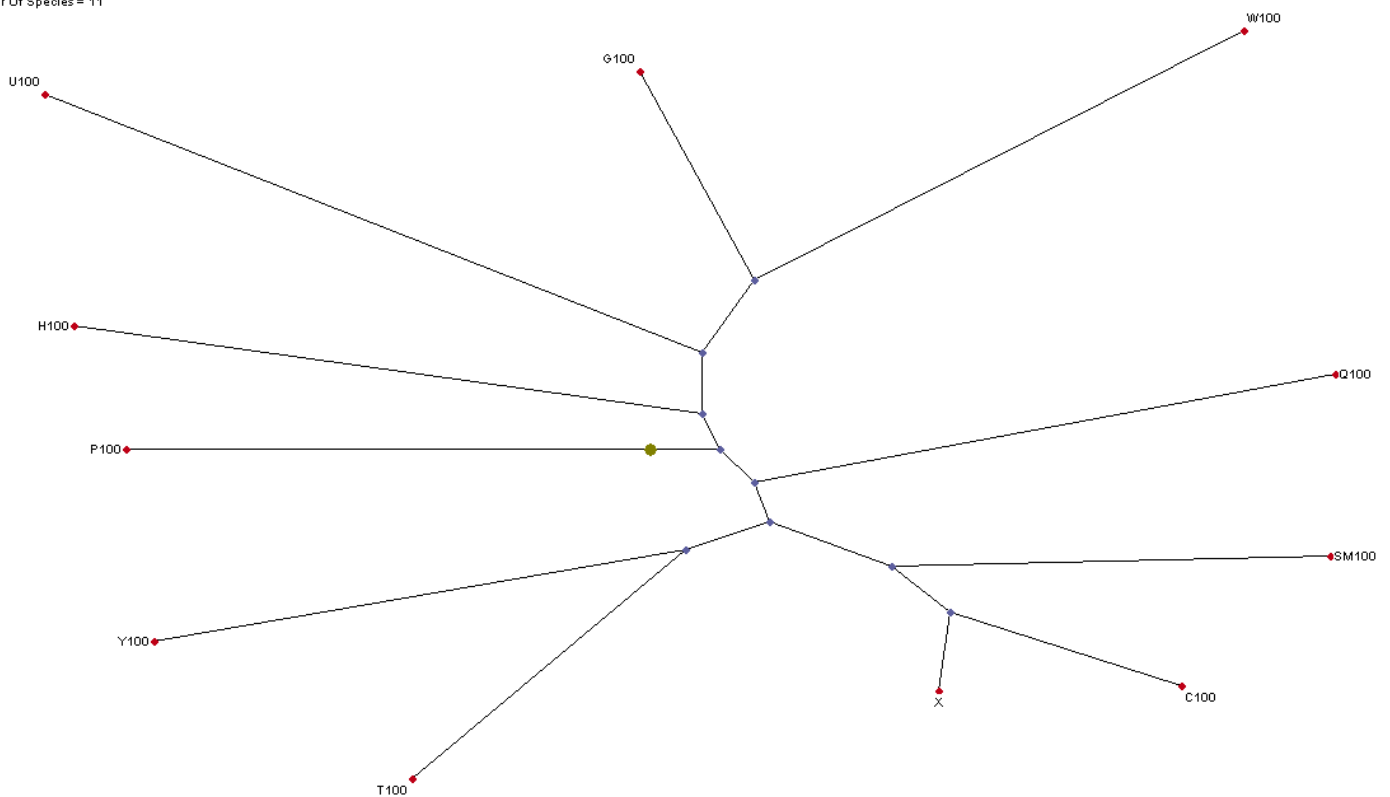


Figure 5.11 Shared Allele phylogenetic tree.

5.2 Estimation and validation paternity testing by microsatellite loci in selected cattle breeds

| | CEP 1 | CEP 2 | CEP 3 | PIC |
|-----------|----------|----------|----------|-------|
| C100 | 0.998123 | 0.988668 | 0.999998 | 0.709 |
| H100 | 0.997526 | 0.987431 | 0.999997 | 0.698 |
| T100 | 0.986215 | 0.976321 | 0.999986 | 0.620 |
| SM100 | 0.980362 | 0.964416 | 0.999969 | 0.630 |
| P100 | 0.999183 | 0.993402 | 0.999999 | 0.728 |
| Y100 | 0.997262 | 0.986747 | 0.999991 | 0.693 |
| U100 | 0.953371 | 0.948479 | 0.999907 | 0.589 |
| Q100 | 0.997591 | 0.987107 | 0.999997 | 0.702 |
| W100 | 0.972259 | 0.961617 | 0.999948 | 0.604 |
| G100 | 0.991806 | 0.974167 | 0.999982 | 0.661 |
| Whole set | 0.998385 | 0.990697 | 0.99999 | 0.663 |

Table 5.17 Results of combined probabilities of paternity exclusion and polymorphic information content by breeds calculated for general dataset.

By using software environment created to handle large sets of genotype data, results summarized in table 5.17 were obtained for general dataset. All three scenarios of combined exclusion of paternity were calculated across all of 10 MS loci. Best probability for exclusion of one parent (CEP 1), when the genotypes of both parents and ancestor are known, was reached for Piedmontese breed (0.999183). For the whole general dataset, CEP 1 was calculated as 0.998385. The worst value was obtained for Hereford breed (0.953371).

In case of CEP 2 scenario, when one of parents genotype is unknown, whole set probability was calculated as 0.990697, best one was reached for Piedmontese breed (0.993402), the worst one for Hereford (0.948479).

CEP 3 results, when all of three genotypes are known (parents and offspring), but we want to know what is the probability of exclusion of both parents, show that for whole dataset CEP 3=0.999990, the best was reached for Piedmontese (0.999999), the worst one for Hereford breed (0.999907). At all, results reached are fully comparable to (Putnova et al., 2011; Radko, 2010). Polymorphic information content ranges from 0.589 in Hereford to 0.728 in Piedmontese, the value calculated for the whole general dataset (n=3300) equals 0.663 what is less than (D'Andrea et al., 2011) observed.

Results of combined probabilities and polymorphic information content reflect real situation and breed strategies in all of observed breeds. Not well defined breeds or breeds with "weak" acceptance of breeding animals pedigree show higher values in all of parameters as result of higher genetic variability within breeds. Anyway, results proved that panel of microsatellite loci used for genotyping of general dataset fullfills recommendations on paternity exclusion as well as studies of genetic diversity of selected cattle breeds (International Society for Animal Genetics; FAO - Measuremens of Domestic Animal Diversity). Nowadays, as normaly done by ISAG, recommended panel of microsatellite loci was extended to 17 loci (*SPS113, BM1818, RM067, ILST006, MGTG4B, CSSM66, CSRM60* were added for routine testing) thanks to reduced genetic variability in whole world spread, well controlled cattle populations like Holstein is.

5.3 Creation of the software support for routine genotyping of microsatellite loci under the reference laboratory conditions

Following chapters describe created software solution for storing, operating and reporting issues with microsatellite data under accredited laboratory conditions. Text is aimed to explain detailed solution, processes and software structure in many aspects.

5.3.1 Network model

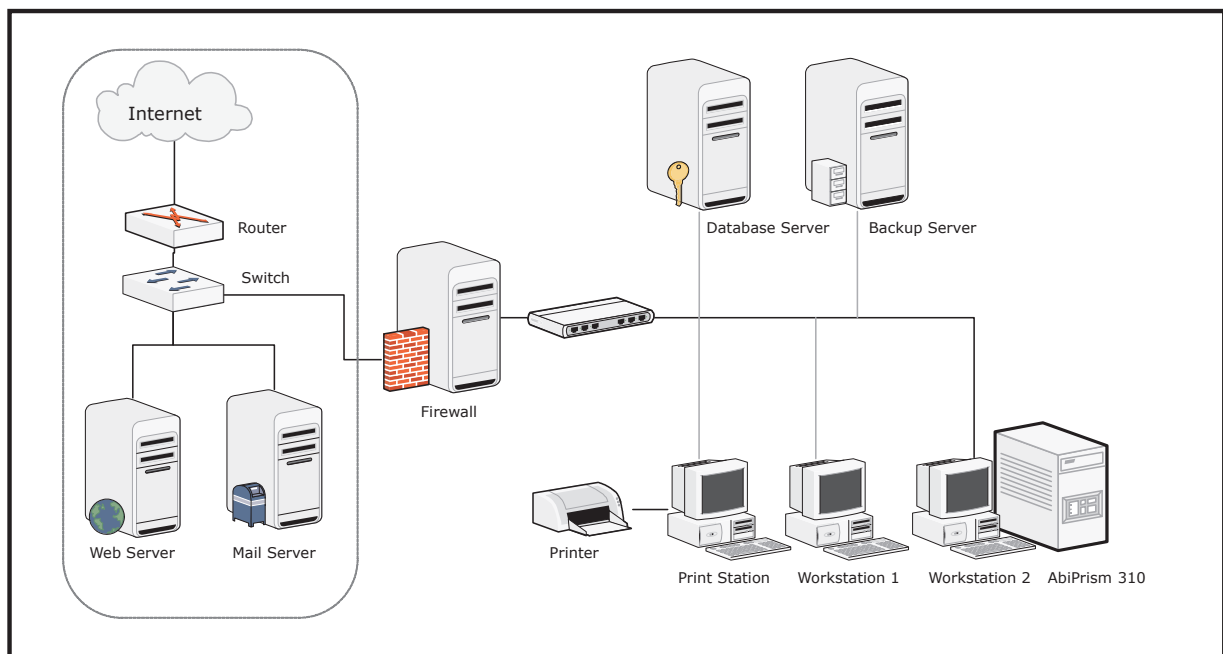


Figure 5.13 Schema of network model of the system.

Recommended network connection for which applications is designed is shown on figure 5.13. There is a public section of network which can be used especially for reporting to laboratory customers as well as for automatization of data input in future. In case of security, there is no public IP address allowed in private network. So, there is no chance for computers from Internet area to access any of computers included in private network.

The whole system can operate on 1..n computers according to laboratory needs. There must be at least one computer which can represent Workstations, Print Station and provides connection to AbiPrism 310 genetic analyser. In this case computer is used as database and backup server as well. Normaly, we can assume Workstation 1 is used for manual data input, Workstation 2 works with data from genetic analyser, Print Station is used for protocols and database issues and another computer in network is dedicated as Backup Server, where database is saved daily.

5.3.2 Application model

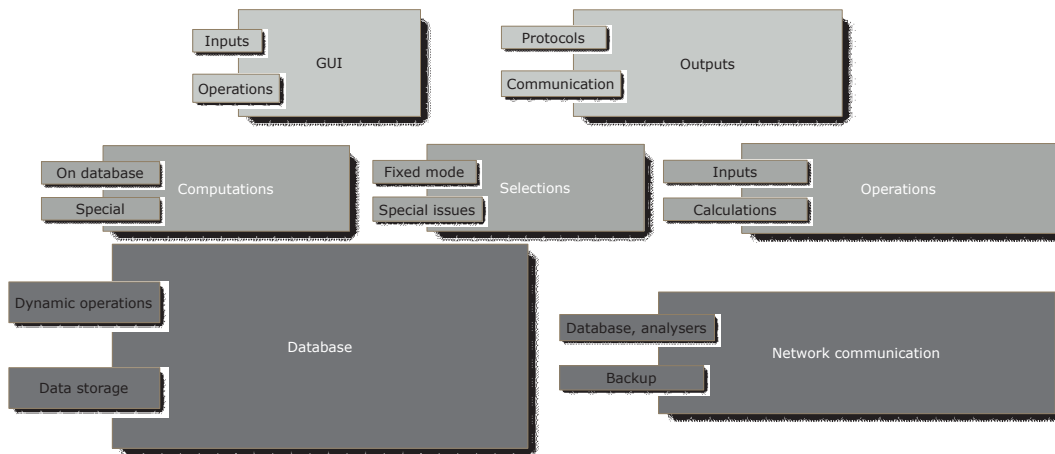


Figure 5.14 Application design.

The software application is designed in three layers. Top layer operates with data outputs and inputs thorough graphical user interface, second layer performs user inputs, communicates with database and performs software calculations and manipulations over data. Lowest layer operates directly with database interface thorough SQL and with network - backups and security issues. Top layer is half implemented in Borland Delphi (Borland Inc., 2007), half with usage of reporting system - Fast Reports (Fast Reports Inc., 2007).

Middle layer of the system is implemented fully In Borland Delphi environment and coded in Object Pascal. Database server used is represent by Firebird SQL Server XXX. Network layer is implemented in Object Pascal and in assembler code.

5.3.3 Key processes

For identification, description and definition of key processes, UML case diagram syntax was used as it offers simple, readable and repeatable tool for complex software design based on processes in system. Use cases syntax operates with only three basic principles:

- use case - ellipse, defined as process on the chosen level of abstraction which creates closed group of activities and can be more specify on more detail level of abstraction. On the mottom levels processes can be easily converted to code classes in case of object programming. On the oppostie, on the top level of abstraction, system is created by only one process and actors.
- actor - person, operator, who interacts with the system, resp. who interacts with the system thorough interaction processes. Role of actors is very useful because they allow everybody to imagine easily who and how interact with system.
- relation - oriented line, by which type of relation is defined.

As a system model is quite simple, lets skip top levels of abstraction and then, three basic processes with their particular subprocesses, actors and relation can be identified. The first one is represent by most simple case - order for genetic type identification whom use case diagram is displyed on figure 5.15. Order comes from farmer/breeder, who communicates

with lab manager directly by phone, on trade fairs, by email etc. When they agree on sample type and sample of tissue is sent by breeder to the lab, lab manager processes ID data as well as sample. He inputs ID data into the database, initiates farmers data (or selects them from databes when they are already present) and marks sample with unique ID lab number. This laboratory unique identifier comes with the sample thorough the whole process in lab and it is stored with results in database too. It identifies sample, pairing with breeders data, with the results of analysis, with animal ID. When lab ID is set, lab manager passes the sample to lab, where laboratory operator starts routine process doing DNA analysis on sample. When raw data (labeled with previously set unique lab ID) are ready as output of DNA analyser, then process of their joining in database is started automaticly. Lab manager who see actual state of database then contorls results of analysis and issues analysis protocol which is sent to farmer finally. When farmer re-orders analysis/protocol, he communicates with lab manager again and he just performs database search for given animal ID (which is already present in database) and re-issue protocol again. So, farmers can also run e.g. paternity testing based on protocols re-issued manually to perform self-control as they can obtain complete pedigree of their animals in the moment.

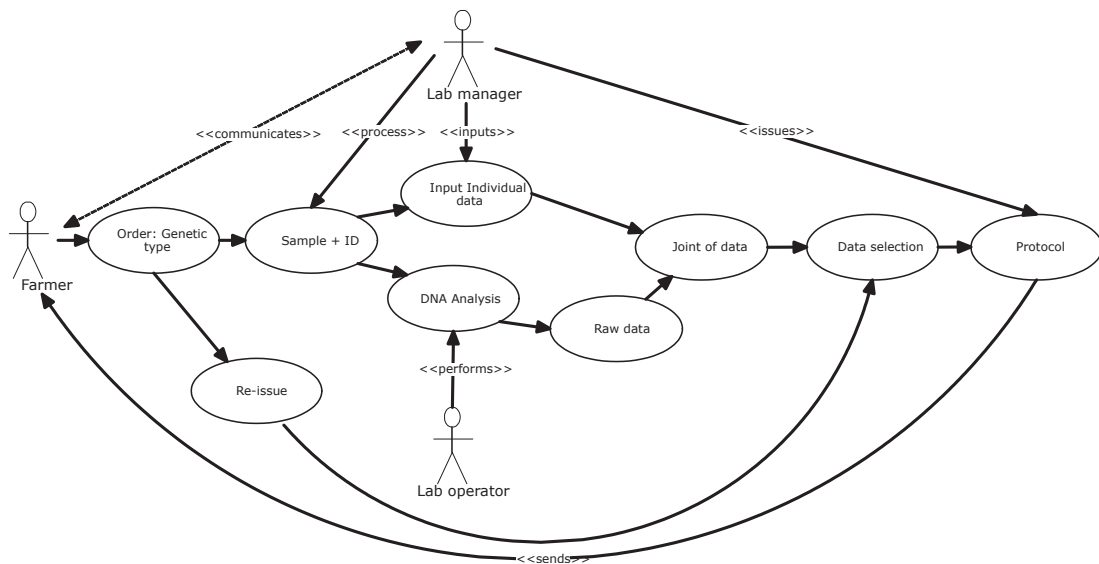


Figure 5.15 Use case diagram of genetic type test issue.

Second use case displays process of paternity testing issue (figure 5.16). When farmer sends an order, with individual sample which should be tested, he can send samples of desired parents as well, or he can supply their animal IDs when he wants to test individuals tested previously which are considered to be parents of testing individual. Also, he can do combination of both or he can supply set of possible mothers/fathers separately or he can ask to find possible parents from animals inserted in the database. Next, the whole process of data inputs, analysis and actors interactions is simmlar to previous use case described above. Only more than one analysis is done typically. Post processing of data consist of selection of possible parents sets:

- automatically by database engine, when parents animal IDs (or names, or lab numbers) are given,
- manually, when number of possible parents is given by breeder, considered by lab management.

Then paternity validation process is execute on all of combinations of desired mothers/fathers for one individual. Depending on paternity testing description wanted by farmer (exclusion of paternity, testing of parents, finding of possible parents, ...) protocol for selected task is issued then.

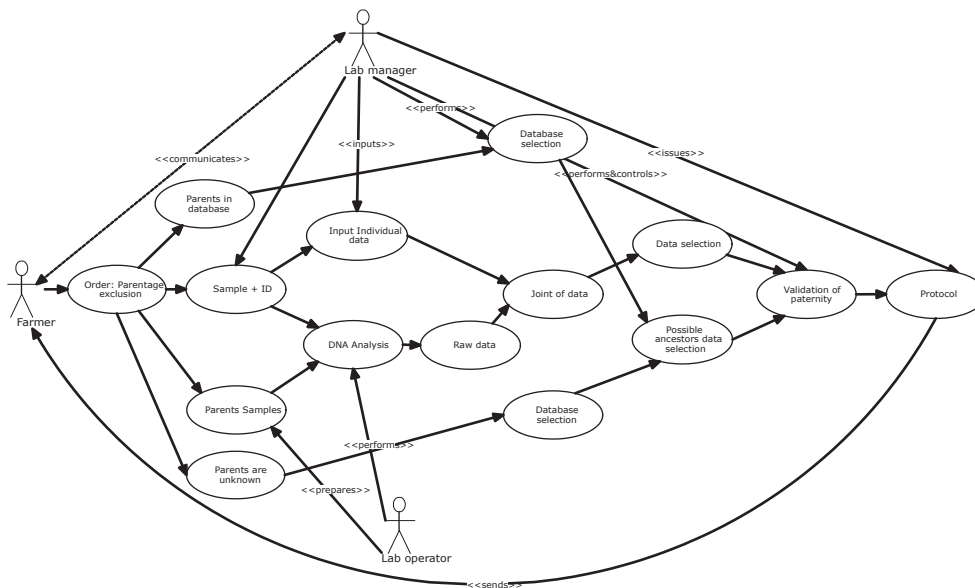


Figure 5.16 Use case diagram of paternity testing issue.

As Lamgen laboratory is accredited by Czech accreditation institute as well as it is part of ISAG/FAO round tests and lab network and issues protocols with DNA profiles and performs paternity testing, it is highly recommended to control usability of chosen micro-sattelite panel for noted tasks. As probability of the same genetic profile, probability of paternity exclusions etc. are based on the results obtained for microsatellite data analyzed for selected group (breed, animals, etc.) previously they can be proved only for “results based” state which is also widely used when genetic profiles play roles in forensics. As inbreeding, special breeding strategies can reduce genetic variability dramatically, set given by ISAG/FAO is updated based on routine genotyping results time to time. For this purpose, the software system contains calculations for used panel evaluation as well. This evaluation process could be described by diagram on figure 5.17. Farmer/customer has feedback of his results given by statistical calculation done over the whole database day-by-day. Authorities have a direct feedback and lab can be connected to world lab networks which contribute on evaluation of panels recommended. At the end, outputs of described statistics can be used for research in genetic variability as well. Simple and fast calculations can bring reliable results on selected groups of animals in comparison with the others on very large database which is filling by daily routine lab operations.

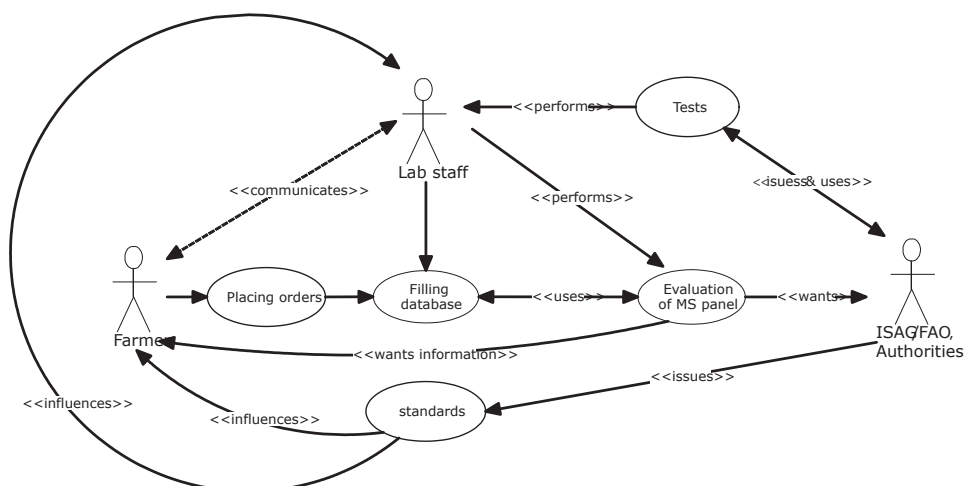


Figure 5.17 Use case diagram of microsatellite panel usage evaluation.

5.3.4 Database

SQL database design used for SW creation is displayed on figure 5.18 in notation of ERD diagram. Basis of database design is created by table Individual with unique key (primary key) LabID. By using this primary key, all of data are paired in the database. Individual table contains then animal ID (issued by authority), breed declaration, desired parents/grandparents IDs, sex of individual, customer IDs and notes. Individual table should be filled by lab manager whenever he receives sample and order manually.

When Individual data are set, Sample table is created for individual (can be created manually as well, when Individual data are unavailable). To individual account, one or many sample tables can be attached, as sample can be resent many times by customer for any reason. Opposite relationship Sample-Individual is unique as the same sample can be link with one and only one individual. Sample table includes LabID as well as identification of sample storage (if it is stored in genetic bank) according to lab identification, dates of sample receiving and testing (input automatically by genetic analyser). Type of sample (blood, tissue, hair, ...) is also stored.

Genetic profile table and Functional genes table are created automatically with relationship one and only one (both directions) when Sample table is created for individual sample. These tables are filled automatically, when sequencing machine outputs results of analysis with LabIDs included in these tables. Tables for sets of desired parents deduced from Individual table are also used in database design with zero-to many bi-directional relationships. These tables are used then for paternity testing when it is recommended and can be filled both manually and semi- or full- automatically.

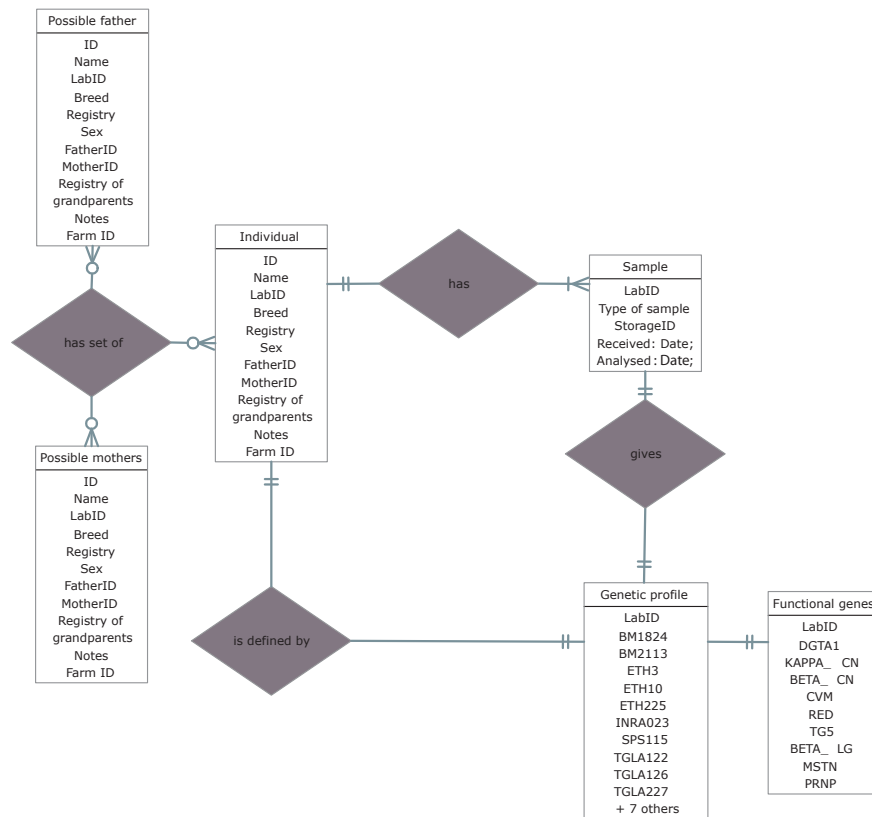


Figure 5.18 ERD diagram of database used for storage of samples data.

Datatypes used for data storage

```
AnimalID VARCHAR (25) CHARACTER SET WIN1250 NOT NULL COLLATE WIN1250,  
AnimalID_corrected VARCHAR (20) CHARACTER SET WIN1250 COLLATE WIN1250,  
Name VARCHAR (40) CHARACTER SET WIN1250 COLLATE WIN1250,  
LAB_ID VARCHAR (15) CHARACTER SET WIN1250 COLLATE WIN1250,  
GEN_BANK_ID VARCHAR (20) CHARACTER SET WIN1250 COLLATE WIN1250,  
GEN_BANK_2ND_ID VARCHAR (5) CHARACTER SET WIN1250 COLLATE WIN1250,  
FARM_ID VARCHAR (45) CHARACTER SET WIN1250 COLLATE WIN1250,  
ANIMAL_REGISTRY VARCHAR (7) CHARACTER SET WIN1250 COLLATE WIN1250,  
SEX VARCHAR (1) CHARACTER SET WIN1250 COLLATE WIN1250,  
BREED VARCHAR (16) CHARACTER SET WIN1250 COLLATE WIN1250,  
REC_DATE DATE,  
DATE_OF_TEST DATE,  
SAMPLE_TYPE VARCHAR (8) CHARACTER SET WIN1250 COLLATE WIN1250,  
PROTOCOL_ISSUED VARCHAR (1) CHARACTER SET WIN1250 COLLATE WIN1250,  
FATHER_ID VARCHAR (25) CHARACTER SET WIN1250 COLLATE WIN1250,  
FATHER_REGISTRY VARCHAR (7) CHARACTER SET WIN1250 COLLATE WIN1250,  
MOTHER_ID VARCHAR (25) CHARACTER SET WIN1250 COLLATE WIN1250,  
FATHER_OF_MOTHER_ID VARCHAR (25) CHARACTER SET WIN1250 COLLATE WIN1250,  
FATHER_OF_MOTHER_REGISTRY VARCHAR (7) CHARACTER SET WIN1250 COLLATE WIN1250,  
NAME_MOTHER VARCHAR (35) CHARACTER SET WIN1250 COLLATE WIN1250,  
NOTE VARCHAR (100) CHARACTER SET WIN1250 COLLATE WIN1250,  
FATHER_NAME VARCHAR (35) CHARACTER SET WIN1250 COLLATE WIN1250,  
FATHER_OF_MOTHER_NAME VARCHAR (35) CHARACTER SET WIN1250 COLLATE WIN1250,  
BM1824_1 VARCHAR (3) CHARACTER SET WIN1250 COLLATE WIN1250,  
BM1824_2 VARCHAR (3) CHARACTER SET WIN1250 COLLATE WIN1250,  
BM2113_1 VARCHAR (3) CHARACTER SET WIN1250 COLLATE WIN1250,  
BM2113_2 VARCHAR (3) CHARACTER SET WIN1250 COLLATE WIN1250,  
ETH3_1 VARCHAR (3) CHARACTER SET WIN1250 COLLATE WIN1250,  
ETH3_2 VARCHAR (3) CHARACTER SET WIN1250 COLLATE WIN1250,  
ETH10_1 VARCHAR (3) CHARACTER SET WIN1250 COLLATE WIN1250,  
ETH10_2 VARCHAR (3) CHARACTER SET WIN1250 COLLATE WIN1250,  
ETH225_1 VARCHAR (3) CHARACTER SET WIN1250 COLLATE WIN1250,  
ETH225_2 VARCHAR (3) CHARACTER SET WIN1250 COLLATE WIN1250,  
INRA023_1 VARCHAR (3) CHARACTER SET WIN1250 COLLATE WIN1250,  
INRA023_2 VARCHAR (3) CHARACTER SET WIN1250 COLLATE WIN1250,  
SPS115_1 VARCHAR (3) CHARACTER SET WIN1250 COLLATE WIN1250,  
SPS115_2 VARCHAR (3) CHARACTER SET WIN1250 COLLATE WIN1250,  
TGLA122_1 VARCHAR (3) CHARACTER SET WIN1250 COLLATE WIN1250,  
TGLA122_2 VARCHAR (3) CHARACTER SET WIN1250 COLLATE WIN1250,  
TGLA126_1 VARCHAR (3) CHARACTER SET WIN1250 COLLATE WIN1250,  
TGLA126_2 VARCHAR (3) CHARACTER SET WIN1250 COLLATE WIN1250,  
TGLA227_1 VARCHAR (3) CHARACTER SET WIN1250 COLLATE WIN1250,  
TGLA227_2 VARCHAR (3) CHARACTER SET WIN1250 COLLATE WIN1250,  
DGAT1 VARCHAR (10) CHARACTER SET WIN1250 COLLATE WIN1250,  
KAPPA_CN VARCHAR (10) CHARACTER SET WIN1250 COLLATE WIN1250,  
BETA_CN VARCHAR (10) CHARACTER SET WIN1250 COLLATE WIN1250,  
CVM VARCHAR (10) CHARACTER SET WIN1250 COLLATE WIN1250,  
RED VARCHAR (3) CHARACTER SET WIN1250 COLLATE WIN1250,  
TG5 VARCHAR (3) CHARACTER SET WIN1250 COLLATE WIN1250,  
BETA_LG VARCHAR (3) CHARACTER SET WIN1250 COLLATE WIN1250,  
MSTN VARCHAR (3) CHARACTER SET WIN1250 COLLATE WIN1250,  
PRNP VARCHAR (3) CHARACTER SET WIN1250 COLLATE WIN1250
```

5.3.5 Data operations and SQL queries

This chapter describes key SQL queries used in database interface. As SQL can significantly reduce programming effort, we can demonstrate its power in genetic data as well on following examples. As the system stores data about thousands of individuals and for all of processes described in chapter 5.3.3 selections must be used, searching engine is presented

5 Results and Discussion

in software. User can define search criteria by given numbers, text variables or intervals. Also, there is input field which allows to input SQL query defined by user over tables in the system, so for selecting particular groups of animals the whole SQL syntax can be used. For example, when we need to know actual numbers of individuals from all of breeds included in database, where at least half of microsatellite data are entered, we can use SQL query like:

```
select count(*), plemeno from skot where CHAR_LENGTH (bm1824_1
||bm1824_2||bm2113_1||bm2113_2||eth3_1||eth3_2||ETH10_1||ETH10_
2||ETH225_1||ETH225_2||INRA023_1||INRA023_2||SPS115_1||SPS115_2
||TGLA122_1||TGLA122_2||TGLA126_1||TGLA126_2||TGLA227_1||TGLA22
7_2)>10 and plemeno!='' and plemeno!=' ' group by plemeno;
```

As one would need to operate over the selected group of individuals (e.g. during manual selection of possible parents among individuals 1) from one farm and/or 2) born before 2010, etc.), the whole system works in two modes:

- whole database mode - the whole database of all individuals is used. This mode is basic one. User can do all of things - searching animals in table manually, issuing protocols, creates parents sets, etc. Also, user can define search conditions in searching engine, type in SQL query and run it.
- selection mode - when selection is performed, number of selected animals is changed and selected group is displayed in table below search engine. User can go back to whole database mode using button. All of interface functionality is enable as well in this mode.

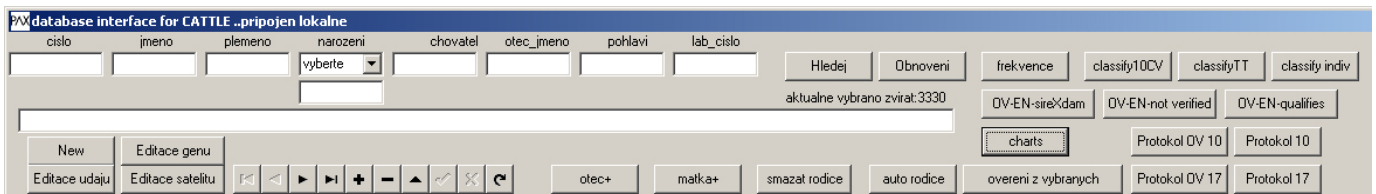


Figure 5.19 Search engine of database interface.

When search conditions are specified, dynamic SQL view is created by database interface. E.g. when user specifies that he wants to select all of bulls with breed declared as pure Czech Fleckvieh from database, interface then runs following SQL query:

```
CREATE VIEW SELECT_VIEW AS select * from SKOT where POHLAVI
like '%M%' and PLEMENO like '%C100%';
```

Following query can be used to obtain table with all of alleles from microsatellite *BM1824*, which are defined and are not unknown.

```
CREATE VIEW SATELIT (SAT) AS SELECT BM1824_1 FROM SKOT where
BM1824_1 !='?' UNION ALL SELECT BM1824_2 FROM SKOT where
BM1824_2 !='?' ;
```

5 Results and Discussion

When we apply next query to previously created dynamic view SATELIT, when bulls from Czech Fleckvieh are selected previously by search engine, table with different alleles, their counts and frequencies will be created for selected group of animals as:

```
select SATELIT.SAT, count(SATELIT.SAT), cast(count(SATELIT.SAT)
as float)/(select count(*) from SATELIT) from SATELIT group by
SAT;
```

This query will output table:

| Allele | Count | Frequency |
|--------|-------|-----------|
| 178 | 98 | 0.200000 |
| 180 | 122 | 0.248979 |
| 182 | 154 | 0.314286 |
| 186 | 1 | 0.002041 |
| 188 | 114 | 0.232665 |
| 190 | 1 | 0.002041 |

Table 5.19 SQL output of query to calculate frequencies and counts of different alleles in defined loci.

Given example is necessary for calculations needed to panel evaluation (CEP, PIC, etc.) as it is described in section 5.3.6.

5.3.6 Algorithms

Main algorithms of software implementation of the system are presented in this section.

Parsing of SQL selection query, enabling selection operational mode.

```
s:='';
flag:=[rfReplaceAll];

if edit1.Text<>' ' then s:=s+'JMENO like \''+edit1.Text+'%'+edit1.Text+'',';
if (edit3.Text<>' ')and(combobox1.Text<>'vyberte') then s:=s+'DATUM_
NAROZENI'+combobox1.Text+''+edit3.Text+'',';
if edit4.Text<>' ' then s:=s+'CHOV like \''+edit4.Text+'%'+edit4.Text+'',';
if edit5.Text<>' ' then s:=s+'OTEC_CISLO like \''+edit5.Text+'%'+edit5.Text+'',';
if edit7.Text<>' ' then s:=s+'LAB_CISLO like \''+edit7.Text+'%'+edit7.Text+'',';
if edit6.Text<>' ' then s:=s+'POHLAVI like \''+edit6.Text+'%'+edit6.Text+'',';
if edit2.Text<>' ' then s:=s+'PLEMENO like \''+edit2.Text+'%'+edit2.Text+'',';
if edit9.Text<>' ' then s:=s+'CISLO like \''+edit9.Text+'%'+edit9.Text+'',';
if s[length(s)]=',' then setlength(s,length(s)-1);
s:=stringreplace(s,',',' ' and ',flag);

if edit8.Text<>' ' then s:=edit8.Text;

ShowMessage(s);

ibquery1.SQL.Clear;
ss:='DROP VIEW '+ipaddr+'SATELIT';
ibquery1.SQL.Add(ss);
try begin ibquery1.open;
```

```
        ibquery1.Active; end except end;

ibquery1.SQL.Clear;
ss:='DROP VIEW '+ipaddr+'SELECT_VIEW';
ibquery1.SQL.Add(ss);
try begin ibquery1.open;
    ibquery1.Active; end except end;

ibquery1.SQL.Clear;
ss:='CREATE VIEW '+ipaddr+'SELECT_VIEW AS '+select * from SKOT where '+s+'';

ibquery1.SQL.Add(ss);
ibquery1.open;
ibquery1.Active;

label16.Caption:='aktualne vybrano zvirat:'+inttostr(ibtable3.recordcount);
ibtable3.first;
view:=true;
end;
```

Calculation of alleles frequencies for all of loci - stored in hash table with Allele lenght and its frequency.

```
var pole_h: array [1..10] of tdihash;
    i,ii,sat,length,pom: integer;
    iii: ^real;
    i1,i2: real;
    ss: string;
    f: textfile;
begin
    for i:=1 to 10 do begin
        pole_h[i]:=tdihash.Create(realhandler,getdicardinalkeyhandler);
    end;
    for i:=1 to 10 do begin
        ss:='DROP VIEW '+ipaddr+'SATELIT';
        ss:='CREATE VIEW '+ipaddr+'SATELIT (SAT) AS SELECT '+form1.combobox3.Items.
Strings[i-1]+'_1 FROM SKOT WHERE '+form1.combobox3.Items.Strings[i-1]+'_1 != ''?''
UNION ALL SELECT '+form1.combobox3.Items.Strings[i-1]+'_2 FROM SKOT WHERE '+form1.com-
bobox3.Items.Strings[i-1]+'_2 != ''?''';
        form1.ibquery2.SQL.Add('select '+ipaddr+'SATELIT.SAT, count('+ipaddr+'SATELIT.SAT),
cast(count('+ipaddr+'SATELIT.SAT)as float)/(select count(*) from '+ipaddr+'SATELIT) from
'+ipaddr+'SATELIT group by SAT;');
        for ii:=1 to form1.ibquery2.RecordCount do begin
            length:=form1.IBQuery2.Fields[0].asinteger;
            iii:=pole_h[i].InsertItemByKey(length);
            iii^:=form1.IBQuery2.Fields[2].AsFloat;
            form1.ibquery2.Next;
        end;
    end;
end;
```

Calculation of Combined Exclusion Probabilities

```
function CEP(cislo,sat: integer): real;
var ss: string;
    x2,x3,x4,x5,x6: real;
    i: integer;
begin
    if view=true then begin
        form1.ibquery1.SQL.Clear;
        ss:='DROP VIEW '+ipaddr+'SATELIT';
        form1.ibquery1.SQL.Clear;
        ss:='CREATE VIEW '+ipaddr+'SATELIT (SAT) AS SELECT '+form1.combobox3.Items.
Strings[sat-1]+'_1 FROM '+ipaddr+'SELECT_VIEW WHERE '+form1.combobox3.Items.
Strings[sat-1]+'_1 != ''?'' UNION ALL SELECT '+form1.combobox3.Items.Strings[sat-1]+'_2
```

5 Results and Discussion

```
FROM '+ipaddr+'SELECT_VIEW WHERE '+form1.combobox3.Items.Strings[sat-1]+'_2 != ''?'';
    form1.ibquery2.SQL.Add('select '+ipaddr+'SATELIT.SAT, count('+ipaddr+'SATELIT.
SAT), cast(count('+ipaddr+'SATELIT.SAT)as float)/(select count(*) from
'+ipaddr+'SATELIT) from '+ipaddr+'SATELIT group by SAT;');
    x2:=0; x3:=0; x4:=0; x5:=0; x6:=0;
    for i:=1 to form1.ibquery2.RecordCount do begin
        x2:=x2+sqr(form1.ibquery2.Fields[2].AsFloat);
        x3:=x3+power(form1.ibquery2.Fields[2].AsFloat,3);
        x4:=x4+power(form1.ibquery2.Fields[2].AsFloat,4);
        x5:=x5+power(form1.ibquery2.Fields[2].AsFloat,5);
        x6:=x6+power(form1.ibquery2.Fields[2].AsFloat,6);
        form1.ibquery2.Next;
    end;

    case cislo of
    1: CEP:=(1-2*x2-(sqr(x2))+x4);
    2: CEP:=(1-4*x2+2*(sqr(x2))+4*x3-3*x4);
    3: CEP:=(1+4*x4-4*x5-3*x6-8*(sqr(x2))+8*x2*x3+2*(sqr(x3)));
    end;
```

Paternity testing condition for one loci

```
if not(((dbgrid3.Fields[23].AsString=dbgrid1.Fields[23].AsString) or (dbgrid3.
Fields[23].AsString=dbgrid1.Fields[24].AsString)) and ((dbgrid3.Fields[24].
AsString=dbgrid2.Fields[23].AsString) or (dbgrid3.Fields[24].AsString=dbgrid2.
Fields[24].AsString)) )
    or
    (((dbgrid3.Fields[23].AsString=dbgrid2.Fields[23].AsString) or (db-
grid3.Fields[23].AsString=dbgrid2.Fields[24].AsString)) and ((dbgrid3.Fields[24].
AsString=dbgrid1.Fields[23].AsString) or (dbgrid3.Fields[24].AsString=dbgrid1.
Fields[24].AsString)) ) )
    then begin
        dbgrid3.Columns.Items[23].Color:=clred;
        dbgrid3.Columns.Items[24].Color:=clred;
        label69.Color:=clred;
        label70.Color:=clred;
        test:=false;
    end;
```

Sorting by allele frequency for one loci

```
if (ibtable1.Fields[25].AsString='') or (ibtable1.Fields[26].AsString='') then
write(f,'?'+#9+'?'+#9) else begin
    pom:=ibtable1.Fields[25].AsInteger;
    i1:=real(pole_h[2].pitomofkey(pom)^);
    pom:=ibtable1.Fields[26].AsInteger;
    i2:=real(pole_h[2].pitomofkey(pom)^);
    if i1<=i2 then begin
        write(f,ibtable1.fields[25].AsString+#9+ibtable1.fields[26].AsString+#9);
    end else begin
        write(f,ibtable1.fields[26].AsString+#9+ibtable1.fields[25].AsString+#9);
    end;
end;
```

5.3.7 Security

As security and safety of data, security of operations and their traceability can be significant in accredited laboratory, several functions in software interface are created. For security and traceability purposes, there is implementation in software of getting IP address from computer which wants to connect database. This procedure is implemented in following code:

IP address catch up.

```
flag:=[rfReplaceAll];
WSAStartup($101, GInitData);
IPAddr:= '';
GetHostName(Buffer, SizeOf(Buffer));
phe :=GetHostByName(buffer);
if phe = nil then Exit;
pptr := PaPInAddr(Phe^.h_addr_list);
i := 0;
while pptr^[i] <> nil do
begin
  ipaddr:=stringreplace(StrPas(inet_ntoa(pptr^[i]^)),'.',' ',flag);
  Inc(i);
end;
WSACleanup;
ipaddr:=stringreplace(ipaddr, ' ', '', flag);
ipaddr:=stringreplace(ipaddr, '0', 'a', flag);
ipaddr:=stringreplace(ipaddr, '1', 'b', flag);
ipaddr:=stringreplace(ipaddr, '2', 'c', flag);
ipaddr:=stringreplace(ipaddr, '3', 'd', flag);
ipaddr:=stringreplace(ipaddr, '4', 'e', flag);
ipaddr:=stringreplace(ipaddr, '5', 'f', flag);
ipaddr:=stringreplace(ipaddr, '6', 'g', flag);
ipaddr:=stringreplace(ipaddr, '7', 'h', flag);
ipaddr:=stringreplace(ipaddr, '8', 'i', flag);
ipaddr:=stringreplace(ipaddr, '9', 'j', flag);
ipaddr:=uppercase(ipaddr);
end;
```

Then, obtained IP address is used as identifier in all of operations which are done with database like dynamic views creations, SQL queries executions etc. With all operations, address is logged like:

```
ss:='CREATE VIEW '+ipaddr+'SELECT_VIEW AS '+'select * from SKOT where '+s+'';
```

so, IP address of machine is logged in database everytime when dynamic view (searching over database) is involved. Administrator can easily control, who operates with, searches and views data plus which data he selected for operations. As well, blocking of operations for undefined IP addresses can be implemented in Firebird SQL server as a trigger and can not allow non-granted users to access database.

Another security issues are connected with network design proposed in section 5.3.1. Open and closed zones plus firewalls offer great security for the system as protection against DoS, external access etc. are. Design respects fully open internal network zone and closed one way link to the Internet.

Database is daily backed up on dedicated server in internal network as files when Workstation 1 belonging to lab manager is started. As well every day it is backed up in database server internally which is treated by Firebird SQL Server, so the rollback on database can be done in daily steps, what can prevent huge data lost.

5.3.8 GUI

Following text describes GUI (Graphical User Interface) during usage of database interface. GUI is designed with respect to needs of operational staff in laboratory as well as small time-saving functions (detailed Tab orders etc.).

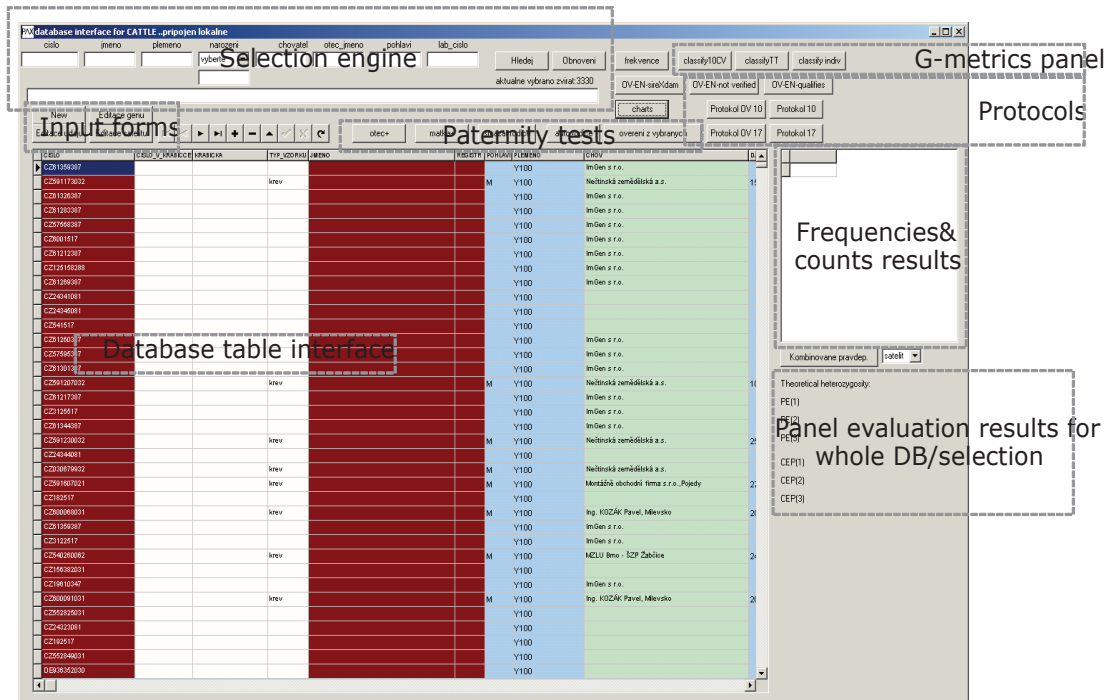


Figure 5.20 Basic user interface of database system.

When user wants to input data (e.g. animal ID data) following GUI is appeared. Form allows to input data for new individual as well as partly for new sample delivered. When <- sign is appeared, the button allows user to search data previously entered to the database to avoid mistyping etc. When OK button is pressed, new item is stored into the database.

The 'editace udaju' dialog box contains several input fields and buttons. The top section has 'lab_císlo' (krabicka), 'císlo_v_krabickce', 'typ_vzorku', and 'datum prijati'. The middle section has 'císlo', 'císlo_puvodni', 'registr', 'jméno', 'pohlaví', 'plemeno', and 'narozeni' (chov). The bottom section has 'otec_císlo', 'otec_registr', 'otec_jméno', 'matka_císlo', 'matka_jméno', 'otec_matky_císlo', 'otec_matky_registr', and 'otec_matky_jméno'. There are 'OK' and 'Cancel' buttons at the bottom.

Figure 5.21 Inserting of new sample.

When new sample is inserted, MS data can be inputted manually as well or they can be corrected after automatical pairing with raw data from genetic analyser.

| | | | |
|----|-----|-----|----------|
| 1 | 178 | 180 | BM1824 |
| 2 | 127 | 135 | BM2113 |
| 3 | 117 | 117 | ETH3 |
| 4 | ? | ? | ETH10 |
| 5 | 148 | 148 | ETH225 |
| 6 | 206 | 206 | INRA023 |
| 7 | ? | ? | SPS115 |
| 8 | 151 | 153 | TGLA122 |
| 9 | 115 | 121 | TGLA126 |
| 10 | 87 | 93 | TGLA227 |
| 11 | | | SPS113 |
| 12 | | | BM1818 |
| 13 | | | RM067 |
| 14 | | | ILSTS006 |
| 15 | | | MGTG4B |
| 16 | | | CSSM66 |
| 17 | | | CSRM60 |

poznamka

OK

Figure 5.22 Editation of genotype data.

Also, genotyped data can be edited in Tab order mode, so user can just press Tab key between each field. Corrections can be done in this way really easily.

Charts button shows G-metric classification results on chosen individual started by Classify indiv button. Classify TT button runs classification performs on the whole database, when classification is validated on the the whole dataset as well - training set equals test set. Classify 10 CV button runs 10 fold cross validation across the whole dataset using G-metric algorithm. Both of button lead to save model to 'trainset' and classification results to 'output' files. 'Trainset' represents saving of dynamic containers used for fast calculation of classification results, 'output' represents tab separated text file contains all of results. Frekvence button outputs *AlleleFrequencyDS* based on displayed selection.

Protocols button group is used to issue, printing and exporting (to PDF) selected protocols described below.

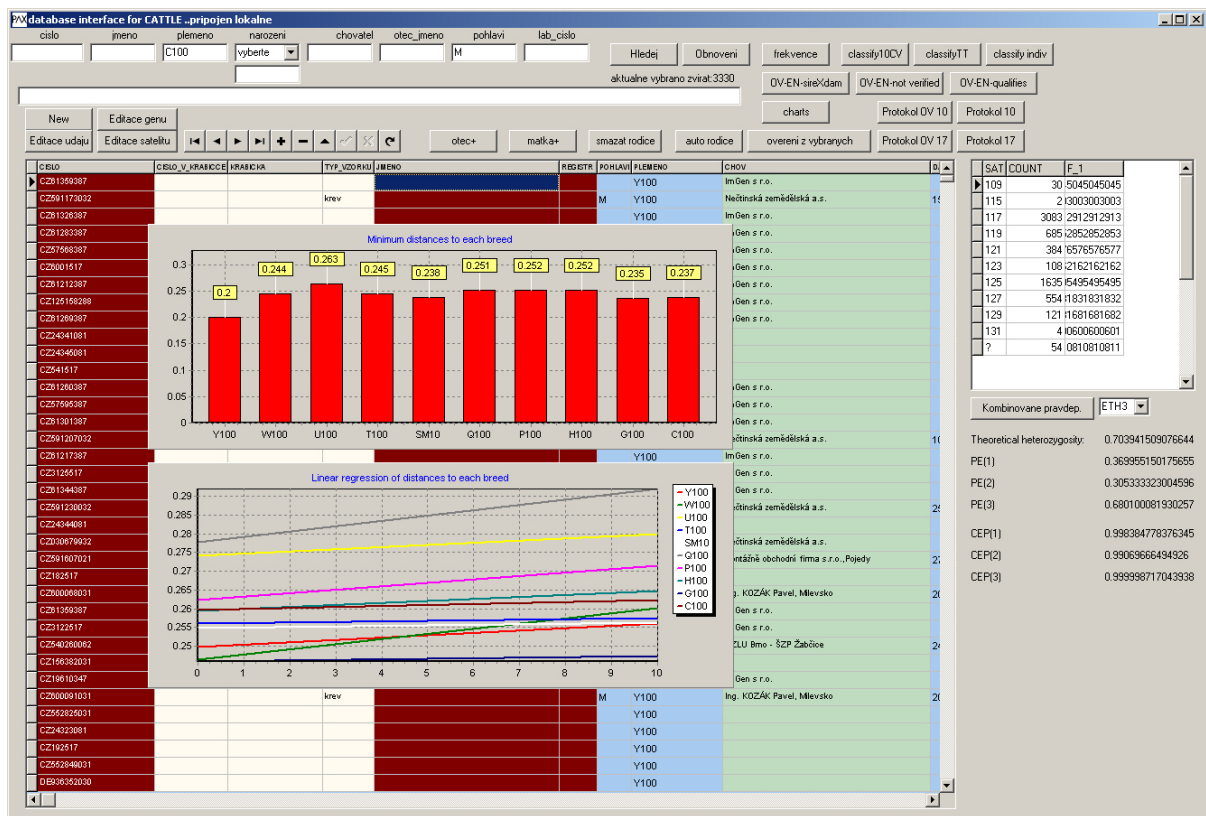


Figure 5.23 Fully used GUI of DB interface.

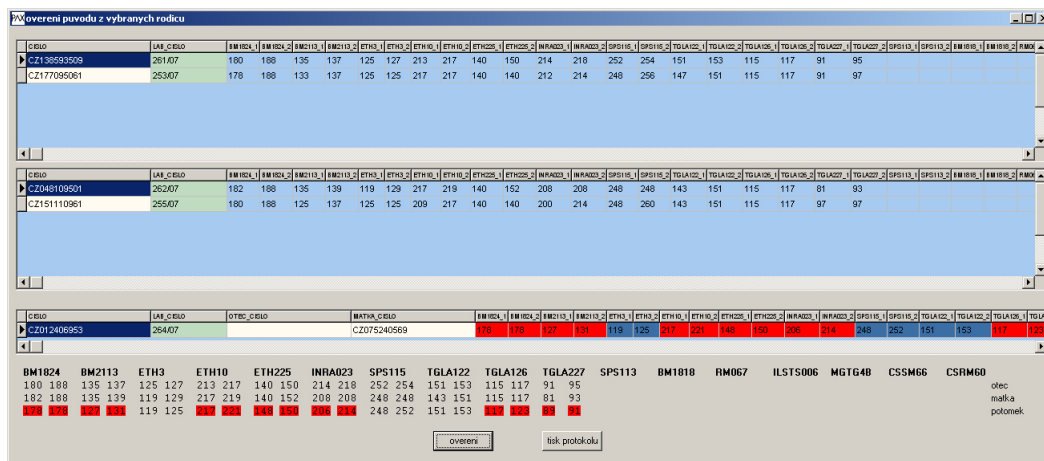



Figure 5.24 Paternity verification screen.

5.3.9 Protocols

On request, following protocols can be issued from SW database interface. All of protocols can be exported to PDF files or printed on Windows installed printer. Dates etc. are filled automatically as well as data from database (whole or selection mode). Fast Reports engine is used to produce protocols. Issues are mainly parentage exclusion protocols, parentage verification protocols, non-exclusion protocols and genetic profile protocols in Czech as well as English versions. Current laws and accreditation prescriptions are accepted in protocols. Also, they can be issued for 10 or 17 MS panels.

5 Results and Discussion

Mendel University in Brno
Laboratory of Agrogenomics
Testing Laboratory No. L 1030.3 accredited CIA by CSN EN ISO/IEC 17025
Zemědělská 1, 613 00 Brno, Czech Republic
ISAG Lab. code: CS/B



L 1030.3

Bovine DNA Type Card

Reg. No.: CZ579314061

Name/Register: C100

Breed: C100

Sex: M

Date of Birth:

Lab. Ref. Number: S 849/07

Date tested: 3.10.2007

Sire:

Reg. No.: HG212

Dam:

Reg. No.: CZ158064964

PARENTAGE ANALYSIS:
Given sire HG 212 can not be excluded, the dam has not be verified.

MICROSATELLITES

| | | |
|----------|-----|-----|
| BM2113 | 127 | 133 |
| ETH00 | 209 | 221 |
| SPS115 | 248 | 248 |
| TGLA227 | 81 | 91 |
| INRA323 | 208 | 214 |
| TGLA122 | 151 | 161 |
| TGLA126 | 113 | 123 |
| BM1824 | 178 | 182 |
| ETH225 | 150 | 150 |
| ETH3 | 117 | 125 |
| SPS113 | | |
| BM1818 | | |
| RM067 | | |
| ILSTS006 | | |
| MCTGD8 | | |
| CSSM66 | | |
| CSRM60 | | |

Breeder: CHDIMPULS, ZOD ČÁSLAVICE, ČÁSLAVICE 120,675 24


Brno 04. 03. 2012

.....
Irena Vrhová, Ing. Bc.
Laboratory Assistant, Manager

1/1

.....
Josef Dvořák, Univ. Prof. Dr.h.c.
Laboratory Manager

Mendel University in Brno
Laboratory of Agrogenomics
Testing Laboratory No. L 1030.3 accredited CIA by CSN EN ISO/IEC 17025
Zemědělská 1, 613 00 Brno, Czech Republic
ISAG Lab. code: CS/B



L 1030.3

Bovine DNA Type Card

Reg. No.: CZ184302961

Name/Register: C100

Breed: C100

Sex: k

Date of Birth:

Lab. Ref. Number: S 934/08

Date tested: 14.10.2008

Sire:

Reg. No.:

Dam:

Reg. No.:

PARENTAGE ANALYSIS:
Parentage not verified.

MICROSATELLITES

| | | |
|----------|-----|-----|
| BM2113 | 131 | 135 |
| ETH00 | 217 | 217 |
| SPS115 | 248 | 248 |
| TGLA227 | 79 | 81 |
| INRA323 | 202 | 208 |
| TGLA122 | 161 | 161 |
| TGLA126 | 115 | 117 |
| BM1824 | 180 | 182 |
| ETH225 | 150 | 150 |
| ETH3 | 119 | 128 |
| SPS113 | | |
| BM1818 | | |
| RM067 | | |
| ILSTS006 | | |
| MCTGD8 | | |
| CSSM66 | | |
| CSRM60 | | |

Breeder: ZOD Čáslavice, 675 24 ČÁSLAVICE

Brno 04. 03. 2012

.....
Irena Vrhová, Ing. Bc.
Laboratory Assistant, Manager

1/1

.....
Josef Dvořák, Univ. Prof. Dr.h.c.
Laboratory Manager

Figure 5.25 “Sire can not be excluded” protocol.

Figure 5.26 “Parentage can not be verified” protocol.

5.4 Proving of usability of machine learning methods in cattle breed discrimination task

This chapter shows and discuss results of usage of machine learning algorithms and their modifications for cattle breed discrimination task. Also, three types of data representations are used to explore which of them is most suitable for described issue, so chapter is divided according this point of view to the whole problem. Latest, usage of the results for genetic diversity description of breeds is discussed in this part of thesis.

Usability of machine learning algorithms for cattle breed discrimination can be evaluated by many criteria. As results of methods are highly dependent on dataset, parameters used for classification algorithm and many parameters are calculated then as the results of final model classification, it is not easy to compare results. Imagine that e.g. one badly classified class (breed in our case, genetically non uniform) can significantly influence results of the whole classification. In this case, results need to be discussed from many points of view. In this section, there are presented all of results obtained for each chosen method and each algorithm. Only best results are presented for each method with parameters set accordingly. As the main parameter of good-of-fitness of classification model, we can assume percentage of correctly classified instances (calculated across 10 folds as average results for all of classes) and Kappa statistic as indicator, how better model is in comparison with random classification based only on dataset character - observed probabilities. Also, overall results as FP Rate (which can express mistakes done by model important for breed discrimination), Precision and F-Measure as overall good-of-fitness parameter should be discussed. Then, we need to discuss how classes are classified by particular method according to whole results as well.

5.4.1 ZeroR

Parameters used

Scheme: weka.classifiers.rules.ZeroR

Results of classification

```

Correctly Classified Instances      730          21.9219 %
Incorrectly Classified Instances    2600         78.0781 %
Kappa statistic                    0
Mean absolute error                 0.1678
Root mean squared error             0.2896
Relative absolute error             100%
Root relative squared error         100%
Total Number of Instances          3330

```

=== Detailed Accuracy By Class ===

| | TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area | Class |
|---------------|---------|---------|-----------|--------|-----------|----------|-------|
| | 0 | 0 | 0 | 0 | 0 | 0.495 | Y100 |
| | 0 | 0 | 0 | 0 | 0 | 0.481 | W100 |
| | 0 | 0 | 0 | 0 | 0 | 0.492 | U100 |
| | 0 | 0 | 0 | 0 | 0 | 0.498 | T100 |
| | 1 | 1 | 0.219 | 1 | 0.36 | 0.5 | SM100 |
| | 0 | 0 | 0 | 0 | 0 | 0.485 | Q100 |
| | 0 | 0 | 0 | 0 | 0 | 0.49 | P100 |
| | 0 | 0 | 0 | 0 | 0 | 0.495 | H100 |
| | 0 | 0 | 0 | 0 | 0 | 0.5 | G100 |
| | 0 | 0 | 0 | 0 | 0 | 0.497 | C100 |
| Weighted Avg. | 0.219 | 0.219 | 0.048 | 0.219 | 0.079 | 0.497 | |

| a | b | c | d | e | f | g | h | i | j | |
|---|---|---|---|-----|---|---|---|---|---|-------|
| 0 | 0 | 0 | 0 | 188 | 0 | 0 | 0 | 0 | 0 | Y100 |
| 0 | 0 | 0 | 0 | 66 | 0 | 0 | 0 | 0 | 0 | W100 |
| 0 | 0 | 0 | 0 | 137 | 0 | 0 | 0 | 0 | 0 | U100 |
| 0 | 0 | 0 | 0 | 705 | 0 | 0 | 0 | 0 | 0 | T100 |
| 0 | 0 | 0 | 0 | 730 | 0 | 0 | 0 | 0 | 0 | SM100 |
| 0 | 0 | 0 | 0 | 73 | 0 | 0 | 0 | 0 | 0 | Q100 |
| 0 | 0 | 0 | 0 | 125 | 0 | 0 | 0 | 0 | 0 | P100 |
| 0 | 0 | 0 | 0 | 243 | 0 | 0 | 0 | 0 | 0 | H100 |
| 0 | 0 | 0 | 0 | 700 | 0 | 0 | 0 | 0 | 0 | G100 |
| 0 | 0 | 0 | 0 | 363 | 0 | 0 | 0 | 0 | 0 | C100 |

Table 5.20 Confusion matrix of ZeroR classifier.

ZeroR algorithm is choosing the most frequent class in dataset and then classifies all of instances to the chosen class. It is mainly used as a basis for comparison of efficiency of the other methods (Witten et al., 2011; Berka, 2001), so we can see how results of other classification results are changed and influenced by basis dataset frequency. For sure, increasing of overall calculated measures of model fitting can show, how the particular algorithm is useable in comparison with ZeroR basis.

Table 5.20 and section 5.4.1 show results of ZeroR classifier on general dataset. The same results should be obtained for all of dataset used in this work (as probabilities of classes ob-

5 Results and Discussion

served are the same in them), so only results for general dataset are presented. From results, it is evident that ZeroR classifier classified all of instances of dataset as a class SM100 which is the most frequent in dataset ($n=730$). In this case, when all of unknown distances will be classified as the most frequent class in the training set we can obtain model with 21.9219 % of correctly classified instances. So, any increase of this basis can show that the model with better parameter of correctly classified instances is more usable for classification. Regarding to this, Kappa statistics calculated for ZeroR model equals 0 and all of other parameters of classification power of the method are calculated accordingly.

5.4.2 J48

5.4.2.1 General Dataset

Parameters used

Scheme: weka.classifiers.trees.J48 -U -M 2

Results of classification

```

Correctly Classified Instances      1776      53.3333 %
Correctly Classified Instances      1776      53.3333 %
Incorrectly Classified Instances    1554      46.6667 %
Kappa statistic                    0.4357
Mean absolute error                 0.1017
Root mean squared error             0.2718
Relative absolute error              60.6031 %
Root relative squared error          93.8604 %
Total Number of Instances          3330
    
```

=== Detailed Accuracy By Class ===

| | TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area | Class |
|---------------|---------|---------|-----------|--------|-----------|----------|-------|
| | 0.277 | 0.039 | 0.297 | 0.277 | 0.287 | 0.732 | Y100 |
| | 0.439 | 0.009 | 0.509 | 0.439 | 0.472 | 0.839 | W100 |
| | 0.518 | 0.023 | 0.49 | 0.518 | 0.504 | 0.813 | U100 |
| | 0.711 | 0.101 | 0.655 | 0.711 | 0.682 | 0.847 | T100 |
| | 0.61 | 0.161 | 0.516 | 0.61 | 0.559 | 0.785 | SM100 |
| | 0.11 | 0.008 | 0.242 | 0.11 | 0.151 | 0.697 | Q100 |
| | 0.072 | 0.014 | 0.167 | 0.072 | 0.101 | 0.604 | P100 |
| | 0.527 | 0.036 | 0.538 | 0.527 | 0.532 | 0.797 | H100 |
| | 0.634 | 0.127 | 0.57 | 0.634 | 0.6 | 0.818 | G100 |
| | 0.245 | 0.044 | 0.403 | 0.245 | 0.305 | 0.674 | C100 |
| Weighted Avg. | 0.533 | 0.095 | 0.513 | 0.533 | 0.518 | 0.785 | |

| a | b | c | d | e | f | g | h | i | j | |
|----|----|----|-----|-----|---|---|-----|-----|----|-------|
| 52 | 3 | 15 | 11 | 26 | 2 | 8 | 15 | 42 | 14 | Y100 |
| 3 | 29 | 0 | 10 | 1 | 0 | 2 | 4 | 17 | 0 | W100 |
| 8 | 2 | 71 | 6 | 13 | 1 | 1 | 4 | 28 | 3 | U100 |
| 12 | 6 | 6 | 501 | 90 | 2 | 5 | 11 | 63 | 9 | T100 |
| 19 | 0 | 12 | 93 | 445 | 8 | 9 | 16 | 66 | 62 | SM100 |
| 4 | 1 | 1 | 6 | 23 | 8 | 0 | 6 | 16 | 8 | Q100 |
| 15 | 3 | 7 | 12 | 25 | 1 | 9 | 13 | 31 | 9 | P100 |
| 10 | 1 | 4 | 16 | 27 | 5 | 6 | 128 | 32 | 14 | H100 |
| 27 | 11 | 23 | 77 | 75 | 3 | 5 | 22 | 444 | 13 | G100 |
| 25 | 1 | 6 | 33 | 138 | 3 | 9 | 19 | 40 | 89 | C100 |

Table 5.21 Confusion matrix of J48 classifier for *generalDS* dataset.

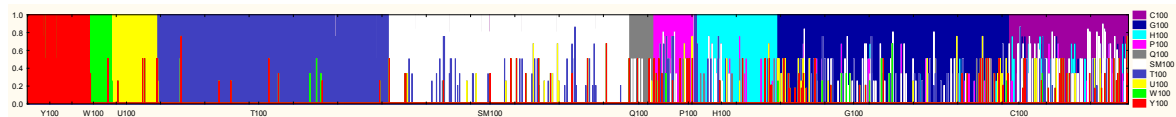


Figure 5.29 Graph of predicted probabilities by J48 classifier on the training *generalDS* set.

5.4.2.2 Allele Length Dataset

Parameters used

Scheme: weka.classifiers.trees.J48 -U -M 2

Results of classification

```

Correctly Classified Instances      1963      58.9489 %
Incorrectly Classified Instances    1367      41.0511 %
Kappa statistic                    0.5056
Mean absolute error                 0.0898
Root mean squared error             0.2609
Relative absolute error             53.5139 %
Root relative squared error         90.0824 %
Total Number of Instances          3330
    
```

=== Detailed Accuracy By Class ===

| | TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area | Class |
|---------------|---------|---------|-----------|--------|-----------|----------|-------|
| | 0.367 | 0.038 | 0.367 | 0.367 | 0.367 | 0.753 | Y100 |
| | 0.394 | 0.009 | 0.481 | 0.394 | 0.433 | 0.73 | W100 |
| | 0.62 | 0.013 | 0.68 | 0.62 | 0.649 | 0.873 | U100 |
| | 0.695 | 0.088 | 0.679 | 0.695 | 0.687 | 0.845 | T100 |
| | 0.655 | 0.14 | 0.568 | 0.655 | 0.609 | 0.806 | SM100 |
| | 0.178 | 0.012 | 0.245 | 0.178 | 0.206 | 0.595 | Q100 |
| | 0.184 | 0.021 | 0.258 | 0.184 | 0.215 | 0.655 | P100 |
| | 0.519 | 0.027 | 0.603 | 0.519 | 0.558 | 0.796 | H100 |
| | 0.773 | 0.083 | 0.713 | 0.773 | 0.742 | 0.887 | G100 |
| | 0.309 | 0.06 | 0.386 | 0.309 | 0.343 | 0.683 | C100 |
| Weighted Avg. | 0.589 | 0.079 | 0.578 | 0.589 | 0.581 | 0.805 | |

| | a | b | c | d | e | f | g | h | i | j | |
|--|----|----|----|-----|-----|---|---|-----|-----|----|-------|
| | 52 | 3 | 15 | 11 | 26 | 2 | 8 | 15 | 42 | 14 | Y100 |
| | 3 | 29 | 0 | 10 | 1 | 0 | 2 | 4 | 17 | 0 | W100 |
| | 8 | 2 | 71 | 6 | 13 | 1 | 1 | 4 | 28 | 3 | U100 |
| | 12 | 6 | 6 | 501 | 90 | 2 | 5 | 11 | 63 | 9 | T100 |
| | 19 | 0 | 12 | 93 | 445 | 8 | 9 | 16 | 66 | 62 | SM100 |
| | 4 | 1 | 1 | 6 | 23 | 8 | 0 | 6 | 16 | 8 | Q100 |
| | 15 | 3 | 7 | 12 | 25 | 1 | 9 | 13 | 31 | 9 | P100 |
| | 10 | 1 | 4 | 16 | 27 | 5 | 6 | 128 | 32 | 14 | H100 |
| | 27 | 11 | 23 | 77 | 75 | 3 | 5 | 22 | 444 | 13 | G100 |
| | 25 | 1 | 6 | 33 | 138 | 3 | 9 | 19 | 40 | 89 | C100 |

Table 5.22 Confusion matrix for J48 classifier for *allelelengthDS*.

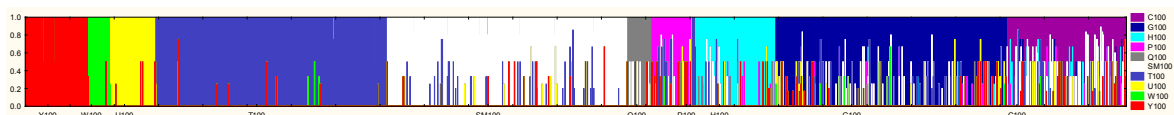


Figure 5.30 Graph of predictions by J48 classifier on the training *allelelengthDS* set.

5.4.2.3 Allele Frequency Dataset

Parameters used

Scheme: weka.classifiers.trees.J48 -U -M 2

Results of classification

```

Correctly Classified Instances      1890           56.7568 %
Incorrectly Classified Instances    1440           43.2432 %
Kappa statistic                    0.4813
Mean absolute error                 0.093
Root mean squared error            0.2669
Relative absolute error             55.454 %
Root relative squared error        92.1533 %
Total Number of Instances         3330
    
```

=== Detailed Accuracy By Class ===

| | TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area | Class |
|---------------|---------|---------|-----------|--------|-----------|----------|-------|
| | 0.351 | 0.038 | 0.357 | 0.351 | 0.354 | 0.746 | Y100 |
| | 0.47 | 0.011 | 0.47 | 0.47 | 0.47 | 0.786 | W100 |
| | 0.613 | 0.017 | 0.604 | 0.613 | 0.609 | 0.857 | U100 |
| | 0.661 | 0.094 | 0.653 | 0.661 | 0.657 | 0.838 | T100 |
| | 0.638 | 0.134 | 0.572 | 0.638 | 0.603 | 0.788 | SM100 |
| | 0.247 | 0.015 | 0.265 | 0.247 | 0.255 | 0.731 | Q100 |
| | 0.216 | 0.024 | 0.262 | 0.216 | 0.237 | 0.632 | P100 |
| | 0.486 | 0.031 | 0.551 | 0.486 | 0.516 | 0.782 | H100 |
| | 0.739 | 0.08 | 0.711 | 0.739 | 0.725 | 0.877 | G100 |
| | 0.267 | 0.068 | 0.324 | 0.267 | 0.293 | 0.63 | C100 |
| Weighted Avg. | 0.568 | 0.08 | 0.559 | 0.568 | 0.562 | 0.793 | |

| | a | b | c | d | e | f | g | h | i | j | |
|----|----|----|-----|-----|----|----|-----|-----|----|---|-------|
| 66 | 1 | 3 | 12 | 34 | 3 | 6 | 9 | 36 | 18 | | Y100 |
| 4 | 31 | 3 | 8 | 6 | 1 | 0 | 1 | 12 | 0 | | W100 |
| 5 | 1 | 84 | 5 | 15 | 1 | 4 | 6 | 11 | 5 | | U100 |
| 7 | 9 | 7 | 466 | 102 | 8 | 9 | 7 | 53 | 37 | | T100 |
| 22 | 1 | 12 | 86 | 466 | 11 | 11 | 20 | 22 | 79 | | SM100 |
| 4 | 1 | 1 | 8 | 11 | 18 | 3 | 6 | 10 | 11 | | Q100 |
| 9 | 0 | 4 | 11 | 18 | 7 | 27 | 16 | 12 | 21 | | P100 |
| 6 | 5 | 11 | 13 | 30 | 5 | 18 | 118 | 22 | 15 | | H100 |
| 32 | 13 | 8 | 64 | 20 | 9 | 6 | 15 | 517 | 16 | | G100 |
| 30 | 4 | 6 | 41 | 113 | 5 | 19 | 16 | 32 | 97 | | C100 |

Table 5.23 Confusion matrix for J48 classifier for *allelefrequencyDS*.

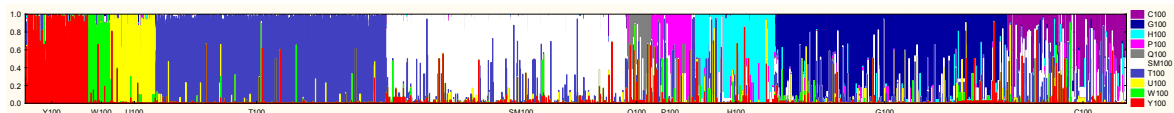


Figure 5.31 Graph of predictions by J48 classifier on the training *allelefrequencyDS* set.

5.4.2.4 Discussion of J48 results

Best results for J48 algorithm used for decision trees induction were obtained for its parameters set as unpruned tree with minimum 2 instances in each tree leaf as section 5.4.2 shows.

For general dataset, algorithm outputs results of classification as it is shown in section 5.4.2.1, table 5.21 and figure 5.29. Over all 10 folds cross validation, model has 53.33 % of correctly classified instances and Kappa statistic equals to 0.4357. Overall FP Rate calculated across all of classes as a weighted average equals 0.095, so with regard to table 5.21 and number of correctly classified instances, it is evident that model at all predicts false negatives rather than false positives values. So, it is better at all because of results predicted are reliable with 53.33 % probability, but there is smaller chance to obtain false positive result than false negative one.

Overall weighted precision is calculated as 0.513, F-measure as overall indicator of model fitting equals 0.518.

The best classified class according to number of correctly classified instances is G100 one with TP Rate=0.634, the worst one is P100 with TP Rate=0.072. FP rate is calculated as the smallest one for W100 class, as biggest for SM100 as 0.161. The most precise classified class is T100 with Precision=0.655 and F-Measure=0.682, the worst one is P100 with Precision=0.167, F-Measure=0.101.

Figure 5.29 shows prediction of model on the whole dataset. It is evident, that Q100, P100, G100 and C100 are not genetically uniform as the other classes in given dataset what corresponds with results above. We can see as well mixture between SM100 and C100 as C100 class is under big influence of portions of probabilities predicted for SM100 class.

For allele length dataset, better overall results were obtained (Section 5.4.2.2). Correctly classified instances percentage calculated for 10 fold cross validation is 58.95 %. Kappa statistic equals 0.5056. Overall FP Rate=0.079. Weighted precision for the whole model is calculated as 0.578, F-measure as 0.581.

The best classified class according to TP Rate is G100 (0.773), the worst is Q100 (0.178). Best FP Rate was calculated for W100 (0.009), the worst one for SM100 (0.14). G100 with Precision=0.713, F-Measure=0.742 is seemed to be as best classified class, in opposite, Q100 with Precision=0.245 and F-Measure=0.206 as the worst one.

Figure 5.30 shows the similar results as were calculated for general dataset. Big admixture of predicted probabilities on training set could be observed mainly in P100 and C100 classes as well as in SM100, G100 and Q100 ones.

For allele frequency dataset (Section 5.4.2.3), overall results were obtained as follows. Correctly classified instances percentage calculated for 10 fold cross validation is 56.76 %. Kappa statistic equals 0.4813. Overall FP Rate=0.08. Weighted precision for the whole model is calculated as 0.559, F-measure=0.562.

The best classified class in TP Rate parameter is G100 (0.739), the worst is P100 (0.216). Best FP Rate was calculated for W100 (0.011), the worst one for SM100 (0.134). G100 with Precision=0.711, F-Measure=0.725 is the best classified class, in opposite, P100 with Precision=0.262 and F-Measure=0.237 as the worst one.

Figure 5.31 shows bigger portion of admixture calculated over all of classes that was produced in general and allele length datasets. Anyway, it corresponds with more balanced spread of admixture for all of classes, so from robustness point of view, this model should be recommended for breed discrimination by J48 as the best one.

5.4.3 JRip

5.4.3.1 General Dataset

Parameters used

Scheme: `weka.classifiers.rules.JRip -F 3 -N 2.0 -O 2 -S 1`

Results of classification

```

Correctly Classified Instances      1758           52.7928 %
Incorrectly Classified Instances    1572           47.2072 %
Kappa statistic                    0.4174
Mean absolute error                0.1195
Root mean squared error            0.2569
Relative absolute error             71.2514 %
Root relative squared error        88.7228 %
Total Number of Instances          3330
    
```

=== Detailed Accuracy By Class ===

| | TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area | Class |
|---------------|---------|---------|-----------|--------|-----------|----------|-------|
| | 0.096 | 0.01 | 0.367 | 0.096 | 0.152 | 0.75 | Y100 |
| | 0.379 | 0.005 | 0.61 | 0.379 | 0.467 | 0.806 | W100 |
| | 0.569 | 0.014 | 0.639 | 0.569 | 0.602 | 0.854 | U100 |
| | 0.687 | 0.087 | 0.679 | 0.687 | 0.683 | 0.841 | T100 |
| | 0.652 | 0.342 | 0.349 | 0.652 | 0.455 | 0.704 | SM100 |
| | 0.192 | 0.004 | 0.5 | 0.192 | 0.277 | 0.723 | Q100 |
| | 0.136 | 0.005 | 0.531 | 0.136 | 0.217 | 0.732 | P100 |
| | 0.502 | 0.018 | 0.689 | 0.502 | 0.581 | 0.816 | H100 |
| | 0.696 | 0.082 | 0.694 | 0.696 | 0.695 | 0.845 | G100 |
| | 0.102 | 0.022 | 0.363 | 0.102 | 0.159 | 0.684 | C100 |
| Weighted Avg. | 0.528 | 0.116 | 0.546 | 0.528 | 0.507 | 0.781 | |

| a | b | c | d | e | f | g | h | i | j | |
|----|----|----|-----|-----|----|----|-----|-----|----|-------|
| 18 | 1 | 1 | 7 | 125 | 2 | 2 | 6 | 22 | 4 | Y100 |
| 0 | 25 | 0 | 1 | 15 | 0 | 1 | 1 | 23 | 0 | W100 |
| 0 | 4 | 78 | 5 | 42 | 0 | 0 | 1 | 7 | 0 | U100 |
| 6 | 4 | 2 | 484 | 128 | 1 | 3 | 3 | 66 | 8 | T100 |
| 4 | 0 | 13 | 121 | 476 | 5 | 3 | 28 | 42 | 38 | SM100 |
| 1 | 0 | 4 | 2 | 42 | 14 | 0 | 1 | 4 | 5 | Q100 |
| 3 | 0 | 0 | 9 | 79 | 1 | 17 | 4 | 10 | 2 | P100 |
| 3 | 1 | 7 | 12 | 81 | 1 | 1 | 122 | 12 | 3 | H100 |
| 7 | 6 | 5 | 29 | 153 | 2 | 1 | 5 | 487 | 5 | G100 |
| 7 | 0 | 12 | 43 | 223 | 2 | 4 | 6 | 29 | 37 | C100 |

Table 5.24 Confusion matrix for JRip classifier for *generalDS*.

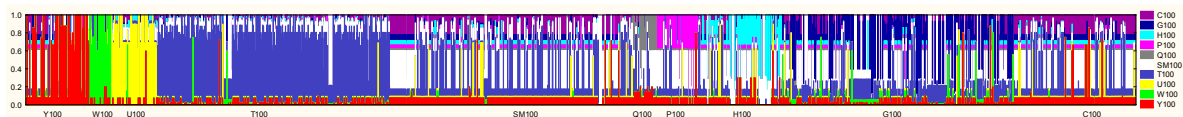


Figure 5.32 Graph of predictions by JRip classifier on the training *generalDS* set.

5.4.3.2 Allele Length Dataset

Parameters used

Scheme: weka.classifiers.rules.JRip -F 3 -N 2.0 -O 2 -S 1

Results of classification

```

Correctly Classified Instances      2054          61.6817 %
Incorrectly Classified Instances    1276          38.3183 %
Kappa statistic                    0.5361
Mean absolute error                 0.0987
Root mean squared error             0.2421
Relative absolute error             58.8129 %
Root relative squared error         83.6095 %
Total Number of Instances          3330
    
```

=== Detailed Accuracy By Class ===

| | TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area | Class |
|---------------|---------|---------|-----------|--------|-----------|----------|-------|
| | 0.399 | 0.028 | 0.457 | 0.399 | 0.426 | 0.779 | Y100 |
| | 0.47 | 0.005 | 0.66 | 0.47 | 0.549 | 0.799 | W100 |
| | 0.73 | 0.014 | 0.685 | 0.73 | 0.707 | 0.877 | U100 |
| | 0.681 | 0.068 | 0.729 | 0.681 | 0.704 | 0.871 | T100 |
| | 0.708 | 0.214 | 0.481 | 0.708 | 0.573 | 0.797 | SM100 |
| | 0.425 | 0.007 | 0.574 | 0.425 | 0.488 | 0.761 | Q100 |
| | 0.328 | 0.012 | 0.506 | 0.328 | 0.398 | 0.713 | P100 |
| | 0.638 | 0.017 | 0.749 | 0.638 | 0.689 | 0.843 | H100 |
| | 0.743 | 0.068 | 0.743 | 0.743 | 0.743 | 0.881 | G100 |
| | 0.287 | 0.032 | 0.523 | 0.287 | 0.37 | 0.75 | C100 |
| Weighted Avg. | 0.617 | 0.083 | 0.626 | 0.617 | 0.611 | 0.827 | |

| | a | b | c | d | e | f | g | h | i | j | |
|----|----|-----|-----|-----|----|----|-----|-----|-----|-------|--|
| 75 | 1 | 5 | 15 | 58 | 3 | 8 | 4 | 12 | 7 | Y100 | |
| 1 | 31 | 0 | 3 | 17 | 0 | 0 | 2 | 11 | 1 | W100 | |
| 5 | 2 | 100 | 3 | 16 | 1 | 0 | 0 | 9 | 1 | U100 | |
| 1 | 7 | 8 | 480 | 117 | 2 | 9 | 4 | 69 | 8 | T100 | |
| 13 | 1 | 9 | 75 | 517 | 8 | 4 | 16 | 32 | 55 | SM100 | |
| 6 | 0 | 3 | 6 | 19 | 31 | 0 | 1 | 6 | 1 | Q100 | |
| 10 | 1 | 4 | 7 | 36 | 2 | 41 | 10 | 4 | 10 | P100 | |
| 6 | 0 | 5 | 5 | 41 | 1 | 4 | 155 | 18 | 8 | H100 | |
| 26 | 4 | 3 | 30 | 92 | 5 | 9 | 7 | 520 | 4 | G100 | |
| 21 | 0 | 9 | 34 | 161 | 1 | 6 | 8 | 19 | 104 | C100 | |

Table 5.25 Confusion matrix for JRip classifier for *allelelengthDS*.

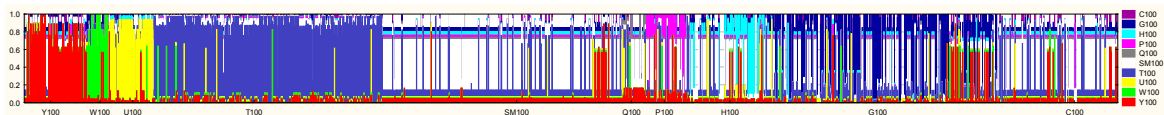


Figure 5.33 Graph of predictions by JRip classifier on the training *allelelengthDS* set.

5.4.3.3 Allele Frequency Dataset

Parameters used

```
weka.classifiers.rules.JRip -F 3 -N 2.0 -O 2 -S 1
```

Results of classification

```
Correctly Classified Instances      2053      61.6517 %
Incorrectly Classified Instances    1277      38.3483 %
Kappa statistic                    0.5363
Mean absolute error                 0.0973
Root mean squared error            0.2448
Relative absolute error             58.0066 %
Root relative squared error        84.5522 %
Total Number of Instances         3330
```

=== Detailed Accuracy By Class ===

| | TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area | Class |
|---------------|---------|---------|-----------|--------|-----------|----------|-------|
| | 0.356 | 0.023 | 0.482 | 0.356 | 0.41 | 0.765 | Y100 |
| | 0.485 | 0.011 | 0.478 | 0.485 | 0.481 | 0.776 | W100 |
| | 0.664 | 0.018 | 0.607 | 0.664 | 0.634 | 0.887 | U100 |
| | 0.691 | 0.071 | 0.724 | 0.691 | 0.707 | 0.867 | T100 |
| | 0.712 | 0.202 | 0.498 | 0.712 | 0.586 | 0.795 | SM100 |
| | 0.384 | 0.008 | 0.519 | 0.384 | 0.441 | 0.755 | Q100 |
| | 0.288 | 0.013 | 0.456 | 0.288 | 0.353 | 0.721 | P100 |
| | 0.584 | 0.019 | 0.703 | 0.584 | 0.638 | 0.824 | H100 |
| | 0.766 | 0.059 | 0.776 | 0.766 | 0.771 | 0.895 | G100 |
| | 0.314 | 0.039 | 0.496 | 0.314 | 0.384 | 0.728 | C100 |
| Weighted Avg. | 0.617 | 0.08 | 0.621 | 0.617 | 0.61 | 0.824 | |

| | a | b | c | d | e | f | g | h | i | j | |
|----|----|----|-----|-----|----|----|-----|-----|-----|-------|--|
| 67 | 2 | 6 | 15 | 51 | 7 | 6 | 5 | 19 | 10 | Y100 | |
| 1 | 32 | 0 | 6 | 12 | 0 | 2 | 1 | 10 | 2 | W100 | |
| 0 | 2 | 91 | 6 | 23 | 1 | 0 | 4 | 10 | 0 | U100 | |
| 1 | 10 | 10 | 487 | 115 | 5 | 5 | 3 | 56 | 13 | T100 | |
| 14 | 8 | 15 | 73 | 520 | 4 | 7 | 15 | 19 | 55 | SM100 | |
| 3 | 0 | 3 | 7 | 21 | 28 | 4 | 2 | 3 | 2 | Q100 | |
| 8 | 2 | 1 | 9 | 44 | 4 | 36 | 8 | 3 | 10 | P100 | |
| 2 | 1 | 9 | 12 | 39 | 0 | 7 | 142 | 18 | 13 | H100 | |
| 19 | 7 | 7 | 24 | 83 | 2 | 4 | 7 | 536 | 11 | G100 | |
| 24 | 3 | 8 | 34 | 137 | 3 | 8 | 15 | 17 | 114 | C100 | |

Table 5.26 Confusion matrix for JRip classifier for *allelefrequencyDS*.

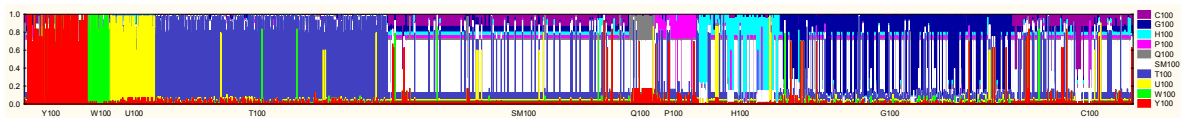


Figure 5.34 Graph of predictions by JRip classifier on the training *allelefrequencyDS* set.

5.4.3.4 Discussion of JRip results

JRip algorithm with parameters set as is described in section 5.4.3 was used to induce decision rules on given three datasets. Number of folds was set to 3, pruning of rules was allowed, minimal number of instances covered by rule was set to 2, two optimizations were performed for each set of rules, and seed number was set to 1.

On the general dataset, 52.79 % of correctly classified instances and Kappa statistic=0.4174 were reached. Average FP Rate was calculated as 0.116. Overall weighed precision equals 0.546 and F-measure equals 0.507.

The best classified class in the meaning of TP Rate is G100 (0.696), the one with the worst TP Rate is Y100 (0.096). Best FP rate was calculated for Q100 (0.004), the worst one for SM100 class (0.342). Precision equals to 0.694 was reached for G100 (F-Measure=0.695) and in opposite, Precision=0.363 (F-Measure=0.159) were reached for C100 class.

From results available in previous paragraph, we can see that the most well predicted classes are W100, U100, T100, H100, G100. Also, Y100, SM100 and C100 are not well predicted, how table 5.24 and figure 5.32 show with respect of evidence that SM100, Y100 and C100 classes have big admixture, especially C100 and Y100 are mainly classified as SM100 individuals.

For the allele length dataset, following parameters were calculated. Number of correctly classified instances is 61.68 %, Kappa statistic=0.5361, FP Rate=0.083. So, the overall results seems to be better than for general dataset. Overall precision of the model is evaluated as 0.626, F-Measure calculated as 0.611.

The best TP Rate was calculated for G100 class (0.743), the worst result was obtained for C100 class (0.287). FP Rate ranges from 0.005 (W100) to 0.214 (SM100). Class with the best precision calculated is H100 in this case (0.749, F-Measure=0.689), the worst results (Precision=0.457, F-Measure=0.426) were obtained for Y100 class.

The graph 5.33 shows big admixture of SM100 class according to results of classification to classes Q100, P100, W100, G100 and C100 as well. In this case, model is significantly influenced by most frequent class as well as SM100 breed genetic definition.

Results which were obtained for allele frequency dataset for JRip classification algorithm shows 61.65 % of correctly classified instances and Kappa statistic equals 0.5363. Overall FP Rate=0.08, Precision=0.621, F-Measure=0.61. So, overall results on all of three datasets are very similar, and there is no significant difference between all of dataset as in J48 method in previous chapter.

Best classified breed according TP Rate is G100 (0.766) as the opposite of 0.288 in P100. FP Rate ranges from 0.008 for Q100 to 0.202 in SM100. Best precision was calculated for G100 class (0.776, F-Measure=0.771), the worst one as 0.482 for Y100 class (F-Measure=0.41).

The same results could be concluded from graph 5.34 as for allele length dataset. Big admixture between SM100, Q100, P100, H100, G100 and C100 classes is displayed on the graph and fully respects results in confusion matrix for this prediction model.

Based on results, we can conclude that JRip model for allele length dataset should be recommended as the most usable one for breed discrimination task from decision rules models and datasets proved by this work as it is the most robust, with the acceptable FP Rate. Anyway, obtained results are not suitable for general purposes as 61.68 % of correctly classified instances do not fulfill needs for robust prediction on unknown samples.

5.4.4 Naive Bayes Classifier

5.4.4.1 General Dataset

Parameters used

Scheme: weka.classifiers.bayes.NaiveBayes

Results of classification

```

Correctly Classified Instances      2749      82.5526 %
Incorrectly Classified Instances    581      17.4474 %
Kappa statistic                    0.7893
Mean absolute error                0.046
Root mean squared error            0.1605
Relative absolute error             27.3989 %
Root relative squared error        55.414 %
Total Number of Instances          3330
    
```

=== Detailed Accuracy By Class ===

| | TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area | Class |
|---------------|---------|---------|-----------|--------|-----------|----------|-------|
| | 0.771 | 0.012 | 0.797 | 0.771 | 0.784 | 0.984 | Y100 |
| | 0.682 | 0.001 | 0.918 | 0.682 | 0.783 | 0.997 | W100 |
| | 0.905 | 0.007 | 0.855 | 0.905 | 0.879 | 0.98 | U100 |
| | 0.916 | 0.036 | 0.872 | 0.916 | 0.893 | 0.985 | T100 |
| | 0.892 | 0.09 | 0.736 | 0.892 | 0.807 | 0.965 | SM100 |
| | 0.397 | 0.001 | 0.879 | 0.397 | 0.547 | 0.971 | Q100 |
| | 0.408 | 0.002 | 0.879 | 0.408 | 0.557 | 0.961 | P100 |
| | 0.79 | 0.006 | 0.914 | 0.79 | 0.848 | 0.919 | H100 |
| | 0.959 | 0.022 | 0.919 | 0.959 | 0.938 | 0.992 | G100 |
| | 0.537 | 0.035 | 0.654 | 0.537 | 0.59 | 0.921 | C100 |
| Weighted Avg. | 0.826 | 0.037 | 0.828 | 0.826 | 0.818 | 0.969 | |

| a | b | c | d | e | f | g | h | i | j | |
|-----|----|-----|-----|-----|----|----|-----|-----|-----|-------|
| 145 | 0 | 0 | 7 | 17 | 0 | 1 | 1 | 3 | 14 | Y100 |
| 0 | 45 | 1 | 5 | 1 | 0 | 0 | 0 | 11 | 3 | W100 |
| 0 | 0 | 124 | 3 | 3 | 0 | 1 | 0 | 3 | 3 | U100 |
| 3 | 1 | 4 | 646 | 33 | 0 | 0 | 0 | 14 | 4 | T100 |
| 2 | 0 | 0 | 33 | 651 | 0 | 1 | 4 | 4 | 35 | SM100 |
| 8 | 1 | 4 | 4 | 10 | 29 | 1 | 2 | 3 | 11 | Q100 |
| 13 | 1 | 3 | 9 | 17 | 1 | 51 | 4 | 5 | 21 | P100 |
| 3 | 0 | 7 | 7 | 21 | 1 | 0 | 192 | 7 | 5 | H100 |
| 3 | 0 | 1 | 9 | 7 | 1 | 0 | 1 | 671 | 7 | G100 |
| 5 | 1 | 1 | 18 | 124 | 1 | 3 | 6 | 9 | 195 | C100 |

Table 5.27 Confusion matrix for Naive Bayes classifier for *generalDS*.

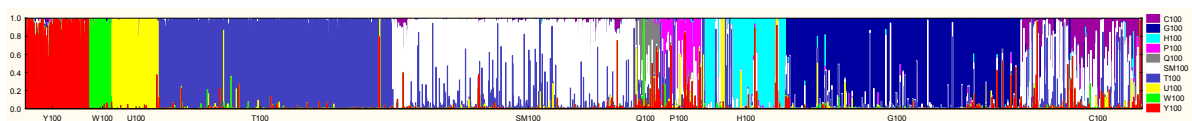


Figure 5.35 Graph of predictions by Naive Bayes classifier on the training *generalDS* set.

5.4.4.2 Allele Length Dataset

Parameters used

Scheme: weka.classifiers.bayes.NaiveBayes

Results of classification

```

Correctly Classified Instances      1869           56.1261 %
Incorrectly Classified Instances    1461           43.8739 %
Kappa statistic                    0.4748
Mean absolute error                 0.1057
Root mean squared error            0.247
Relative absolute error             63.0086 %
Root relative squared error        85.2994 %
Total Number of Instances          3330
    
```

=== Detailed Accuracy By Class ===

| | TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area | Class |
|---------------|---------|---------|-----------|--------|-----------|----------|-------|
| | 0.431 | 0.029 | 0.474 | 0.431 | 0.451 | 0.899 | Y100 |
| | 0.47 | 0.019 | 0.337 | 0.47 | 0.392 | 0.94 | W100 |
| | 0.788 | 0.03 | 0.527 | 0.788 | 0.632 | 0.951 | U100 |
| | 0.715 | 0.112 | 0.632 | 0.715 | 0.671 | 0.897 | T100 |
| | 0.536 | 0.134 | 0.528 | 0.536 | 0.532 | 0.795 | SM100 |
| | 0.274 | 0.013 | 0.328 | 0.274 | 0.299 | 0.882 | Q100 |
| | 0.16 | 0.011 | 0.364 | 0.16 | 0.222 | 0.816 | P100 |
| | 0.539 | 0.046 | 0.478 | 0.539 | 0.507 | 0.846 | H100 |
| | 0.679 | 0.089 | 0.671 | 0.679 | 0.675 | 0.898 | G100 |
| | 0.298 | 0.04 | 0.478 | 0.298 | 0.367 | 0.776 | C100 |
| Weighted Avg. | 0.561 | 0.083 | 0.553 | 0.561 | 0.552 | 0.858 | |

| | a | b | c | d | e | f | g | h | i | j | |
|----|----|-----|-----|-----|----|----|-----|-----|-----|---|-------|
| 81 | 5 | 3 | 8 | 23 | 6 | 5 | 6 | 37 | 14 | | Y100 |
| 0 | 31 | 2 | 7 | 11 | 0 | 0 | 5 | 10 | 0 | | W100 |
| 1 | 0 | 108 | 4 | 12 | 0 | 1 | 4 | 5 | 2 | | U100 |
| 13 | 16 | 7 | 504 | 80 | 11 | 9 | 9 | 43 | 13 | | T100 |
| 6 | 12 | 29 | 143 | 391 | 6 | 3 | 40 | 54 | 46 | | SM100 |
| 6 | 1 | 7 | 9 | 9 | 20 | 2 | 4 | 11 | 4 | | Q100 |
| 10 | 0 | 6 | 12 | 28 | 9 | 20 | 15 | 18 | 7 | | P100 |
| 6 | 7 | 11 | 15 | 28 | 3 | 5 | 131 | 20 | 17 | | H100 |
| 29 | 15 | 19 | 58 | 57 | 4 | 4 | 24 | 475 | 15 | | G100 |
| 19 | 5 | 13 | 38 | 101 | 2 | 6 | 36 | 35 | 108 | | C100 |

Table 5.28 Confusion matrix for Naive Bayes classifier for *allelelengthDS*.

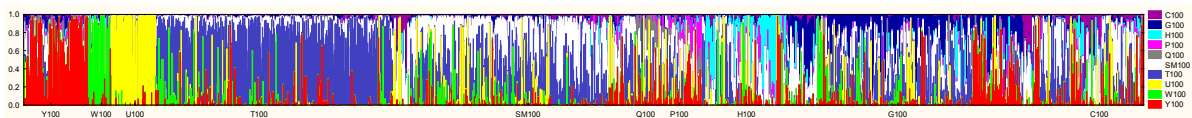


Figure 5.36 Graph of predictions by Naive Bayes classifier on the training *allelelengthDS* set.

5.4.4.3 Allele Frequency Dataset

Parameters used

Scheme: `weka.classifiers.bayes.NaiveBayes`

Results of classification

```

Correctly Classified Instances      1944      58.3784 %
Incorrectly Classified Instances    1386      41.6216 %
Kappa statistic                    0.498
Mean absolute error                 0.1011
Root mean squared error             0.2405
Relative absolute error             60.2626 %
Root relative squared error         83.052 %
Total Number of Instances          3330
    
```

=== Detailed Accuracy By Class ===

| | TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area | Class |
|---------------|---------|---------|-----------|--------|-----------|----------|-------|
| | 0.399 | 0.021 | 0.528 | 0.399 | 0.455 | 0.905 | Y100 |
| | 0.348 | 0.009 | 0.426 | 0.348 | 0.383 | 0.898 | W100 |
| | 0.774 | 0.022 | 0.602 | 0.774 | 0.677 | 0.961 | U100 |
| | 0.746 | 0.125 | 0.615 | 0.746 | 0.674 | 0.901 | T100 |
| | 0.614 | 0.138 | 0.554 | 0.614 | 0.583 | 0.826 | SM100 |
| | 0.288 | 0.011 | 0.368 | 0.288 | 0.323 | 0.889 | Q100 |
| | 0.208 | 0.011 | 0.426 | 0.208 | 0.28 | 0.854 | P100 |
| | 0.535 | 0.031 | 0.575 | 0.535 | 0.554 | 0.851 | H100 |
| | 0.7 | 0.08 | 0.699 | 0.7 | 0.7 | 0.902 | G100 |
| | 0.273 | 0.051 | 0.396 | 0.273 | 0.323 | 0.757 | C100 |
| Weighted Avg. | 0.584 | 0.085 | 0.571 | 0.584 | 0.572 | 0.866 | |

| a | b | c | d | e | f | g | h | i | j | |
|----|----|-----|-----|-----|----|----|-----|-----|----|-------|
| 75 | 2 | 4 | 7 | 27 | 2 | 7 | 10 | 31 | 23 | Y100 |
| 3 | 23 | 1 | 12 | 10 | 0 | 1 | 2 | 14 | 0 | W100 |
| 1 | 0 | 106 | 7 | 11 | 0 | 2 | 2 | 2 | 6 | U100 |
| 2 | 7 | 8 | 526 | 84 | 9 | 3 | 3 | 50 | 13 | T100 |
| 14 | 7 | 16 | 144 | 448 | 2 | 2 | 15 | 39 | 43 | SM100 |
| 2 | 1 | 6 | 9 | 8 | 21 | 4 | 4 | 13 | 5 | Q100 |
| 11 | 0 | 4 | 16 | 21 | 7 | 26 | 11 | 10 | 19 | P100 |
| 4 | 3 | 14 | 19 | 21 | 0 | 11 | 130 | 21 | 20 | H100 |
| 17 | 6 | 8 | 70 | 57 | 11 | 0 | 19 | 490 | 22 | G100 |
| 13 | 5 | 9 | 45 | 121 | 5 | 5 | 30 | 31 | 99 | C100 |

Table 5.29 Confusion matrix for Naive Bayes classifier for *allelefrequencyDS*.

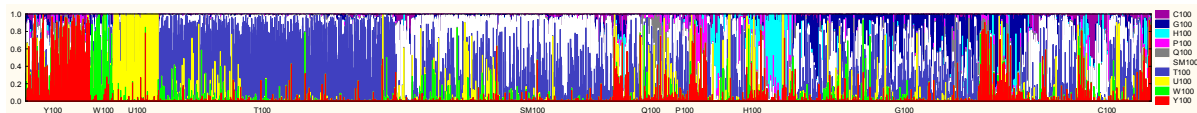


Figure 5.37 Graph of predictions by Naive Bayes classifier on the training *allelefrequencyDS* set.

5.4.4.4 Discussion of Naive Bayes results

Naive Bayes algorithm was examined for breed discrimination task on all of three datasets as well. For the general dataset, 82.55 % of correctly classified instances and Kappa statistic=0.7893 parameters were obtained. Overall FP Rate calculated as weighted mean across all of classes equals 0.037, overall precision is 0.828 and F-Measure=0.818.

Best precision was reached for G100 class (0.919, F-Measure=0.938), the worst one for C100 class (0.654, F-measure=0.59). FP Rate ranges from 0.001 (Q100, W100) to 0.09 (SM100). TP Rate ranges from 0.959 (G100) to 0.397 (Q100).

In accordance with results described above, figure 5.35 shows, that there is admixture of class probabilities present mainly between SM100 and C100 classes, and G100 and SM100 classes as well. Naive Bayes classifier performed on general dataset has one of the best results reached for breed classification task among all of tested algorithms.

Usability of Naive Bayes classifier was not proved by the results for allele length and allele frequency datasets. For allele length dataset, 56.13 % of instances were classified correctly by model, Kappa statistic equals 0.4748 and average FP Rate is 0.083, overall precision is 0.5536 and F-Measure equals 0.552. Similar results were obtained for classes themselves. The best TP Rate was obtained for G100 class (0.679), the worst for P100 one (0.16). FT Rate ranges from 0.011 for P100 to 0.134 for SM100 class. It is quite interesting result, TP Rate and FP Rate calculated for P100 class. Small TP Rate value shows that P100 individuals can not be classified correctly (they are classified as another breed), but on the other hand small FP Rate shows, there are not misclassification from another breeds individuals. So, we can say that P100 can not create valid breed but is different that the others under the Naive Bayes classification on allele length dataset.

As graph 5.36 shows, Naive Bayes classifier on allele length dataset is not suitable for good breed definition as in SM100, Q100, P100, H100, G100 and C100 groups, there is big portion of probabilities of other breeds equally spread for all of noticed breeds.

In the section 5.4.4.3 there are results obtained for Naive Bayes classifier on allele frequency dataset. 58.38 % of instances were classified correctly by this model. Kappa statistic for the whole 10 fold cross validation equals 0.498. Overall weighted FP Rate is 0.085, precision equals 0.571 and F-Measure=0.572.

The best TP Rate=0.774 was reached for U100 class, the worst one (0.273) for C100 class. FP Rate ranges from 0.009 (W100) to 0.138 in SM100. The best precision was reached for G100 class (0.699, F-Measure=0.7), the worst one for Q100 class (0.368, F-Measure=0.323).

Graph of predictions on the training set shows better results than in case of allele length dataset, anyway results for SM100 and C100 breeds are not satisfactory. Also, big mixture of predicted probabilities is evident for T100 (however this class is predicted clearly) and SM100 and G100 classes.

5.4.5 Bayes Net

5.4.5.1 General Dataset

Parameters used

```
Scheme:          weka.classifiers.bayes.BayesNet -D -Q weka.classifiers.bayes.net.search.local.K2 -- -P 1 -S BAYES -E weka.classifiers.bayes.net.estimate.SimpleEstimator -- -A 0.5
```

Results of classification

```
Correctly Classified Instances      2786          83.6637 %
Incorrectly Classified Instances     544          16.3363 %
Kappa statistic                     0.8035
Mean absolute error                 0.043
Root mean squared error             0.1558
Relative absolute error             25.6494 %
Root relative squared error         53.7972 %
Total Number of Instances          3330
```

=== Detailed Accuracy By Class ===

| | TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area | Class |
|---------------|---------|---------|-----------|--------|-----------|----------|-------|
| | 0.809 | 0.011 | 0.822 | 0.809 | 0.815 | 0.985 | Y100 |
| | 0.758 | 0.002 | 0.909 | 0.758 | 0.826 | 0.998 | W100 |
| | 0.92 | 0.006 | 0.869 | 0.92 | 0.894 | 0.982 | U100 |
| | 0.915 | 0.032 | 0.885 | 0.915 | 0.9 | 0.986 | T100 |
| | 0.87 | 0.077 | 0.761 | 0.87 | 0.812 | 0.966 | SM100 |
| | 0.562 | 0.002 | 0.891 | 0.562 | 0.689 | 0.974 | Q100 |
| | 0.48 | 0.004 | 0.811 | 0.48 | 0.603 | 0.962 | P100 |
| | 0.802 | 0.006 | 0.915 | 0.802 | 0.855 | 0.921 | H100 |
| | 0.96 | 0.019 | 0.929 | 0.96 | 0.944 | 0.993 | G100 |
| | 0.579 | 0.039 | 0.644 | 0.579 | 0.61 | 0.925 | C100 |
| Weighted Avg. | 0.837 | 0.033 | 0.837 | 0.837 | 0.833 | 0.97 | |

| a | b | c | d | e | f | g | h | i | j | |
|-----|----|-----|-----|-----|----|----|-----|-----|-----|-------|
| 152 | 0 | 0 | 6 | 12 | 0 | 1 | 1 | 2 | 14 | Y100 |
| 0 | 50 | 1 | 4 | 0 | 0 | 0 | 0 | 9 | 2 | W100 |
| 0 | 0 | 126 | 3 | 3 | 0 | 1 | 0 | 1 | 3 | U100 |
| 3 | 1 | 4 | 645 | 31 | 0 | 1 | 1 | 13 | 6 | T100 |
| 4 | 0 | 0 | 32 | 635 | 0 | 3 | 4 | 4 | 48 | SM100 |
| 5 | 1 | 3 | 0 | 7 | 41 | 1 | 1 | 1 | 13 | Q100 |
| 10 | 1 | 2 | 9 | 13 | 2 | 60 | 4 | 5 | 19 | P100 |
| 3 | 0 | 7 | 7 | 17 | 1 | 1 | 195 | 7 | 5 | H100 |
| 2 | 1 | 1 | 8 | 7 | 1 | 1 | 1 | 672 | 6 | G100 |
| 6 | 1 | 1 | 15 | 109 | 1 | 5 | 6 | 9 | 210 | C100 |

Table 5.30 Confusion matrix for Bayes Net classifier for *generalDS*.

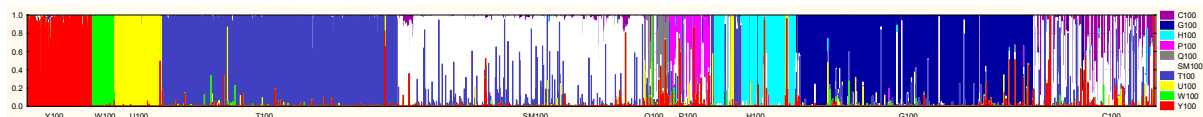


Figure 5.38 Graph of predictions by Bayes Net classifier on the training *generalDS* set.

5.4.5.2 Allele Length Dataset

Parameters used

```
Scheme:          weka.classifiers.bayes.BayesNet -D -Q weka.classifiers.bayes.net.search.local.K2 -- -P 1 -S BAYES -E weka.classifiers.bayes.net.estimate.SimpleEstimator -- -A 0.5
```

Results of classification

```
Correctly Classified Instances      2809          84.3544 %
Incorrectly Classified Instances     521          15.6456 %
Kappa statistic                     0.8128
Mean absolute error                  0.0399
Root mean squared error              0.1523
Relative absolute error              23.7714 %
Root relative squared error          52.6101 %
Total Number of Instances           3330
```

=== Detailed Accuracy By Class ===

| | TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area | Class |
|---------------|---------|---------|-----------|--------|-----------|----------|-------|
| | 0.809 | 0.013 | 0.792 | 0.809 | 0.8 | 0.984 | Y100 |
| | 0.924 | 0.002 | 0.897 | 0.924 | 0.91 | 0.999 | W100 |
| | 0.905 | 0.006 | 0.867 | 0.905 | 0.886 | 0.979 | U100 |
| | 0.919 | 0.032 | 0.886 | 0.919 | 0.903 | 0.985 | T100 |
| | 0.829 | 0.064 | 0.784 | 0.829 | 0.806 | 0.965 | SM100 |
| | 0.658 | 0.006 | 0.716 | 0.658 | 0.686 | 0.972 | Q100 |
| | 0.632 | 0.007 | 0.782 | 0.632 | 0.699 | 0.971 | P100 |
| | 0.798 | 0.008 | 0.886 | 0.798 | 0.84 | 0.921 | H100 |
| | 0.954 | 0.013 | 0.952 | 0.954 | 0.953 | 0.994 | G100 |
| | 0.634 | 0.035 | 0.687 | 0.634 | 0.659 | 0.936 | C100 |
| Weighted Avg. | 0.844 | 0.029 | 0.842 | 0.844 | 0.842 | 0.972 | |

| a | b | c | d | e | f | g | h | i | j | |
|-----|----|-----|-----|-----|----|----|-----|-----|-----|-------|
| 152 | 2 | 2 | 5 | 14 | 1 | 2 | 1 | 1 | 8 | Y100 |
| 0 | 61 | 0 | 2 | 1 | 0 | 0 | 0 | 2 | 0 | W100 |
| 2 | 0 | 124 | 2 | 4 | 0 | 1 | 2 | 1 | 1 | U100 |
| 4 | 2 | 4 | 648 | 29 | 2 | 1 | 0 | 11 | 4 | T100 |
| 6 | 0 | 1 | 34 | 605 | 9 | 3 | 6 | 4 | 62 | SM100 |
| 4 | 1 | 3 | 4 | 3 | 48 | 2 | 1 | 0 | 7 | Q100 |
| 6 | 0 | 2 | 6 | 8 | 1 | 79 | 8 | 3 | 12 | P100 |
| 4 | 1 | 6 | 7 | 13 | 2 | 4 | 194 | 5 | 7 | H100 |
| 6 | 0 | 1 | 7 | 10 | 0 | 3 | 1 | 668 | 4 | G100 |
| 8 | 1 | 0 | 16 | 85 | 4 | 6 | 6 | 7 | 230 | C100 |

Table 5.31 Confusion matrix for Bayes Net classifier for *allelelengthDS*.

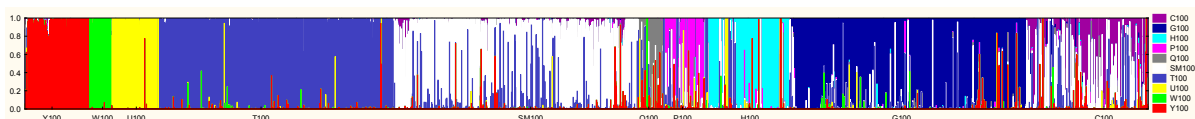


Figure 5.39 Graph of predictions by Bayes Net classifier on the training *allelelengthDS* set.

5.4.5.3 Allele Frequency Dataset

Parameters used

```
Scheme:          weka.classifiers.bayes.BayesNet -D -Q weka.classifiers.bayes.net.search.local.K2 -- -P 1 -S BAYES -E weka.classifiers.bayes.net.estimate.BMAEstimator -- -A 0.5
```

Results of classification

```
Correctly Classified Instances      2824          84.8048 %
Incorrectly Classified Instances    506           15.1952 %
Kappa statistic                    0.8181
Mean absolute error                 0.0392
Root mean squared error             0.1509
Relative absolute error             23.3925 %
Root relative squared error         52.1192 %
Total Number of Instances          3330
```

=== Detailed Accuracy By Class ===

| | TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area | Class |
|---------------|---------|---------|-----------|--------|-----------|----------|-------|
| | 0.814 | 0.011 | 0.814 | 0.814 | 0.814 | 0.987 | Y100 |
| | 0.894 | 0.002 | 0.908 | 0.894 | 0.901 | 0.999 | W100 |
| | 0.912 | 0.005 | 0.893 | 0.912 | 0.903 | 0.976 | U100 |
| | 0.928 | 0.033 | 0.883 | 0.928 | 0.905 | 0.986 | T100 |
| | 0.837 | 0.057 | 0.805 | 0.837 | 0.821 | 0.967 | SM100 |
| | 0.685 | 0.005 | 0.769 | 0.685 | 0.725 | 0.975 | Q100 |
| | 0.704 | 0.006 | 0.815 | 0.704 | 0.755 | 0.972 | P100 |
| | 0.79 | 0.007 | 0.893 | 0.79 | 0.838 | 0.921 | H100 |
| | 0.951 | 0.016 | 0.941 | 0.951 | 0.946 | 0.994 | G100 |
| | 0.623 | 0.039 | 0.663 | 0.623 | 0.642 | 0.934 | C100 |
| Weighted Avg. | 0.848 | 0.029 | 0.847 | 0.848 | 0.847 | 0.972 | |

| a | b | c | d | e | f | g | h | i | j | |
|-----|----|-----|-----|-----|----|----|-----|-----|-----|-------|
| 153 | 0 | 0 | 7 | 8 | 3 | 2 | 1 | 3 | 11 | Y100 |
| 0 | 59 | 0 | 3 | 0 | 0 | 0 | 0 | 4 | 0 | W100 |
| 1 | 0 | 125 | 2 | 4 | 0 | 0 | 1 | 2 | 2 | U100 |
| 2 | 1 | 3 | 654 | 24 | 2 | 2 | 1 | 13 | 3 | T100 |
| 4 | 0 | 0 | 31 | 611 | 3 | 3 | 3 | 4 | 71 | SM100 |
| 6 | 1 | 3 | 3 | 1 | 50 | 1 | 1 | 0 | 7 | Q100 |
| 3 | 0 | 1 | 6 | 7 | 1 | 88 | 8 | 2 | 9 | P100 |
| 5 | 1 | 6 | 7 | 14 | 2 | 4 | 192 | 5 | 7 | H100 |
| 3 | 2 | 2 | 9 | 8 | 0 | 4 | 1 | 666 | 5 | G100 |
| 11 | 1 | 0 | 19 | 82 | 4 | 4 | 7 | 9 | 226 | C100 |

Table 5.32 Confusion matrix for Bayes Net classifier for *allelefrequencyDS*.

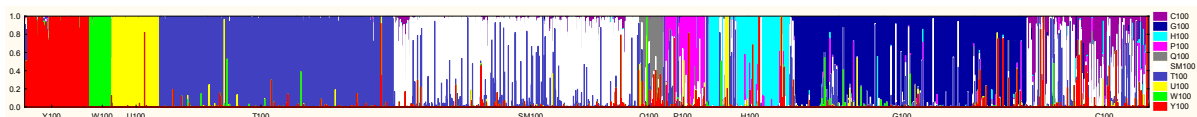


Figure 5.40 Graph of predictions by Bayes Net classifier on the training *allelefrequencyDS* set.

5.4.5.4 Discussion of Bayes Net results

Bayes Net classifier was examined in the same way as was previously described for other classifiers. Parameters set were: estimator - Simple Estimator with $\alpha=0.5$, search algorithm as K2 (hill climbing one) with maximum 1 parent and initial state as Naive Bayes classifier. On the general dataset, we reached 83.66 % of correctly classified instances and Kappa statistic=0.8035. Overall calculated FP Rate equals 0.033, Precision=0.837 and F-Measure 0.833.

The best TP Rate was reached for G100 class (0.960), the worst one for P100 class (0.480). FP ranges from 0.002 (Q100) to 0.077 (SM100). Best precision was calculated for G100 class (0.929, F-Measure=0.944) and the worst one for C100 (0.644, F-Measure=0.610). When we examined confusion matrix for this classifier (table 5.30) we can observe the main misclassified individuals in C100 and SM100 classes. 48 individuals of SM100 breed were classified as C100 ones, and 109 individuals from C100 were classified as SM100 respectively. In this case, model is not able to divide these two breeds effectively, what should reflect breeding strategy of Czech Fleckvieh and similar genetic basis of both breeds.

On the figure 5.38 there are displayed results of predictions of Bayes Net on the training set. Graph shows clearly very good level of prediction as well as described similarities between C100 and SM100 groups of individuals.

Results available for Bayes Net classifier ran on the allele length dataset show very good classification power as Bayes Net on general dataset has. Number of correctly classified instances is 84.35 % and Kappa statistic equals to 0.8128, FP Rate=0.029, Precision=0.842 and F-Measure=0.842.

The best classified class according to TP Rate is G100 with TP Rate=0.954, the worst classified is P100 with 0.632 TP Rate. FP Rate obtained ranges from 0.02 for W100 to 0.064 in SM100 group of individuals. Best precision was reached for G100 breed (0.952, F-Measure=0.953), the lowest one for C100 (0.687, F-Measure=0.659).

Also, graph of predictions on dataset shows very clearly predicted probabilities of individuals breed (with exception of C100 breed).

Bayes Net classifier reached on allele frequency dataset 84.81 % of correctly classified instances what is the best results all over the classifiers and datasets combinations. Kappa statistic equals 0.8181 in this case. Overall calculated FP Rate is 0.029, Precision=0.847 and F-Measure=0.847. The best TP Rate (0.951) was reached for G100 class, the worst one (0.623) for C100. FP Rate ranges from 0.002 in W100 to 0.057 in SM100. Best precision was calculated for G100 class (0.941, F-Measure=0.946), the worst one for C100 as 0.663, F-Measure=0.642. As results obtained for Bayes Net classifier for allele frequency dataset are the best obtained from all of classifiers, graph of predictions (figure 5.40) gives one of the most reliable picture of classification methods limits in cattle breeds calculated on given dataset. As well, when the model is 10 fold cross validated, the similar results can be expected on unknown samples as well. Surprisingly, results are comparable to horse breed discrimination ones presented in (Burocziova, Riha; 2009). As the horses breeding strategies as well as parentage and pedigree control is completely different (more strict), we can expect better results of discrimination as well. On the other hand, Bayes Net classifier result show, that we can find good performed classifier suitable for cattle breeds discrimination as well. Additionally, results obtained for Naive Bayes Classifier, Bayes Net allows to reach very good results of classification on all of three datasets (and the best results for allele frequency dataset).

5.4.6 IB1

5.4.6.1 General Dataset

Parameters used

Scheme: `weka.classifiers.lazy.IB1`

Results of classification

```

Correctly Classified Instances      1999      60.03 %
Incorrectly Classified Instances    1331      39.97 %
Kappa statistic                    0.5155
Mean absolute error                 0.0799
Root mean squared error             0.2827
Relative absolute error              47.6532 %
Root relative squared error          97.6367 %
Total Number of Instances          3330
    
```

=== Detailed Accuracy By Class ===

| | TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area | Class |
|---------------|---------|---------|-----------|--------|-----------|----------|-------|
| | 0.33 | 0.021 | 0.488 | 0.33 | 0.394 | 0.655 | Y100 |
| | 0.5 | 0.006 | 0.647 | 0.5 | 0.564 | 0.747 | W100 |
| | 0.708 | 0.012 | 0.719 | 0.708 | 0.713 | 0.848 | U100 |
| | 0.748 | 0.116 | 0.634 | 0.748 | 0.686 | 0.816 | T100 |
| | 0.659 | 0.149 | 0.554 | 0.659 | 0.602 | 0.755 | SM100 |
| | 0.274 | 0.007 | 0.455 | 0.274 | 0.342 | 0.633 | Q100 |
| | 0.192 | 0.013 | 0.358 | 0.192 | 0.25 | 0.589 | P100 |
| | 0.444 | 0.027 | 0.568 | 0.444 | 0.499 | 0.709 | H100 |
| | 0.756 | 0.079 | 0.718 | 0.756 | 0.736 | 0.838 | G100 |
| | 0.325 | 0.054 | 0.423 | 0.325 | 0.368 | 0.635 | C100 |
| Weighted Avg. | 0.6 | 0.084 | 0.587 | 0.6 | 0.588 | 0.758 | |

| a | b | c | d | e | f | g | h | i | j | |
|----|----|----|-----|-----|----|----|-----|-----|-----|-------|
| 62 | 2 | 4 | 34 | 35 | 3 | 4 | 13 | 14 | 17 | Y100 |
| 1 | 33 | 0 | 10 | 1 | 0 | 0 | 1 | 19 | 1 | W100 |
| 1 | 1 | 97 | 7 | 13 | 1 | 1 | 7 | 5 | 4 | U100 |
| 6 | 2 | 6 | 527 | 86 | 2 | 2 | 5 | 45 | 24 | T100 |
| 6 | 1 | 5 | 97 | 481 | 5 | 11 | 18 | 39 | 67 | SM100 |
| 2 | 1 | 2 | 3 | 25 | 20 | 5 | 1 | 8 | 6 | Q100 |
| 13 | 0 | 3 | 20 | 25 | 4 | 24 | 4 | 20 | 12 | P100 |
| 7 | 3 | 6 | 27 | 39 | 4 | 7 | 108 | 27 | 15 | H100 |
| 15 | 8 | 6 | 52 | 48 | 3 | 6 | 18 | 529 | 15 | G100 |
| 14 | 0 | 6 | 54 | 116 | 2 | 7 | 15 | 31 | 118 | C100 |

Table 5.33 Confusion matrix for IB1 classifier for *generalDS*.

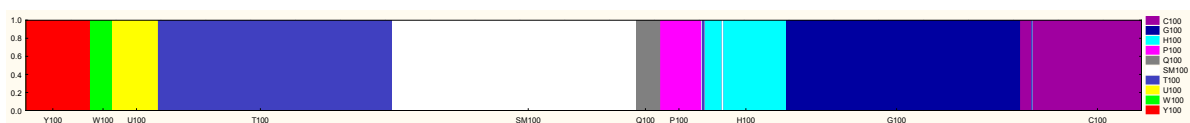


Figure 5.41 Graph of predictions by IB1 classifier on the training *generalDS* set.

5.4.6.2 Allele Length Dataset

Parameters used

Scheme: weka.classifiers.lazy.IB1

Results of classification

```

Correctly Classified Instances      1588           47.6877 %
Incorrectly Classified Instances    1742           52.3123 %
Kappa statistic                    0.3669
Mean absolute error                 0.1046
Root mean squared error            0.3235
Relative absolute error             62.368 %
Root relative squared error        111.6987 %
Total Number of Instances         3330
    
```

=== Detailed Accuracy By Class ===

| | TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area | Class |
|---------------|---------|---------|-----------|--------|-----------|----------|-------|
| | 0.346 | 0.028 | 0.428 | 0.346 | 0.382 | 0.659 | Y100 |
| | 0.152 | 0.008 | 0.27 | 0.152 | 0.194 | 0.572 | W100 |
| | 0.672 | 0.012 | 0.702 | 0.672 | 0.687 | 0.83 | U100 |
| | 0.603 | 0.142 | 0.533 | 0.603 | 0.566 | 0.73 | T100 |
| | 0.508 | 0.179 | 0.444 | 0.508 | 0.474 | 0.665 | SM100 |
| | 0.096 | 0.007 | 0.241 | 0.096 | 0.137 | 0.545 | Q100 |
| | 0.136 | 0.024 | 0.183 | 0.136 | 0.156 | 0.556 | P100 |
| | 0.317 | 0.027 | 0.478 | 0.317 | 0.381 | 0.645 | H100 |
| | 0.611 | 0.124 | 0.567 | 0.611 | 0.588 | 0.744 | G100 |
| | 0.264 | 0.082 | 0.284 | 0.264 | 0.274 | 0.591 | C100 |
| Weighted Avg. | 0.477 | 0.11 | 0.466 | 0.477 | 0.467 | 0.684 | |

| | a | b | c | d | e | f | g | h | i | j | |
|----|----|----|-----|-----|---|----|----|-----|----|----|-------|
| 65 | 1 | 2 | 20 | 22 | 4 | 9 | 5 | 42 | 18 | 18 | Y100 |
| 2 | 10 | 0 | 18 | 10 | 2 | 0 | 3 | 18 | 3 | 3 | W100 |
| 3 | 1 | 92 | 11 | 15 | 3 | 0 | 3 | 5 | 4 | 4 | U100 |
| 11 | 8 | 7 | 425 | 121 | 4 | 11 | 16 | 63 | 39 | 39 | T100 |
| 15 | 2 | 9 | 125 | 371 | 1 | 11 | 19 | 85 | 92 | 92 | SM100 |
| 4 | 0 | 6 | 14 | 15 | 7 | 9 | 1 | 11 | 6 | 6 | Q100 |
| 11 | 0 | 1 | 29 | 24 | 1 | 17 | 6 | 20 | 16 | 16 | P100 |
| 4 | 5 | 10 | 24 | 45 | 2 | 12 | 77 | 38 | 26 | 26 | H100 |
| 21 | 6 | 1 | 70 | 98 | 2 | 15 | 21 | 428 | 38 | 38 | G100 |
| 16 | 4 | 3 | 62 | 115 | 3 | 9 | 10 | 45 | 96 | 96 | C100 |

Table 5.34 Confusion matrix for IB1 classifier for *allelelengthtDS*.

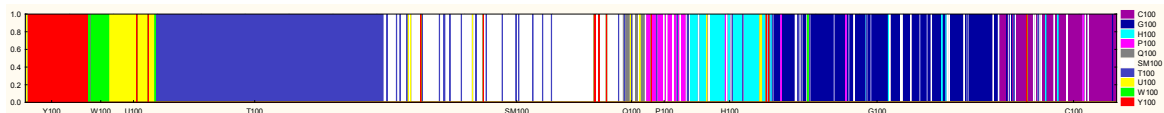


Figure 5.42 Graph of predictions by IB1 classifier on the training *allelelengthtDS* set.

5.4.6.3 Allele Frequency Dataset

Parameters used

Scheme: weka.classifiers.lazy.IB1

Results of classification

```

Correctly Classified Instances      1458           43.7838 %
Incorrectly Classified Instances    1872           56.2162 %
Kappa statistic                    0.3183
Mean absolute error                 0.1124
Root mean squared error            0.3353
Relative absolute error            67.0223 %
Root relative squared error        115.7916 %
Total Number of Instances          3330
    
```

=== Detailed Accuracy By Class ===

| | TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area | Class |
|---------------|---------|---------|-----------|--------|-----------|----------|-------|
| | 0.191 | 0.025 | 0.31 | 0.191 | 0.237 | 0.583 | Y100 |
| | 0.227 | 0.016 | 0.227 | 0.227 | 0.227 | 0.606 | W100 |
| | 0.65 | 0.013 | 0.685 | 0.65 | 0.667 | 0.818 | U100 |
| | 0.55 | 0.166 | 0.47 | 0.55 | 0.507 | 0.692 | T100 |
| | 0.46 | 0.146 | 0.469 | 0.46 | 0.465 | 0.657 | SM100 |
| | 0.027 | 0.005 | 0.111 | 0.027 | 0.044 | 0.511 | Q100 |
| | 0.144 | 0.02 | 0.22 | 0.144 | 0.174 | 0.562 | P100 |
| | 0.263 | 0.034 | 0.376 | 0.263 | 0.31 | 0.615 | H100 |
| | 0.611 | 0.181 | 0.473 | 0.611 | 0.534 | 0.715 | G100 |
| | 0.226 | 0.074 | 0.271 | 0.226 | 0.246 | 0.576 | C100 |
| Weighted Avg. | 0.438 | 0.119 | 0.42 | 0.438 | 0.424 | 0.659 | |

| | a | b | c | d | e | f | g | h | i | j | |
|--|----|----|----|-----|-----|---|----|----|-----|----|-------|
| | 36 | 4 | 2 | 25 | 20 | 4 | 10 | 15 | 57 | 15 | Y100 |
| | 1 | 15 | 0 | 20 | 6 | 0 | 1 | 1 | 19 | 3 | W100 |
| | 1 | 1 | 89 | 13 | 7 | 1 | 2 | 7 | 11 | 5 | U100 |
| | 14 | 11 | 5 | 388 | 114 | 3 | 3 | 14 | 100 | 53 | T100 |
| | 10 | 9 | 9 | 149 | 336 | 1 | 15 | 26 | 99 | 76 | SM100 |
| | 4 | 1 | 6 | 13 | 10 | 2 | 3 | 2 | 29 | 3 | Q100 |
| | 10 | 0 | 4 | 16 | 26 | 2 | 18 | 12 | 29 | 8 | P100 |
| | 6 | 8 | 7 | 28 | 47 | 2 | 10 | 64 | 55 | 16 | H100 |
| | 24 | 14 | 4 | 91 | 65 | 3 | 12 | 17 | 428 | 42 | G100 |
| | 10 | 3 | 4 | 82 | 85 | 0 | 8 | 12 | 77 | 82 | C100 |

Table 5.35 Confusion matrix for IB1 classifier for *allelefrequencyDS*.

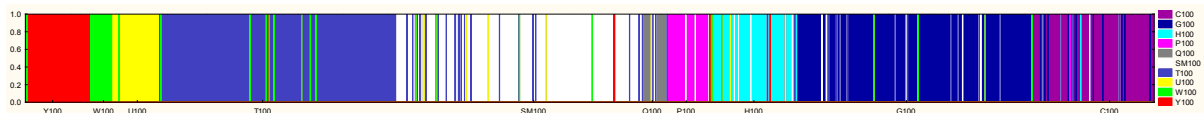


Figure 5.43 Graph of predictions by IB1 classifier on the training *allelefrequencyDS* set.

5.4.6.4 Discussion of IB1 results

Percentage of correctly classified instances for IB1 classifier from Weka-3-6-6 obtained for general dataset was 60.03 % and Kappa statistic=0.5155. Average FP Rate is set on 0.084, Precision=0.587 and F-Measure=0.588.

The best TP Rate among all of classes was reached for G100 (0.756), the worst one for P100 (0.192) class. FP Rate ranges from 0.006 for W100 to 0.149 for SM100 breed. Best precision was reached for U100 class (0.719, F-Measure=0.713), 0.358 (F-Measure=0.25) as the worst result for P100 was obtained.

As TP Rate for all of classes individually as well as for the whole dataset is quite low, we can see quite a lot of misclassified instances in confusion matrix of cross validation (table 5.33).

Otherwise, results of 10 fold cross validation are not good enough, graph of predictions on training set seems to offer very good results (figure 5.41). Anyway, it shows typical problem of lazy classifiers with their robustness. On the training set, we can always obtain results with precision near 1.0, when results are validated by cross validation they are usually much worse. Only when huge datasets are provided with repetitions, the results obtained from both 10 fold cross validation and validation on the whole dataset becomes similar. In our case, it is evident that given database is not big enough to build a robust model based on IB1 classifier. In fact described, we also can obtain better results on more frequent classes than less frequent ones, besides less frequent ones could be better specified (uniform) than more frequent ones.

In case of allele length dataset, where information about individual is doubled, the problem is more evident also for the whole dataset evaluation where misclassification is present as well (figure 5.42).

Correctly classified instances in this case represent 47.69 %, Kappa statistic is 0.3669, overall FP Rate is 0.11, Precision equals 0.466 and F-Measure=0.467.

The best results among classes for TP Rate was obtained for U100 (0.672), the worst one was obtained for Q100 (0.096), what corresponds with less frequency of Q100 individuals in dataset, so classifier can not create sufficient database which can classify unknown instances properly. Best precision was reached for U100 class as well (0.702, F-Measure=0.687), the worst one for P100 class (0.183, F-Measure=0.156).

The worst results were obtained for IB1 classifier and allele frequency dataset, where doubled information for individual is sorted additionally. In this case, it seems that database supplied is significantly not big enough to build good classification model by IB1 algorithm. Correctly classified instances represent only 43.78 %, Kappa statistic is 0.3183, overall FP Rate 0.119, Precision=0.42 and F-Measure equals 0.424.

U100 class has the best TP Rate (0.650) and Q100 represents the one with lowest TP Rate value (0.027). FP Rate ranges from 0.005 in Q100 (so, there are minimum instances misclassified as Q100) otherwise the most of Q100 instances are misclassified as the others (table 5.35) to 0.181 in G100. Best precision was reached for U100 class (0.685, F-Measure=0.667), the worst one for Q100 (0.111, F-Measure=0.044).

Especially, there are a lot of false negative results and false positives in all of three datasets results, so the classifier tends to misclassify instances in the way of badly voted probabilities. We need to note at this place that IB1 sets probabilities to 0 or 1 only depending on the voting performed by 1-NN algorithm. In this case, IB1 classifier can not be recommended as the one which is suitable for breed discrimination in cattle.

5.4.7 IB5

5.4.7.1 General Dataset

Parameters used

```
Scheme:          weka.classifiers.lazy.IBk -K 5 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""
```

Results of classification

```
Correctly Classified Instances      2177          65.3754 %
Incorrectly Classified Instances    1153          34.6246 %
Kappa statistic                    0.5701
Mean absolute error                 0.1018
Root mean squared error             0.2209
Relative absolute error             60.6705 %
Root relative squared error         76.2933 %
Total Number of Instances          3330
```

=== Detailed Accuracy By Class ===

| | TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area | Class |
|---------------|---------|---------|-----------|--------|-----------|----------|-------|
| | 0.287 | 0.007 | 0.701 | 0.287 | 0.408 | 0.916 | Y100 |
| | 0.288 | 0 | 0.95 | 0.288 | 0.442 | 0.963 | W100 |
| | 0.766 | 0.009 | 0.778 | 0.766 | 0.772 | 0.975 | U100 |
| | 0.901 | 0.142 | 0.631 | 0.901 | 0.742 | 0.951 | T100 |
| | 0.838 | 0.195 | 0.547 | 0.838 | 0.662 | 0.907 | SM100 |
| | 0.014 | 0 | 0.5 | 0.014 | 0.027 | 0.829 | Q100 |
| | 0.048 | 0.001 | 0.6 | 0.048 | 0.089 | 0.795 | P100 |
| | 0.428 | 0.004 | 0.897 | 0.428 | 0.579 | 0.897 | H100 |
| | 0.847 | 0.07 | 0.764 | 0.847 | 0.804 | 0.962 | G100 |
| | 0.132 | 0.007 | 0.706 | 0.132 | 0.223 | 0.8 | C100 |
| Weighted Avg. | 0.654 | 0.089 | 0.68 | 0.654 | 0.605 | 0.914 | |

| a | b | c | d | e | f | g | h | i | j | |
|----|----|-----|-----|-----|---|---|-----|-----|----|-------|
| 54 | 0 | 2 | 38 | 60 | 0 | 0 | 2 | 28 | 4 | Y100 |
| 2 | 19 | 1 | 15 | 8 | 0 | 0 | 1 | 20 | 0 | W100 |
| 1 | 0 | 105 | 7 | 14 | 0 | 0 | 0 | 9 | 1 | U100 |
| 5 | 0 | 5 | 635 | 42 | 0 | 0 | 0 | 17 | 1 | T100 |
| 0 | 1 | 2 | 91 | 612 | 0 | 0 | 1 | 21 | 2 | SM100 |
| 2 | 0 | 3 | 16 | 40 | 1 | 1 | 0 | 7 | 3 | Q100 |
| 5 | 0 | 2 | 30 | 52 | 1 | 6 | 5 | 23 | 1 | P100 |
| 2 | 0 | 7 | 40 | 59 | 0 | 2 | 104 | 23 | 6 | H100 |
| 3 | 0 | 4 | 64 | 33 | 0 | 0 | 1 | 593 | 2 | G100 |
| 3 | 0 | 4 | 71 | 199 | 0 | 1 | 2 | 35 | 48 | C100 |

Table 5.36 Confusion matrix for IB5 classifier for *generalDS*.

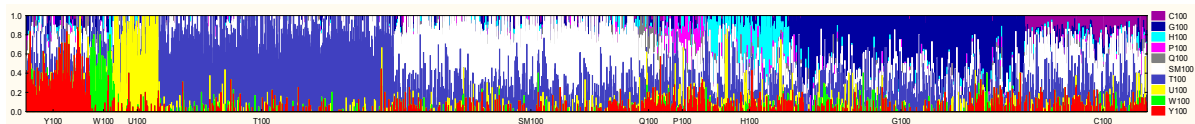


Figure 5.44 Graph of predictions by IB5 classifier on the training *generalDS* set.

5.4.7.2 Allele Length Dataset

Parameters used

```
Scheme:          weka.classifiers.lazy.IBk -K 5 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""
```

Results of classification

```
Correctly Classified Instances      1661          49.8799 %
Incorrectly Classified Instances    1669          50.1201 %
Kappa statistic                    0.3849
Mean absolute error                 0.1202
Root mean squared error            0.2618
Relative absolute error            71.6336 %
Root relative squared error        90.4197 %
Total Number of Instances         3330
```

=== Detailed Accuracy By Class ===

| | TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area | Class |
|---------------|---------|---------|-----------|--------|-----------|----------|-------|
| | 0.41 | 0.039 | 0.385 | 0.41 | 0.397 | 0.781 | Y100 |
| | 0.152 | 0.005 | 0.37 | 0.152 | 0.215 | 0.682 | W100 |
| | 0.708 | 0.012 | 0.713 | 0.708 | 0.711 | 0.902 | U100 |
| | 0.695 | 0.192 | 0.493 | 0.695 | 0.577 | 0.822 | T100 |
| | 0.558 | 0.198 | 0.442 | 0.558 | 0.493 | 0.764 | SM100 |
| | 0.027 | 0.002 | 0.286 | 0.027 | 0.05 | 0.598 | Q100 |
| | 0.072 | 0.007 | 0.3 | 0.072 | 0.116 | 0.629 | P100 |
| | 0.272 | 0.014 | 0.606 | 0.272 | 0.375 | 0.739 | H100 |
| | 0.639 | 0.126 | 0.574 | 0.639 | 0.604 | 0.85 | G100 |
| | 0.154 | 0.024 | 0.438 | 0.154 | 0.228 | 0.633 | C100 |
| Weighted Avg. | 0.499 | 0.117 | 0.49 | 0.499 | 0.471 | 0.775 | |

| | a | b | c | d | e | f | g | h | i | j | |
|----|----|----|-----|-----|---|---|----|-----|----|---|-------|
| 77 | 2 | 2 | 26 | 29 | 2 | 4 | 5 | 40 | 1 | | Y100 |
| 1 | 10 | 1 | 20 | 10 | 0 | 0 | 3 | 21 | 0 | | W100 |
| 3 | 0 | 97 | 9 | 18 | 0 | 2 | 0 | 5 | 3 | | U100 |
| 12 | 3 | 5 | 490 | 115 | 0 | 2 | 6 | 62 | 10 | | T100 |
| 18 | 3 | 14 | 181 | 407 | 0 | 1 | 7 | 74 | 25 | | SM100 |
| 15 | 0 | 3 | 17 | 18 | 2 | 1 | 0 | 16 | 1 | | Q100 |
| 18 | 0 | 0 | 31 | 27 | 2 | 9 | 4 | 22 | 12 | | P100 |
| 15 | 5 | 8 | 25 | 65 | 0 | 5 | 66 | 45 | 9 | | H100 |
| 25 | 4 | 2 | 107 | 91 | 1 | 5 | 7 | 447 | 11 | | G100 |
| 16 | 0 | 4 | 87 | 141 | 0 | 1 | 11 | 47 | 56 | | C100 |

Table 5.37 Confusion matrix for IB5 classifier for *allelelengthDS*.



Figure 5.45 Graph of predictions by IB5 classifier on the training *allelelengthDS* set.

5.4.7.3 Allele Frequency Dataset

Parameters used

```
Scheme:          weka.classifiers.lazy.IBk -K 5 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""
```

Results of classification

```
Correctly Classified Instances      1512          45.4054 %
Incorrectly Classified Instances    1818          54.5946 %
Kappa statistic                    0.3291
Mean absolute error                 0.1256
Root mean squared error            0.2716
Relative absolute error            74.8573 %
Root relative squared error        93.7769 %
Total Number of Instances         3330
```

=== Detailed Accuracy By Class ===

| | TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area | Class |
|---------------|---------|---------|-----------|--------|-----------|----------|-------|
| | 0.234 | 0.031 | 0.314 | 0.234 | 0.268 | 0.691 | Y100 |
| | 0.258 | 0.014 | 0.266 | 0.258 | 0.262 | 0.766 | W100 |
| | 0.657 | 0.021 | 0.573 | 0.657 | 0.612 | 0.895 | U100 |
| | 0.633 | 0.219 | 0.437 | 0.633 | 0.517 | 0.773 | T100 |
| | 0.515 | 0.16 | 0.475 | 0.515 | 0.494 | 0.744 | SM100 |
| | 0 | 0.001 | 0 | 0 | 0 | 0.568 | Q100 |
| | 0.064 | 0.009 | 0.216 | 0.064 | 0.099 | 0.658 | P100 |
| | 0.206 | 0.012 | 0.568 | 0.206 | 0.302 | 0.725 | H100 |
| | 0.647 | 0.185 | 0.482 | 0.647 | 0.553 | 0.814 | G100 |
| | 0.077 | 0.02 | 0.318 | 0.077 | 0.124 | 0.624 | C100 |
| Weighted Avg. | 0.454 | 0.127 | 0.429 | 0.454 | 0.419 | 0.747 | |

| a | b | c | d | e | f | g | h | i | j | |
|----|----|----|-----|-----|---|---|----|-----|----|-------|
| 44 | 3 | 5 | 38 | 15 | 0 | 6 | 4 | 68 | 5 | Y100 |
| 1 | 17 | 0 | 19 | 5 | 0 | 1 | 1 | 21 | 1 | W100 |
| 4 | 0 | 90 | 14 | 13 | 0 | 1 | 1 | 12 | 2 | U100 |
| 14 | 12 | 11 | 446 | 108 | 2 | 1 | 3 | 100 | 8 | T100 |
| 13 | 14 | 17 | 185 | 376 | 0 | 3 | 6 | 99 | 17 | SM100 |
| 5 | 0 | 5 | 17 | 14 | 0 | 2 | 2 | 26 | 2 | Q100 |
| 15 | 2 | 3 | 34 | 24 | 1 | 8 | 7 | 28 | 3 | P100 |
| 10 | 5 | 12 | 38 | 50 | 0 | 5 | 50 | 60 | 13 | H100 |
| 22 | 7 | 5 | 119 | 70 | 1 | 6 | 8 | 453 | 9 | G100 |
| 12 | 4 | 9 | 111 | 117 | 0 | 4 | 6 | 72 | 28 | C100 |

Table 5.38 Confusion matrix for IB5 classifier for *allelefrequencyDS*.

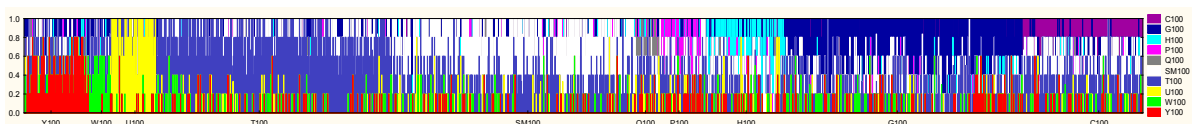


Figure 5.46 Graph of predictions by IB5 classifier on the training *allelefrequencyDS* set.

5.4.7.4 Discussion of IB5 results

Instead of 0 or 1 probability set by IB1 classifier, IB5 classifier set the probabilities of prediction according to voting 5 nearest neighbour instances (according to Euclidean distance) and their classes. So, there is better chance to predict proper class on the smaller database.

On general dataset, IB5 classifier classified correctly 65.38 % of instances with Kappa statistic equals to 0.5701. Overall FP Rate is calculated as 0.089, Precision equals 0.68 and F-Measure is 0.605 as weighted means across all of classes based on cross validation results.

Surprisingly, well defined class according to TP Rate is T100 in this case with TP Rate=0.901, the worst defined one is Q100 with TP Rate=0.014 when small frequency of Q100 in dataset plays a role rather than its similarity to other classes. FP Rate ranges from 0.00 (W100, Q100) to 0.195 for SM100. Anyway, small FP Rates can be useful for specific tasks in breed discrimination problem. When model classifies with small FP Rate for predicted class, we can say that predicted class is mostly sure well predicted. Best precision was reached for W100 class (0.950, F-Measure=0.442), the worst one for Q100 (0.500, F-Measure=0.027)

Graph 5.44 displays better portions of predicted probabilities done by IB5 than IB1 classifier on the whole training set. Anyway, huge admixture is present between T100 and G100 classes as well as between SM100 and C100 classes. We need to imagine that in case of IB5 classifier, when we can find three similar individuals in class which is not actual class of classified individual, it will be misclassified with 3/5 probability.

This fact is evident on results calculated on allele length and allele frequency datasets (figure 5.45 and 5.46). Graphs of predictions are equally divided according to portions predicted by similar individuals from different classes, what is caused (as in case of IB1) by double and sorting of information which define individual in those two datasets.

Only 49.88 % and 45.41 % of correctly classified instances were reached for allele length and allele frequency dataset with Kappa statistic equals 0.3849 resp. 0.3291, overall FP Rate 0.117 resp. 0.127, Precision=0.49 resp. 0.429, F-Measure=0.471 resp. 0.419.

Best TP Rate was calculated for U100 class (0.708) in allele length dataset, 0.657 for U100 in allele frequency dataset. Worst results are - 0.27 for Q100, resp. 0.00 in Q100. FP Rate ranges from 0.002 (Q100) to 0.198 (SM100) in allele length and from 0.001 in Q100 to 0.219 in SM100 in allele frequency dataset.

Best precision was reached for U100 class in both of datasets (0.713, F-Measure=0.711; 0.573, F-Measure=0.612) the worst for Q100 class (0.3, F-Measure=0.116; 0.00, F-Measure=0.00).

When one of all classes is misclassified completely, we can not recommend IB5 as a classifier for breed discrimination task. Anyway, IBk concept of classifiers offers quite a lot of possibilities for modifications (as metrics used for similarity calculations, voting modifications on classes etc.), so we can not reject the whole concept how is shown below in case of G-metric classification based on this concept as well.

5.4.8 Support Vector Machines

5.4.8.1 General Dataset

Parameters used

```
Scheme:          weka.classifiers.functions.SMO -C 1.0 -L 0.0010 -P 1.0E-12 -N 0 -V -1 -W 1
-K "weka.classifiers.functions.supportVector.PolyKernel -C 250007 -E 1.0"
```

Results of classification

```
Correctly Classified Instances      2609          78.3483 %
Incorrectly Classified Instances    721           21.6517 %
Kappa statistic                     0.7403
Mean absolute error                 0.1621
Root mean squared error            0.2754
Relative absolute error             96.6255 %
Root relative squared error        95.0954 %
Total Number of Instances         3330
```

=== Detailed Accuracy By Class ===

| | TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area | Class |
|---------------|---------|---------|-----------|--------|-----------|----------|-------|
| | 0.723 | 0.025 | 0.638 | 0.723 | 0.678 | 0.947 | Y100 |
| | 0.742 | 0.002 | 0.875 | 0.742 | 0.803 | 0.994 | W100 |
| | 0.854 | 0.012 | 0.76 | 0.854 | 0.804 | 0.969 | U100 |
| | 0.887 | 0.047 | 0.836 | 0.887 | 0.86 | 0.958 | T100 |
| | 0.811 | 0.076 | 0.75 | 0.811 | 0.779 | 0.928 | SM100 |
| | 0.562 | 0.005 | 0.732 | 0.562 | 0.636 | 0.91 | Q100 |
| | 0.456 | 0.01 | 0.633 | 0.456 | 0.53 | 0.872 | P100 |
| | 0.654 | 0.014 | 0.791 | 0.654 | 0.716 | 0.902 | H100 |
| | 0.92 | 0.027 | 0.902 | 0.92 | 0.911 | 0.976 | G100 |
| | 0.521 | 0.04 | 0.612 | 0.521 | 0.563 | 0.846 | C100 |
| Weighted Avg. | 0.783 | 0.04 | 0.78 | 0.783 | 0.779 | 0.935 | |

| | a | b | c | d | e | f | g | h | i | j | |
|-----|----|-----|-----|-----|----|----|-----|-----|-----|-------|--|
| 136 | 0 | 3 | 3 | 10 | 2 | 5 | 6 | 7 | 16 | Y100 | |
| 2 | 49 | 1 | 5 | 0 | 1 | 0 | 1 | 7 | 0 | W100 | |
| 3 | 1 | 117 | 2 | 4 | 0 | 1 | 2 | 4 | 3 | U100 | |
| 9 | 1 | 5 | 625 | 38 | 0 | 1 | 5 | 15 | 6 | T100 | |
| 7 | 0 | 7 | 44 | 592 | 3 | 3 | 4 | 7 | 63 | SM100 | |
| 5 | 1 | 4 | 5 | 8 | 41 | 5 | 0 | 0 | 4 | Q100 | |
| 12 | 0 | 2 | 10 | 16 | 2 | 57 | 5 | 7 | 14 | P100 | |
| 6 | 1 | 8 | 9 | 23 | 3 | 10 | 159 | 13 | 11 | H100 | |
| 11 | 3 | 3 | 19 | 8 | 1 | 2 | 6 | 644 | 3 | G100 | |
| 22 | 0 | 4 | 26 | 90 | 3 | 6 | 13 | 10 | 189 | C100 | |

Table 5.39 Confusion matrix for SMO classifier for *generalDS*.

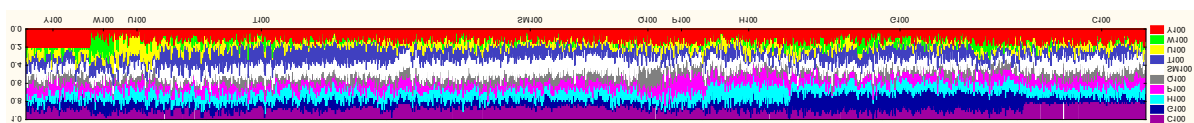


Figure 5.47 Graph of predictions by SMO classifier on the training *generalDS* set.

5.4.8.2 Allele Length Dataset

Parameters used

```
Scheme:          weka.classifiers.functions.SMO -C 1.0 -L 0.0010 -P 1.0E-12 -N 0 -V -1 -W 1
-K "weka.classifiers.functions.supportVector.PolyKernel -C 250007 -E 1.0"
```

Results of classification

```
Correctly Classified Instances      1690          50.7508 %
Incorrectly Classified Instances    1640          49.2492 %
Kappa statistic                    0.394
Mean absolute error                 0.1656
Root mean squared error             0.2821
Relative absolute error             98.702 %
Root relative squared error         97.4057 %
Total Number of Instances          3330
```

=== Detailed Accuracy By Class ===

| | TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area | Class |
|---------------|---------|---------|-----------|--------|-----------|----------|-------|
| | 0.346 | 0.026 | 0.442 | 0.346 | 0.388 | 0.853 | Y100 |
| | 0.015 | 0 | 1 | 0.015 | 0.03 | 0.849 | W100 |
| | 0.672 | 0.015 | 0.657 | 0.672 | 0.664 | 0.933 | U100 |
| | 0.682 | 0.163 | 0.529 | 0.682 | 0.596 | 0.828 | T100 |
| | 0.593 | 0.208 | 0.445 | 0.593 | 0.509 | 0.751 | SM100 |
| | 0.164 | 0.001 | 0.75 | 0.164 | 0.27 | 0.851 | Q100 |
| | 0.184 | 0.008 | 0.479 | 0.184 | 0.266 | 0.728 | P100 |
| | 0.358 | 0.017 | 0.621 | 0.358 | 0.454 | 0.802 | H100 |
| | 0.646 | 0.156 | 0.524 | 0.646 | 0.579 | 0.797 | G100 |
| | 0.121 | 0.017 | 0.473 | 0.121 | 0.193 | 0.692 | C100 |
| Weighted Avg. | 0.508 | 0.118 | 0.523 | 0.508 | 0.479 | 0.791 | |

| | a | b | c | d | e | f | g | h | i | j | |
|----|---|----|-----|-----|----|----|----|-----|----|-------|--|
| 65 | 0 | 2 | 6 | 32 | 1 | 11 | 2 | 65 | 4 | Y100 | |
| 1 | 1 | 0 | 15 | 12 | 0 | 0 | 2 | 35 | 0 | W100 | |
| 4 | 0 | 92 | 13 | 23 | 0 | 0 | 2 | 3 | 0 | U100 | |
| 7 | 0 | 9 | 481 | 121 | 0 | 1 | 5 | 78 | 3 | T100 | |
| 12 | 0 | 10 | 168 | 433 | 0 | 1 | 8 | 85 | 13 | SM100 | |
| 9 | 0 | 4 | 13 | 11 | 12 | 4 | 3 | 15 | 2 | Q100 | |
| 10 | 0 | 4 | 24 | 37 | 2 | 23 | 4 | 17 | 4 | P100 | |
| 11 | 0 | 9 | 18 | 50 | 0 | 2 | 87 | 53 | 13 | H100 | |
| 15 | 0 | 6 | 102 | 100 | 0 | 2 | 13 | 452 | 10 | G100 | |
| 13 | 0 | 4 | 70 | 154 | 1 | 4 | 14 | 59 | 44 | C100 | |

Table 5.40 Confusion matrix for SMO classifier for *allelelengthDS*.

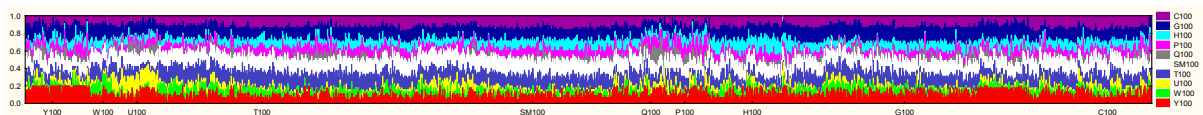


Figure 5.48 Graph of predictions by SMO classifier on the training *allelelengthDS* set.

5.4.8.3 Allele Frequency Dataset

Parameters used

```
Scheme:          weka.classifiers.functions.SMO -C 1.0 -L 0.0010 -P 1.0E-12 -N 0 -V -1 -W 1
-K "weka.classifiers.functions.supportVector.PolyKernel -C 250007 -E 1.0"
```

Results of classification

```
Correctly Classified Instances      1637          49.1592 %
Incorrectly Classified Instances    1693          50.8408 %
Kappa statistic                    0.3735
Mean absolute error                 0.1657
Root mean squared error             0.2823
Relative absolute error             98.7887 %
Root relative squared error         97.4948 %
Total Number of Instances          3330
```

=== Detailed Accuracy By Class ===

| | TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area | Class |
|---------------|---------|---------|-----------|--------|-----------|----------|-------|
| | 0.309 | 0.021 | 0.464 | 0.309 | 0.371 | 0.841 | Y100 |
| | 0 | 0 | 0 | 0 | 0 | 0.82 | W100 |
| | 0.569 | 0.013 | 0.645 | 0.569 | 0.605 | 0.942 | U100 |
| | 0.645 | 0.177 | 0.495 | 0.645 | 0.56 | 0.806 | T100 |
| | 0.595 | 0.186 | 0.473 | 0.595 | 0.527 | 0.763 | SM100 |
| | 0.123 | 0.003 | 0.45 | 0.123 | 0.194 | 0.832 | Q100 |
| | 0.104 | 0.006 | 0.406 | 0.104 | 0.166 | 0.713 | P100 |
| | 0.395 | 0.019 | 0.623 | 0.395 | 0.484 | 0.828 | H100 |
| | 0.65 | 0.179 | 0.492 | 0.65 | 0.56 | 0.789 | G100 |
| | 0.107 | 0.026 | 0.333 | 0.107 | 0.163 | 0.644 | C100 |
| Weighted Avg. | 0.492 | 0.122 | 0.472 | 0.492 | 0.461 | 0.782 | |

| a | b | c | d | e | f | g | h | i | j | |
|----|---|----|-----|-----|---|----|----|-----|----|-------|
| 58 | 0 | 2 | 10 | 28 | 0 | 8 | 5 | 72 | 5 | Y100 |
| 0 | 0 | 1 | 8 | 15 | 0 | 0 | 2 | 40 | 0 | W100 |
| 4 | 0 | 78 | 14 | 27 | 1 | 1 | 1 | 6 | 5 | U100 |
| 1 | 0 | 4 | 455 | 134 | 0 | 0 | 1 | 103 | 7 | T100 |
| 11 | 0 | 6 | 170 | 434 | 0 | 0 | 7 | 86 | 16 | SM100 |
| 5 | 0 | 4 | 16 | 11 | 9 | 5 | 1 | 20 | 2 | Q100 |
| 13 | 0 | 8 | 25 | 28 | 4 | 13 | 3 | 27 | 4 | P100 |
| 14 | 0 | 7 | 23 | 40 | 1 | 3 | 96 | 41 | 18 | H100 |
| 10 | 0 | 4 | 102 | 88 | 3 | 0 | 17 | 455 | 21 | G100 |
| 9 | 0 | 7 | 96 | 112 | 2 | 2 | 21 | 75 | 39 | C100 |

Table 5.41 Confusion matrix for SMO classifier for *allelefrequencyDS*.

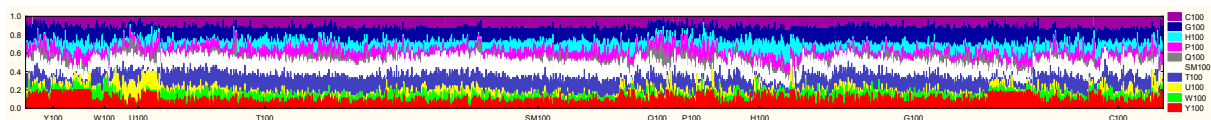


Figure 5.49 Graph of predictions by SMO classifier on the training *allelefrequencyDS* set.

5.4.8.4 Discussion of Support Vector Machine results

SMO implementation of Support Vector Machines principle in Weka was run with parameters displayed in section 5.4.8.1. On the general dataset, 78.35 % of instances were classified correctly by this method. Kappa statistic of the whole model verified by 10 fold cross validation was 0.7403. Overall weighted parameters of classification was calculated as FP Rate=0.04, Precision=0.78 and F-Measure=0.779.

The best classified class by the meaning of TP Rate was G100 (0.920), the worst one was P100 (0.456). FP Rate ranged from 0.002 in W100 to 0.076 in SM100. Precision and F-Measure results ranged from 0.902; 0.911 (G100) to 0.612; 0.563 for C100 class.

Figure 5.47 displays predictions of SMO model on the whole dataset. For this dataset, however predictions and derived hyperplanes (and support vectors) are very closed, model can predict class very accurate. Connected probabilities in each breed are concluded according to proper breed.

Unfortunately, when we run the same method on allele length dataset, classification power is dropping significantly. It can be caused mainly by “weak” definition of individual based on frequency data. Only 50.75 % of correctly classified instances is reported on allele length dataset with model Kappa statistic=0.3940. Overall FP Rate is then 0.118, Precision=0.523 and F-Measure=0.479. The best TP Rate was obtained for T100 class (0.682), the smallest value for W100 class (0.015). False positive rate ranges from 0 (W100) to 0.208 for SM100 class. The best precision was obtained for W100 (1.00, F-Measure=0.03), the worst one for Y100 class (0.442, F-Measure=0.388).

Similar results were obtained for allele frequency dataset - 49.16 % of correctly classified instances, Kappa statistic equals to 0.3735, overall FP Rate=0.122, Precision=0.472, F-Measure=0.461.

G100 class has a best TP Rate (0.65), W100 a worst one equals 0. FP Rate was calculated as lowest in W100 class (0.00) and highest in SM100 (0.186). We can see that any instance of W100 was classified to its class. Precision ranges from 0.00 (W100, F-Measure=0.00) to 0.645 (U100, F-Measure=0.56).

5.4.9 Voted classifier

5.4.9.1 General Dataset

Parameters used

```
weka.classifiers.meta.Vote -S 1 -B "weka.classifiers.bayes.BayesNet -D -Q weka.classifiers.bayes.net.search.local.K2 -- -P 1 -S BAYES -E weka.classifiers.bayes.net.estimate.SimpleEstimator -- -A 0.5" -B "weka.classifiers.bayes.NaiveBayes " -B "weka.classifiers.functions.SMO\""" -R AVG
```

Results of classification

```
Correctly Classified Instances      2764           83.003 %
Incorrectly Classified Instances    566           16.997 %
Kappa statistic                    0.7951
Mean absolute error                 0.0837
Root mean squared error            0.1715
Relative absolute error             49.8913 %
Root relative squared error        59.2282 %
Total Number of Instances         3330
```

=== Detailed Accuracy By Class ===

| | TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area | Class |
|---------------|---------|---------|-----------|--------|-----------|----------|-------|
| | 0.787 | 0.011 | 0.809 | 0.787 | 0.798 | 0.981 | Y100 |
| | 0.727 | 0.001 | 0.923 | 0.727 | 0.814 | 0.998 | W100 |
| | 0.912 | 0.006 | 0.868 | 0.912 | 0.89 | 0.978 | U100 |
| | 0.915 | 0.034 | 0.878 | 0.915 | 0.896 | 0.984 | T100 |
| | 0.881 | 0.085 | 0.745 | 0.881 | 0.807 | 0.965 | SM100 |
| | 0.479 | 0.001 | 0.897 | 0.479 | 0.625 | 0.944 | Q100 |
| | 0.456 | 0.004 | 0.826 | 0.456 | 0.588 | 0.94 | P100 |
| | 0.79 | 0.006 | 0.914 | 0.79 | 0.848 | 0.933 | H100 |
| | 0.959 | 0.022 | 0.922 | 0.959 | 0.94 | 0.992 | G100 |
| | 0.551 | 0.036 | 0.651 | 0.551 | 0.597 | 0.917 | C100 |
| Weighted Avg. | 0.83 | 0.036 | 0.831 | 0.83 | 0.825 | 0.968 | |

| a | b | c | d | e | f | g | h | i | j | |
|-----|----|-----|-----|-----|----|----|-----|-----|-----|-------|
| 148 | 0 | 0 | 6 | 15 | 0 | 1 | 1 | 3 | 14 | Y100 |
| 0 | 48 | 1 | 5 | 0 | 0 | 0 | 0 | 10 | 2 | W100 |
| 0 | 0 | 125 | 3 | 3 | 0 | 1 | 0 | 2 | 3 | U100 |
| 3 | 1 | 4 | 645 | 33 | 0 | 1 | 0 | 14 | 4 | T100 |
| 4 | 0 | 0 | 33 | 643 | 0 | 2 | 3 | 4 | 41 | SM100 |
| 6 | 1 | 3 | 2 | 10 | 35 | 1 | 2 | 3 | 10 | Q100 |
| 11 | 1 | 2 | 9 | 13 | 1 | 57 | 5 | 5 | 21 | P100 |
| 3 | 0 | 7 | 7 | 20 | 1 | 1 | 192 | 7 | 5 | H100 |
| 2 | 0 | 1 | 9 | 7 | 1 | 1 | 1 | 671 | 7 | G100 |
| 6 | 1 | 1 | 16 | 119 | 1 | 4 | 6 | 9 | 200 | C100 |

Table 5.42 Confusion matrix for Voted classifier for *generalDS*.

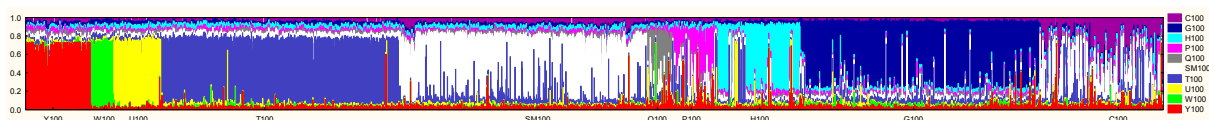


Figure 5.50 Graph of predictions by Voted classifier on the training *generalDS* set.

5.4.9.2 Allele Length Dataset

Parameters used

```
weka.classifiers.meta.Vote -S 1 -B "weka.classifiers.bayes.BayesNet -D -Q weka.classifiers.bayes.net.search.local.K2 -- -P 1 -S BAYES -E weka.classifiers.bayes.net.estimate.SimpleEstimator -- -A 0.5" -B "weka.classifiers.bayes.NaiveBayes " -B "weka.classifiers.functions.SMO -C 1.0 -L 0.0010 -P 1.0E-12 -N 0 -V -1 -W 1 -K \"weka.classifiers.functions.supportVector.PolyKernel -C 250007 -E 1.0\" \" -R AVG
```

Results of classification

```
Correctly Classified Instances      2709      81.3514 %
Incorrectly Classified Instances    621      18.6486 %
Kappa statistic                    0.7762
Mean absolute error                0.1037
Root mean squared error            0.198
Relative absolute error             61.8274 %
Root relative squared error        68.3803 %
Total Number of Instances          3330
```

=== Detailed Accuracy By Class ===

| | TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area | Class |
|---------------|---------|---------|-----------|--------|-----------|----------|-------|
| | 0.761 | 0.013 | 0.781 | 0.761 | 0.771 | 0.96 | Y100 |
| | 0.833 | 0.003 | 0.859 | 0.833 | 0.846 | 0.993 | W100 |
| | 0.898 | 0.012 | 0.759 | 0.898 | 0.823 | 0.972 | U100 |
| | 0.921 | 0.043 | 0.851 | 0.921 | 0.884 | 0.976 | T100 |
| | 0.807 | 0.076 | 0.748 | 0.807 | 0.777 | 0.94 | SM100 |
| | 0.575 | 0.005 | 0.737 | 0.575 | 0.646 | 0.937 | Q100 |
| | 0.528 | 0.006 | 0.776 | 0.528 | 0.629 | 0.932 | P100 |
| | 0.786 | 0.016 | 0.799 | 0.786 | 0.793 | 0.91 | H100 |
| | 0.943 | 0.021 | 0.922 | 0.943 | 0.932 | 0.984 | G100 |
| | 0.526 | 0.028 | 0.697 | 0.526 | 0.6 | 0.91 | C100 |
| Weighted Avg. | 0.814 | 0.036 | 0.81 | 0.814 | 0.808 | 0.954 | |

| a | b | c | d | e | f | g | h | i | j | |
|-----|----|-----|-----|-----|----|----|-----|-----|-----|-------|
| 143 | 2 | 3 | 9 | 11 | 1 | 3 | 1 | 7 | 8 | Y100 |
| 0 | 55 | 0 | 4 | 2 | 0 | 0 | 2 | 3 | 0 | W100 |
| 2 | 0 | 123 | 3 | 6 | 0 | 1 | 2 | 0 | 0 | U100 |
| 7 | 2 | 5 | 649 | 25 | 1 | 1 | 0 | 11 | 4 | T100 |
| 5 | 2 | 6 | 45 | 589 | 6 | 3 | 16 | 8 | 50 | SM100 |
| 5 | 1 | 5 | 5 | 4 | 42 | 1 | 1 | 3 | 6 | Q100 |
| 3 | 0 | 2 | 8 | 16 | 4 | 66 | 11 | 6 | 9 | P100 |
| 3 | 1 | 9 | 9 | 19 | 0 | 3 | 191 | 6 | 2 | H100 |
| 6 | 1 | 3 | 8 | 14 | 1 | 1 | 2 | 660 | 4 | G100 |
| 9 | 0 | 6 | 23 | 101 | 2 | 6 | 13 | 12 | 191 | C100 |

Table 5.43 Confusion matrix for Voted classifier for *allelelengthDS*.

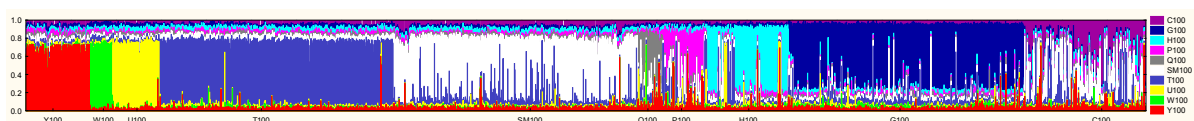


Figure 5.51 Graph of predictions by Voted classifier on the training *allelelengthDS* set.

5.4.9.3 Allele Frequency Dataset

Parameters used

```
weka.classifiers.meta.Vote -S 1 -B "weka.classifiers.bayes.BayesNet -D -Q weka.classifiers.bayes.net.search.local.K2 -- -P 1 -S BAYES -E weka.classifiers.bayes.net.estimate.SimpleEstimator -- -A 0.5" -B "weka.classifiers.bayes.NaiveBayes " -B "weka.classifiers.functions.SMO -C 1.0 -L 0.0010 -P 1.0E-12 -N 0 -V -1 -W 1 -K \"weka.classifiers.functions.supportVector.PolyKernel -C 250007 -E 1.0\" \" -R AVG
```

Results of classification

```
Correctly Classified Instances      2730      81.982 %
Incorrectly Classified Instances    600      18.018 %
Kappa statistic                    0.7832
Mean absolute error                 0.102
Root mean squared error             0.1954
Relative absolute error             60.8146 %
Root relative squared error        67.4764 %
Total Number of Instances          3330
```

=== Detailed Accuracy By Class ===

| | TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area | Class |
|---------------|---------|---------|-----------|--------|-----------|----------|-------|
| | 0.761 | 0.011 | 0.808 | 0.761 | 0.784 | 0.961 | Y100 |
| | 0.818 | 0.001 | 0.947 | 0.818 | 0.878 | 0.993 | W100 |
| | 0.905 | 0.008 | 0.827 | 0.905 | 0.864 | 0.978 | U100 |
| | 0.929 | 0.047 | 0.841 | 0.929 | 0.883 | 0.977 | T100 |
| | 0.822 | 0.077 | 0.749 | 0.822 | 0.784 | 0.941 | SM100 |
| | 0.548 | 0.002 | 0.851 | 0.548 | 0.667 | 0.954 | Q100 |
| | 0.568 | 0.004 | 0.845 | 0.568 | 0.679 | 0.947 | P100 |
| | 0.782 | 0.011 | 0.848 | 0.782 | 0.814 | 0.914 | H100 |
| | 0.947 | 0.024 | 0.912 | 0.947 | 0.929 | 0.986 | G100 |
| | 0.523 | 0.032 | 0.669 | 0.523 | 0.587 | 0.902 | C100 |
| Weighted Avg. | 0.82 | 0.037 | 0.818 | 0.82 | 0.815 | 0.956 | |

| a | b | c | d | e | f | g | h | i | j | |
|-----|----|-----|-----|-----|----|----|-----|-----|-----|-------|
| 143 | 0 | 0 | 8 | 14 | 1 | 2 | 2 | 5 | 13 | Y100 |
| 0 | 54 | 0 | 5 | 1 | 0 | 0 | 1 | 5 | 0 | W100 |
| 1 | 0 | 124 | 2 | 4 | 0 | 0 | 1 | 2 | 3 | U100 |
| 2 | 0 | 4 | 655 | 27 | 1 | 1 | 1 | 12 | 2 | T100 |
| 4 | 0 | 2 | 51 | 600 | 1 | 1 | 9 | 9 | 53 | SM100 |
| 2 | 2 | 4 | 6 | 4 | 40 | 1 | 1 | 7 | 6 | Q100 |
| 6 | 0 | 3 | 9 | 14 | 1 | 71 | 8 | 4 | 9 | P100 |
| 4 | 1 | 7 | 8 | 17 | 0 | 2 | 190 | 8 | 6 | H100 |
| 7 | 0 | 1 | 11 | 12 | 1 | 1 | 2 | 663 | 2 | G100 |
| 8 | 0 | 5 | 24 | 108 | 2 | 5 | 9 | 12 | 190 | C100 |

Table 5.44 Confusion matrix for Voted classifier for *allelefrequencyDS*.

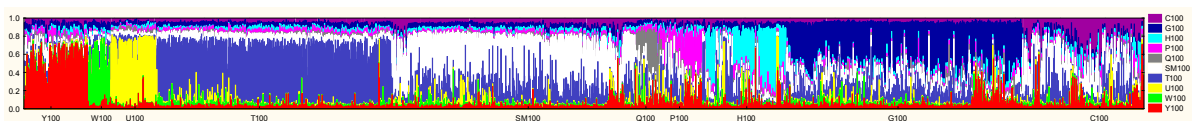


Figure 5.52 Graph of predictions by Voted classifier on the training *allelefrequencyDS* set.

5.4.9.4 Discussion of Voted classifier results

Voted classifier implementing average voting between its basic classifiers was used to explore how the combination of best performance classifier can influence results on all of three examined datasets. As a voted basic classifiers, Naive Baies, Bayes Net and SMO classifiers were selected with parameters described above in each section.

Voted classifier reached on general data set following results - 83.00 % of correctly classified instances, Kappa statistic equals 0.7951 and average FP Rate calculated across folds of cross validation and classes equals 0.036. Weighted average Precision was 0.831, F-Measure=0.825.

Class with assigned highest TP Rate was G100 (0.959), the lowest value appeared for P100 class (0.456). FP Rate ranged from 0.001 (W100, Q100) to 0.085 (SM100). Best Precision was reached for W100 class (0.923; F-Measure=0.814), the worst one for C100 (0.651; F-Measure=0.597). Figure 5.50 displays predicted probabilities on the whole dataset. It shows how classifier works, so we can identify small portion of probabilities of each class predicted for individuals caused by averaging predictions of all three basis classifiers.

On the allele length dataset, similar results were obtained. Percentage of correctly classified instances equals to 81.35 % with Kappa statistic equals 0.7762, average FP Rate=0.036, Precision=0.81 and F-Measure=0.808. Best TP Rate was reached for G100 class (0.943), the worst one for C100 class (0.526). W100 is a class with a lowest FP Rate (0.003), SM100 one with the highest value of FP Rate (0.076). Precision ranged from 0.697 (F-Measure=0.60) for C100 class to 0.922 (F-Measure=0.932) for G100 class.

Within results on allele frequency dataset, we can identify that voted classifier output 81.98 % of correctly classified instances, Kappa statistic equals 0.7832, average FP Rate=0.037, Precision=0.818 and F-Measure=0.815. TP Rate=0.947 was reached for G100 class as the best one, TP Rate=0.523 for C100 class as the lowest one. FP Rate=0.001 represents smallest value for W100 class, FP Rate=0.077 represent the highest one for SM100 class. Precision ranged from 0.669 (C100; F-Measure=0.587) to 0.947 for W100 (F-Measure=0.878).

Big admixture of predicted probabilities is displayed on figure 5.52, especially between T100 and G100 classes and SM100 and C100 classes what is in accordance with results in confusion matrix (table 5.44).

5.4.10 G-metric classification

5.4.10.1 Implementation of G-metric classifier

Datatypes

```
genotype = packed record
  a1: string[3];
  a2: string[3];
end;
```

Key record datatype called genotype is used to store genotypes (allele pair) as two strings of length equals 3. Anyway, this solution is used as alleles are stored in SQL database as the same type, otherwise definition in range of integer values. As '?' is present as sign for unknown value in particular dataset, when we are using string implementation, there is no needing for conversions in routine SW usage as printing protocols, inputation of data etc. Only when calculations on allele length are performed, we need to perform conversion between datatypes as is evident.

G-dis function

```
function Gdis(g1,g2: genotype; sat: integer): real;
var gg1, gg2: real;
begin
  if (g1.a1='?') or (g2.a1='?') then gg1:=1 else begin
    gg1:=(abs(strtoint(g1.a1)-strtoint(g2.a1))/maxdiff[sat]);
  end;
  if (g1.a2='?') or (g2.a2='?') then gg2:=1 else begin
    gg2:=(abs(strtoint(g1.a2)-strtoint(g2.a2))/maxdiff[sat]);
  end;
  gdis:=(gg1+gg2)/2;
end;
```

Function G-dis is implemented according mathematical definition described in section 4.3.4.1. It has two sorted genotypes (alleles are sorted in genotypes according their length) as input at it's output is defined as a distance between input genotypes based on their alleles lengths in pairs sorted according their allele lengths in particular loci. It means we calculate distance between "shorter" alleles from each pair firstly, then "longer" ones are used to compute second part of distance between two genotypes. If one or both of alleles in pair which is used for calculation is unknown, then distance between them is consider as 1 - infinity. When calculation of distance is performed as absolute value from allele length differences, it is "normalized" according to maximum distance available from alleles length in particular loci pre-calculated for whole dataset accross all of individuals. So, on this place, whole dataset creates base for distance calculation and it can be said that results depend on the whole dataset or they are valid under the dataset conditions.

Then, the whole distance of two genotypes in one loci is calculated as average distance between two pairs. As function is defined as sub-function of the whole metric, their values range from 0..1 as well.

G-measure

```
function Gmeasure(indiv1, indiv2: indiv):real;
var i: integer;
begin
  for i:=1 to n of loci
    Gmeasure:=Gmeasure + Gdis(Loci i);
  end;
  Gmeasure:=Gmeasure/10;
end;
```

Function called G-measure calculates distance between two individuals based on maximum distances in each loci pre-calculated for the whole dataset using G-dis function for each loci and pair of individuals.

The result is calculated as average value of distances calculated for each loci, so the results of G-measure function range between 0..1 also.

G-metric classification algorithm - IB1 implementation

Initialization

1. Pre-calculate alleles maximum differences for each loci

```
for each loci
  select all of alleles (their lengths)
  sort output
end
MAXDIFF:= store maximal differences for each loci in list as
MAX LENGHT - MIN LENGHT
```

2. Create WHOLESET as a set of individuals with their breed and alleles sorted in each loci (genotype) according their length
3. Create BREED SETS of individuals for each breed
4. Initialize confusion matrix of whole dataset

10 fold cross validation classification

```
for i:=1 to 10
  Initialize confusion matrix of fold n;
  for all of breeds
    TEST SET:=TEST SET+1/10 of BREED SET;
    TRAIN SET:=TRAIN SET+BREED SET-TEST SET;
  end;
  for each individual 1 from TEST SET
    for each individual 2 from TRAIN SET
      INDIVIDUAL DISTANCES:=Gmeasure(individual 1, individual
2);
      for each BREED
        INDIVIDUAL REGRESSION:=regression on distances of indi-
vidual to all of individuals in breed
        INDIVIDUAL RESULTS:=minimum from INDIVIDUAL DISTANCES;
      end;
    end;
  end;
```

5 Results and Discussion

```
INDIVIDUAL RESULTS;
select minimum from INDIVIDUAL RESULTS;
output INDIVIDUAL RESULTS;
actualize confusion matrix of fold n;
actualize confusion matrix of whole dataset;
end;
output confusin matrix of fold n;
output results for fold n;
end;
output confusin matrix for whole dataset;
output results for whole dataset;
```

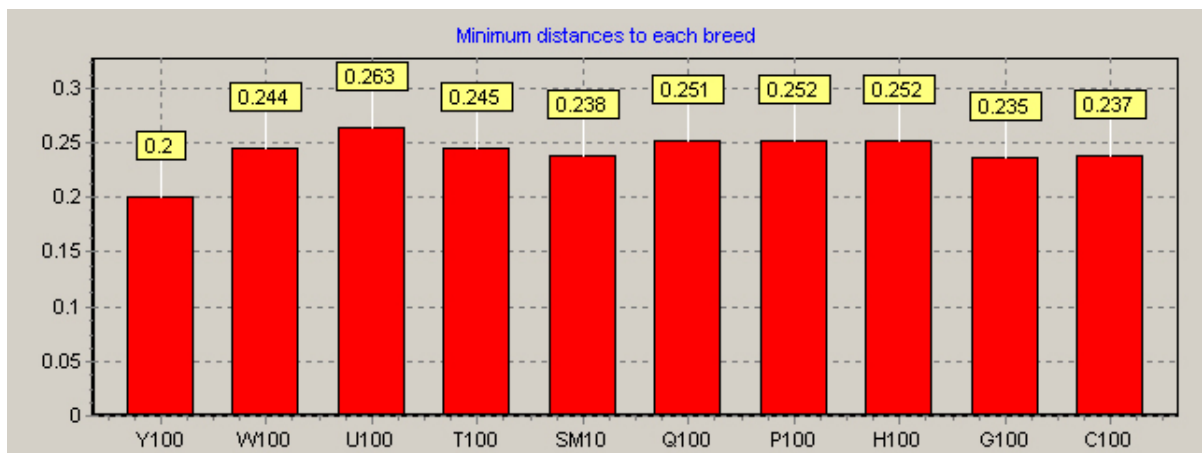


Figure 5.53 Results of G-metric classification of individual by using IB1 implementation.

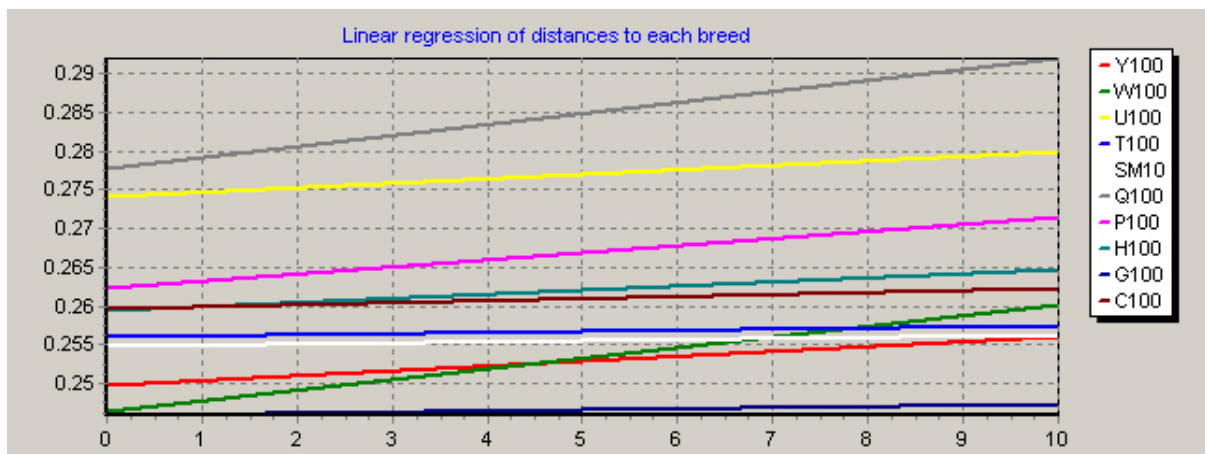


Figure 5.54 Results of G-metric classification of individual by using Slope implementation.

Pseudo code above describes implementation of G-metric classifier. In fact, it is IB1 classifier with usage of Gmeasure and Gdis functions as implementation of G-metric proposed in chapter 4.3.4.1. Algorithm described implements 10 fold cross validation principle, in created software, there is modification when the whole dataset is used as training as well as test set. So, comparable results to algorithms implemented in Weka-3-6-6 framework can be derived.

5 Results and Discussion

After pre-calculation done in case of usage Gmeasure function when maximum differences in allele lengths for all of loci are computed and initialization of global result variables, cycle implements n fold validation is performed. When actual training and test set are established by equal parts of individuals from all of breeds, then for each individual from test set the whole classification model is built as a database of distances from individual to all of individuals in train set, when distance is calculated using Gmeasure function.

As we assume that unknown individual could belongs to each breed used for model building equally, the probability of breed of unknown sample is equal to $1/\text{number of breeds}$ considered. In this case, we just recalculate equal probabilities according to results of distances. As original IB1 classifier uses different strategy when only one instance (with minimal distance) is considered as only one probable, this is a first modification of the whole principle of IB1. IB1 classifier divide probabilities as $0/1$, so the class when closest distance is present has probability of classification equals 1, others have 0.

Then, linear regression is calculated by algorithm over all of computed distances (individual from test set to all of individuals in trainset) for each breed as an expression of summary results for each breed. This can allow to calculate final results of classification based on results of linear regression which can summarize distances of the whole breed to individual effectively as it is described bellow. For linear regression, slope, bias, coefficients of regression and determination plus covariance parameters are calculated based on all of distances sorted.

After that, results for each individual (as a normalized probability of belonging to particular breed) is reported by algorithm.

As results of classification Number of correctly and incorrectly classified instances, Kappa statistics and Confusion matrix are calculated across the whole validation and for each fold. For each class, for each fold and for the whole validation, TP Rate, FP Rate, Precision, Recall, F-Measure and their weighted averages are calculated as well.

The whole concept of implementation offers a lot of possible modifications and it allows usage of a lot of modified principles. When all of distances are computed for classified individual to all of individuals in trainset, we can classify according the whole dataset, or particular breeds datasets. When we choose, e.g. that minimum distance will be calculated thorough the whole training set, we can not conclude probabilities for each breed then. Anyway, given example represent satisfactory condition for implementation of IB1 classifier itself as well as for IB k classifiers with k-NN decision rule of classification. Calculations of classification results by breeds offer quite a lot modifications of basic k-NN algorithm as well as the others ones. We can illustrate it on calculation of INDIVIDUAL RESULTS variable in code above. In this calculation, only minimum distance calculated for each breed can be selected, or we can perform e.g. average distance of k minimals, average distance to all of individuals in breed in training set, selection of best regression parameters for the whole breed etc. Also, normalization of probabilities calculated as corrected equal probabilities of belonging of individual to particular breed can be modified by many ways - e.g. by number of individuals in each breed.

Based on text above, it is evident, that proposed implementation offers effective framework with a lot of possible modifications to use lazy-kind of classification algorithms based on genetic distance measure, which can be subject of modifications as well.

5.4.10.2 General Dataset - Criterion of minimum regression slope on whole breed

distances

Results of classification - 10 CV

```

Correctly Classified Instances      1814      54.4745%
Incorrectly Classified Instances    1516      45.5255%
Kappa statistic                    0.4571
Total number of instances:         3330
    
```

=== Detailed Accuracy By Class ===

| | TP Rate | FP Rate | Precision | Recall | F-Measure | Class |
|---------------|---------|---------|-----------|--------|-----------|-------|
| | 0.830 | 0.227 | 0.243 | 0.830 | 0.376 | Y100 |
| | 0.652 | 0.030 | 0.439 | 0.652 | 0.524 | W100 |
| | 0.861 | 0.054 | 0.551 | 0.861 | 0.672 | U100 |
| | 0.455 | 0.042 | 0.829 | 0.455 | 0.588 | T100 |
| | 0.686 | 0.241 | 0.546 | 0.686 | 0.608 | SM10 |
| | 0.370 | 0.029 | 0.333 | 0.370 | 0.351 | Q100 |
| | 0.080 | 0.006 | 0.500 | 0.080 | 0.138 | P100 |
| | 0.535 | 0.029 | 0.722 | 0.535 | 0.615 | H100 |
| | 0.714 | 0.175 | 0.643 | 0.714 | 0.677 | G100 |
| | 0.022 | 0.002 | 0.667 | 0.022 | 0.043 | C100 |
| Weighted Avg. | 0.545 | 0.117 | 0.627 | 0.545 | 0.522 | |

| | a | b | c | d | e | f | g | h | i | j | |
|-----|----|-----|-----|-----|----|----|---|-----|-----|---|-------|
| 156 | 3 | 3 | 2 | 11 | 2 | 0 | 0 | 0 | 11 | 0 | Y100 |
| 7 | 43 | 0 | 3 | 5 | 0 | 0 | 0 | 2 | 6 | 0 | W100 |
| 8 | 0 | 118 | 2 | 6 | 0 | 0 | 0 | 3 | 0 | 0 | U100 |
| 76 | 18 | 9 | 321 | 136 | 15 | 0 | 6 | 6 | 124 | 0 | T100 |
| 104 | 7 | 21 | 31 | 501 | 8 | 1 | 8 | 8 | 46 | 3 | SM100 |
| 24 | 1 | 10 | 1 | 4 | 27 | 0 | 1 | 1 | 5 | 0 | Q100 |
| 46 | 2 | 6 | 2 | 22 | 13 | 10 | 7 | 7 | 16 | 1 | P100 |
| 35 | 3 | 12 | 6 | 28 | 0 | 6 | 6 | 130 | 23 | 0 | H100 |
| 111 | 15 | 18 | 8 | 28 | 4 | 2 | 2 | 14 | 500 | 0 | G100 |
| 75 | 6 | 17 | 11 | 177 | 12 | 1 | 1 | 9 | 47 | 8 | C100 |

Table 5.45 Confusion matrix of G-metric classifier with minimum regression slope criterion.

Results of classification - Training set

```

Correctly Classified Instances      2030      60.9610%
Incorrectly Classified Instances    1300      39.0390%
Kappa statistic                    0.5345
Total number of instances:         3330
=== Detailed Accuracy By Class ===

```

| | TP Rate | FP Rate | Precision | Recall | F-Measure | Class |
|---------------|---------|---------|-----------|--------|-----------|-------|
| | 0.8617 | 0.1796 | 0.2837 | 0.8617 | 0.4269 | Y100 |
| | 0.7727 | 0.0227 | 0.5258 | 0.7727 | 0.6258 | W100 |
| | 0.8759 | 0.0402 | 0.6000 | 0.8759 | 0.7122 | U100 |
| | 0.5489 | 0.0267 | 0.8958 | 0.5489 | 0.6807 | T100 |
| | 0.7548 | 0.1979 | 0.6015 | 0.7548 | 0.6695 | SM10 |
| | 0.4795 | 0.0221 | 0.4375 | 0.4795 | 0.4575 | Q100 |
| | 0.1520 | 0.0020 | 0.8261 | 0.1520 | 0.2568 | P100 |
| | 0.6132 | 0.0294 | 0.7233 | 0.6132 | 0.6637 | H100 |
| | 0.7771 | 0.1435 | 0.6860 | 0.7771 | 0.7287 | G100 |
| | 0.0331 | 0.0000 | 1.0000 | 0.0331 | 0.0640 | C100 |
| Weighted Avg. | 0.6096 | 0.0942 | 0.7192 | 0.6096 | 0.5850 | |

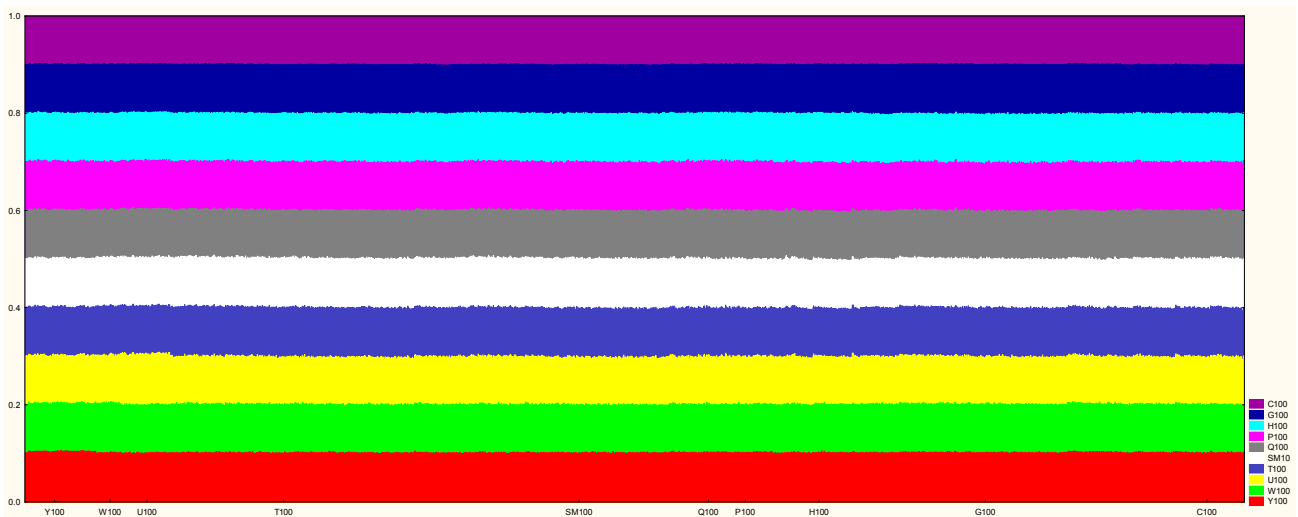


Figure 5.55 Graph of predictions by G-metric classifier with minimum regression slope criterion on the training *generalDS* set.

G-metric classifier run on general dataset as described in section 5.4.10 with voting criterion of minimal regression slope calculated across the whole breed groups in training dataset reported 54.48 % of correctly classified instances for 10 fold cross validation. Kappa statistic was calculated as 0.4571, overall FP Rate=0.117, Precision=0.627 and F-Measure=0.522.

Best classified class by the meaning of TP Rate was U100 (0.861), the worst results obtained in TP Rate were reached for C100 class (0.022). FP Rate ranged from 0.002 (C100) to 0.241 in SM100 class. These results show that classifier is not able to divide C100 class against the others properly, or C100 class could not be defined by used classifier (table 5.45). Precision ranged from 0.243 for Y100 class (F-Measure=0.376) to 0.829 for T100 class (F-Measure=0.588).

As classifier uses different voting method (criterion) for choosing final class of prediction, it reached only 60.96 % of correctly classified instances on the training set. As well, because all of classes are the most probable (probability of belonging of unknown individual to all of classes is equal on the beginning of classification and it is changed in according to results of voting), graph of predictions on the whole training set (figure 5.55) shows, how classifier decides about unknown classified instances. Only small changes then between classes probabilities can caused final decision about individual.

5.4.10.3 General Dataset - IB1 implementation

Results of classification - 10 CV

```

Correctly Classified Instances      1768      53.0931%
Incorrectly Classified Instances    1562      46.9069%
Kappa statistic                    0.4406
Total number of instances:         3330
    
```

=== Detailed Accuracy By Class ===

| | TP Rate | FP Rate | Precision | Recall | F-Measure | Class |
|---------------|---------|---------|-----------|--------|-----------|-------|
| | 0.3989 | 0.0359 | 0.5435 | 0.3989 | 0.4601 | Y100 |
| | 0.4242 | 0.0136 | 0.5385 | 0.4242 | 0.4746 | W100 |
| | 0.7518 | 0.0280 | 0.6821 | 0.7518 | 0.7153 | U100 |
| | 0.6652 | 0.2203 | 0.5610 | 0.6652 | 0.6087 | T100 |
| | 0.5096 | 0.2166 | 0.4908 | 0.5096 | 0.5000 | SM100 |
| | 0.1781 | 0.0107 | 0.4063 | 0.1781 | 0.2476 | Q100 |
| | 0.2320 | 0.0355 | 0.3118 | 0.2320 | 0.2661 | P100 |
| | 0.3868 | 0.0429 | 0.5562 | 0.3868 | 0.4563 | H100 |
| | 0.6943 | 0.1824 | 0.6295 | 0.6943 | 0.6603 | G100 |
| | 0.2727 | 0.1211 | 0.3009 | 0.2727 | 0.2861 | C100 |
| Weighted Avg. | 0.5309 | 0.1538 | 0.5221 | 0.5309 | 0.5220 | |

| a | b | c | d | e | f | g | h | i | j | |
|----|----|-----|-----|-----|----|----|----|-----|-----|-------|
| 75 | 1 | 1 | 22 | 25 | 2 | 6 | 4 | 32 | 20 | Y100 |
| 2 | 28 | 0 | 16 | 2 | 0 | 0 | 1 | 15 | 2 | W100 |
| 1 | 1 | 103 | 10 | 10 | 0 | 0 | 4 | 8 | 0 | U100 |
| 7 | 6 | 7 | 469 | 100 | 1 | 8 | 12 | 63 | 32 | T100 |
| 11 | 2 | 10 | 148 | 372 | 3 | 13 | 15 | 49 | 107 | SM100 |
| 7 | 0 | 6 | 11 | 14 | 13 | 6 | 1 | 7 | 8 | Q100 |
| 10 | 0 | 2 | 19 | 22 | 6 | 29 | 8 | 19 | 10 | P100 |
| 4 | 3 | 10 | 20 | 38 | 0 | 12 | 94 | 45 | 17 | H100 |
| 11 | 8 | 6 | 66 | 55 | 2 | 11 | 21 | 486 | 34 | G100 |
| 10 | 3 | 6 | 55 | 120 | 5 | 8 | 9 | 48 | 99 | C100 |

Table 5.46 Confusion matrix of G-metric classifier with IB1 implementation of voting criterion.

Results of classification - Training set

```

Correctly Classified Instances      3327          99.9099%
Incorrectly Classified Instances    3              0.0901%
Kappa statistic                    0.9989
Total number of instances:        3330
    
```

=== Detailed Accuracy By Class ===

| | TP Rate | FP Rate | Precision | Recall | F-Measure | Class |
|---------------|---------|---------|-----------|--------|-----------|-------|
| | 1.000 | 0.000 | 1.000 | 1.000 | 1.000 | Y100 |
| | 1.000 | 0.000 | 1.000 | 1.000 | 1.000 | W100 |
| | 1.000 | 0.000 | 1.000 | 1.000 | 1.000 | U100 |
| | 1.000 | 0.001 | 0.997 | 1.000 | 0.999 | T100 |
| | 1.000 | 0.000 | 0.999 | 1.000 | 0.999 | SM10 |
| | 1.000 | 0.000 | 1.000 | 1.000 | 1.000 | Q100 |
| | 1.000 | 0.000 | 1.000 | 1.000 | 1.000 | P100 |
| | 0.996 | 0.000 | 1.000 | 0.996 | 0.998 | H100 |
| | 0.999 | 0.000 | 1.000 | 0.999 | 0.999 | G100 |
| | 0.997 | 0.000 | 1.000 | 0.997 | 0.999 | C100 |
| Weighted Avg. | 0.999 | 0.000 | 0.999 | 0.999 | 0.999 | |

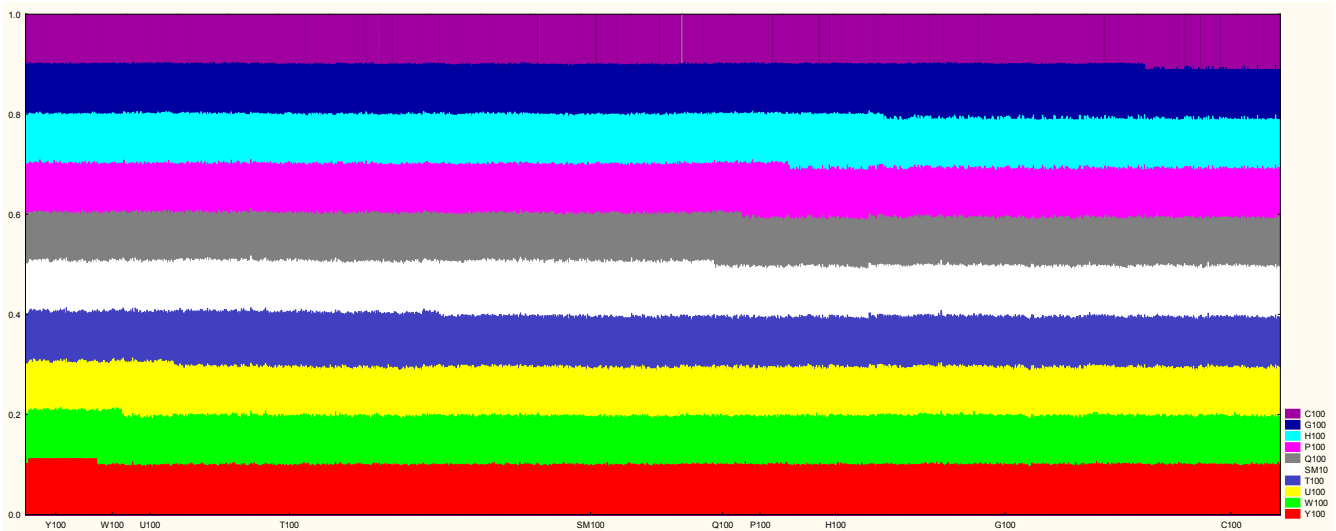


Figure 5.56 Graph of predictions by G-metric classifier with IB1 voting criterion on the training *generalDS* set.



Figure 5.57 Graph of G-metric IB1 model prediction on unknown samples.

G-metric classifier implemented like IB1 classifier (with exception of 1-NN voting algorithm; all of classes of unknown individual has the same probabilities on the beginning) was run on the general dataset with following results of 10 fold cross validation - 53.09 % of correctly classified instances, Kappa statistic=0.4406, overall FP Rate=0.1536, Precision=0.5221, F-Measure=0.5220.

The best TP rate was reached for U100 class (0.7518), the worst one for Q100 class (0.1781). FP Rate ranged from 0.0136 in W100 to 0.1824 in G100. The best Precision value was observed for G100 class (0.6295; F-Measure=0.6603), the worst one for C100 class (0.3009; F-Measure=0.2861).

Figure 5.56 shows predictions of probabilities for the whole data set done by G-metric IB1 classifier. It is evident, that in this case, classifier is able to classify training set as classi-

cal IB1 one, i.e. ~100 % of correctly classified instances. Only where genotype data were not available (3 cases) for particular breed, individuals were misclassified (probabilities equaled, so classifier could not classify them then). On the figure 5.57, it can be inspected detailed, how probabilities are given to individuals.

5.4.11 Algorithms results summary

| | Correctly classified instances % | Kappa statistic | FP Rate | Precision | F-Measure |
|---------------------------------|---|--------------------|---------|-----------|-----------|
| General dataset | | | | | |
| <i>ZeroR</i> | 21.922 | 0.000 | 0.219 | 0.048 | 0.079 |
| J48 | 53.333 | 0.436 | 0.095 | 0.513 | 0.518 |
| Jrip | 52.793 | 0.417 | 0.116 | 0.546 | 0.507 |
| Naive Bayes | 82.553 | 0.789 | 0.037 | 0.828 | 0.818 |
| Bayes Net | 83.664 | 0.804 | 0.032 | 0.837 | 0.833 |
| IB1 | 60.030 | 0.516 | 0.084 | 0.587 | 0.588 |
| IB5 | 65.375 | 0.570 | 0.089 | 0.680 | 0.605 |
| SMO | 78.348 | 0.740 | 0.040 | 0.780 | 0.779 |
| Vote | 83.003 | 0.795 | 0.036 | 0.831 | 0.825 |
| G-metric Slope | 54.475 | 0.457 | 0.117 | 0.627 | 0.522 |
| G-metric IB1 | 53.093 | 0.441 | 0.154 | 0.522 | 0.522 |
| Allele length dataset | | | | | |
| J48 | 58.949 | 0.506 | 0.079 | 0.578 | 0.581 |
| Jrip | 61.682 | 0.536 | 0.082 | 0.626 | 0.611 |
| Naive Bayes | 56.126 | 0.475 | 0.082 | 0.553 | 0.552 |
| Bayes Net | 84.354 | 0.813 | 0.029 | 0.842 | 0.842 |
| IB1 | 47.688 | 0.367 | 0.110 | 0.466 | 0.467 |
| IB5 | 49.880 | 0.385 | 0.117 | 0.490 | 0.471 |
| SMO | 50.751 | 0.394 | 0.118 | 0.523 | 0.479 |
| Vote | 81.351 | 0.776 | 0.036 | 0.810 | 0.808 |
| Allele frequency dataset | | | | | |
| J48 | 56.757 | 0.481 | 0.080 | 0.559 | 0.562 |
| Jrip | 61.652 | 0.536 | 0.080 | 0.621 | 0.610 |
| Naive Bayes | 58.378 | 0.499 | 0.085 | 0.571 | 0.572 |
| Bayes Net | 84.805 | 0.818 | 0.029 | 0.847 | 0.847 |
| IB1 | 43.784 | 0.318 | 0.119 | 0.420 | 0.424 |
| IB5 | 45.405 | 0.329 | 0.127 | 0.429 | 0.419 |

5 Results and Discussion

| | Correctly classified instances % | Kappa statistic | FP Rate | Precision | F-Measure |
|---------------------------------|---|--------------------|---------|-----------|-----------|
| SMO | 49.159 | 0.374 | 0.122 | 0.472 | 0.461 |
| Vote | 81.982 | 0.783 | 0.037 | 0.818 | 0.815 |
| Summary | | | | | |
| General dataset | | | | | |
| min | 52.793 | 0.417 | 0.032 | 0.513 | 0.507 |
| max | 83.664 | 0.804 | 0.154 | 0.837 | 0.833 |
| mean | 66.667 | 0.596 | 0.080 | 0.675 | 0.652 |
| Allele length dataset | | | | | |
| min | 47.688 | 0.367 | 0.029 | 0.466 | 0.467 |
| max | 84.354 | 0.813 | 0.118 | 0.842 | 0.842 |
| mean | 61.348 | 0.531 | 0.082 | 0.611 | 0.601 |
| Allele frequency dataset | | | | | |
| min | 43.784 | 0.318 | 0.029 | 0.420 | 0.419 |
| max | 84.805 | 0.818 | 0.127 | 0.847 | 0.847 |
| mean | 60.240 | 0.517 | 0.085 | 0.592 | 0.589 |
| Overall | | | | | |
| min | 43.784 | 0.318 | 0.029 | 0.420 | 0.419 |
| max | 84.805 | 0.818 | 0.154 | 0.847 | 0.847 |
| mean | 63.053 | 0.552 | 0.082 | 0.630 | 0.617 |

Table 5.47 Summary results of classification results.

Summary results of classification algorithms displayed in table 5.47 are discussed in this chapter. Table 5.47 summarizes results obtained for the best reached parameters set for each classification method run on all of three datasets. Also, it contains summary average results for all of datasets as well as for all of algorithms across datasets and for all of observed classification parameters (Correctly classified instances, Kappa statistic, FP Rate, Precision and F-Measure) obtained by 10 fold cross validation.

As ZeroR as basic classifier usually used as a baseline for comparison of classification power of the others methods reported 21.922 % of correctly classified instances and related measures in accordance with this one, it can be said that none of the others classifiers do not show worse results as a basic one - ZeroR. So, all of classification methods are able to classify unknown individuals to their proper breed better than a random one based on the most frequent class.

At all, 63.053 % of correctly classified instances, Kappa-statistic=0.552, FP Rate=0.082, Precision=0.630 and F-Measure=0.617 were reached across all of classification method and across all of datasets examined. These results shows that classification method selected are able to classify unknown individuals (as results are based on 10 fold cross validation) to their breeds with the power better than basic classifier (which classify to the most frequent class) has.

5 Results and Discussion

Bayes Net algorithm gave the best results in all of datasets examined. The best reached result was obtained for Bayes Net classifier in allele frequency dataset (84.805 % of correctly classified instances, Kappa statistic=0.818, FP Rate=0.029, Precision=0.847, F-Measure=0.847). Similar results were obtained for Bayes Net classifier in allele length and general dataset (84.354 and 83.664 percents of correctly classified instances, see other parameters reached in table 5.47). On the other hand, worst results were obtained in allele frequency dataset as well for IB1 classifier (43.784 % of correctly classified instances, Kappa statistic=0.318, FP Rate=0.119, Precision=0.420 and F-Measure=0.424).

Lets choose percentage of correctly classified instances as a main attribute which says about classification power of particular method. Then, in general dataset results we can establish following ordering of classification method (from the best one to the worst) - Bayes Net, Voted classifier, Naive Bayes, SMO, IB5, IB1, G-metric Slope, J48, G-Metric IB1, JRip. For allele length dataset results, the order is following - Bayes Net, Voted classifier, JRip, J48, Naive Bayes, SMO, IB5, IB1. For allele frequency dataset, we can sort classifiers according to percents of correctly classified instances as follows - Bayes Net, Voted classifier, JRip, Naive Bayes, J48, SMO, IB5, IB1.

| | Correctly classified instances | Kappa statistic | FP Rate | Precision | F-Measure |
|----------------|--------------------------------|-----------------|---------|-----------|-----------|
| J48 | 56.346 | 0.474 | 0.085 | 0.550 | 0.554 |
| Jrip | 58.709 | 0.497 | 0.093 | 0.598 | 0.576 |
| Naive Bayes | 65.686 | 0.588 | 0.068 | 0.651 | 0.647 |
| Bayes Net | 84.274 | 0.811 | 0.030 | 0.842 | 0.841 |
| IB1 | 50.501 | 0.400 | 0.104 | 0.491 | 0.493 |
| IB5 | 53.554 | 0.428 | 0.111 | 0.533 | 0.498 |
| SMO | 59.419 | 0.503 | 0.093 | 0.592 | 0.573 |
| Vote | 82.112 | 0.785 | 0.036 | 0.820 | 0.816 |
| G-metric Slope | 54.475 | 0.457 | 0.117 | 0.627 | 0.522 |
| G-metric IB1 | 53.093 | 0.441 | 0.154 | 0.522 | 0.522 |

Table 5.48 Summary average results of classification power of selected classifiers calculated as means across all of datasets.

5 Results and Discussion

In average, Bayes Net seems to be the most useful method for classification of cattle breeds based on microsatellite genotype data - Bayes Net reached average percentage of correctly classified instances equals 84.274 across all of datasets (table 5.48, figure 5.58), Kappa statistic=0.811, FP Rate=0.030, Precision=0.842 and F-Measure=0.841. On the contrary, IB1, IB 5, JRip and J48 methods do not show acceptable results as well as G-metric Slope and G-metric IB1 algorithms (otherwise, their power will be shown later and they are not fully comparable to common classification methods as they are set up for genetic data mainly under SMM). Voted classifier has obtained very good results, later will be shown, it can be more robust than well-performed Bayes Net methods.

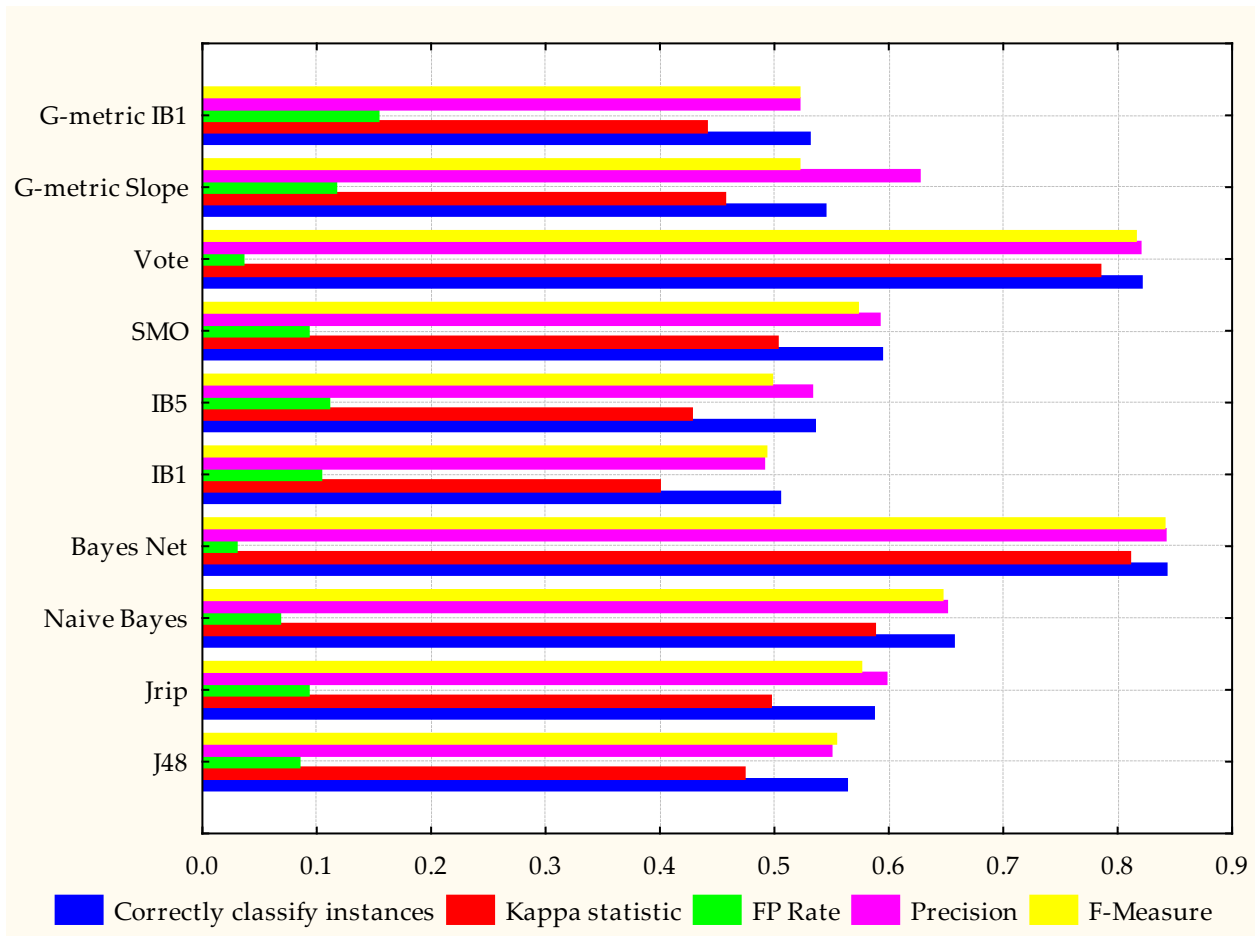


Figure 5.58 Graph of average results of classification power of selected classifiers calculated as means across all of datasets.

Table 5.47 and figure 5.58 show as well that results of classifiers can not be generalized across datasets. For example SMO classifier has good results (78.348 % of correctly classified instances) on general dataset, reduced significantly for allele length and frequency datasets (50.751 %, 49.159 %) as will be discussed later.

Anyway, Bayes Net seems to be best performed classifier suitable to discriminate individuals on the base of their genotype data among all of algorithms examined.

5.4.12 Datasets usability for breed discrimination

As we have results calculated as average values shown in table 5.47 for all of three datasets, and if we assume that all of classifiers reached better than ZeroR classification parameters, we can discuss usability of all datasets used. ZeroR results are not included in calculations performed across datasets. Average results reached as mean of all three datasets has shown, that all of selected classifiers at all has power better than basis one (ZeroR) to classify cattle breeds successfully based on genotype microsatellite data - average percentage of correctly classified instances equals 63.053 - and they are not hardly dependent on type of given dataset, but more dependent on particular classification method.

When we take a look to results in table 5.47 reached in average for all of three datasets, best results for percentage of correctly classified instances calculated across all of methods in each dataset were reached in general dataset (66.667 %), then in allele length dataset (61.348 %), then in allele frequency dataset (60.240 %). These results reflect fact, that all of algorithms except ZeroR are used for these calculations, so results are highly dependent on algorithms selected. This can be easily seen, when the best classification results across all of datasets and algorithms were reached in allele frequency dataset.

On the general dataset, in the meaning of percentage of correctly classified instances, best results were reached by Bayes Net classifier (83.664 %). Bayes Net classifier reached best results as well by using another datasets (84.354 % in allele length dataset, 84.805 % in allele frequency dataset).

The worst performing classifier in general dataset was JRip one with only 52.793 % of correctly classified instances. In allele length and allele frequency datasets, IB1 classifier showed the worst results of all examined methods (47.688 resp. 43.784 %). As was mentioned, the reason of not enough information provided for lazy based learning algorithms for successful classification in case of the both datasets and number of individuals in training sets could cause these results.

At all, algorithms based on frequency and probability data (like Bayes Net, Vote classifier) seem to be robust and independent on genetic data type and can reach acceptable results on more detailed datasets (like allele frequency and allele lengths ones are). In opposite, classification power of lazy based (IB1 and IB5) or data space dividing based (SMO, J48, JRip) algorithms are highly dependent on amount of information to deal with (number of attributes) plus amount of individuals in training sets. As results reached on training sets show, when we give enough information to these algorithm (in the meaning of size of dataset) needed to classify detailed datasets with many attributes, they can perform very well with good prediction results. This indicates the way of their usage on large genotype datasets obtained in routine labs, when this hypothesis will be proved.

5.4.13 Comparison with results in horses

As results of classification algorithms performed on 932 unrelated individuals of 8 breeds for 17 MS loci genotype data (similar to general dataset) were published by Burócziová and Říha (2009), we can discuss results obtained in this paper with results obtained in this thesis. We need just to note, that 10 MS loci genotype dataset was used for this purpose, as well as couple of classification methods were added for this work purposes. Only Naive Bayes, Bayes Net, IB1, IB5, J48 and JRip algorithms implemented in Weka-3-6-6 were examined in the given study.

5 Results and Discussion

Bayes Net algorithm was also evaluated as the best one in the paper - 88 percents of correctly classified instances, Kappa statistic=0.86 (in comparison with 83.664 % in cattle general dataset, Kappa statistic=0.804), then Naive Bayes with 87 % of correctly classified instances and Kappa statistic equals to 0.84 (82.553 % for cattle general dataset, Kappa statistic=0.84), then IB1 classifier with 87 % of correctly classified instances and Kappa statistic=0.84 (60.030 % of correctly classified instances, Kappa statistic=0.516), JRip algorithm (67 % of correctly classified instances, Kappa statistic=0.59) in comparison with 52.793 % of correctly classified instances and Kappa statistic=0.417 in cattle, IB5 classifier with 65 % of correctly classified instances, Kappa statistic=0.56 reched in horses discrimination in comparison with 65.375 % of correctly classified instances, Kappa statistic=0.570 in cattle, J48 with 59 % of correctly classified instances and Kappa statistic=0.51 in horse and 53.333 % of correctly classified instances, Kappa statistic=0.436 in cattle.

In the contrary, IB1 classifier failed in cattle breed discrimination in comparison with horses breed discrimination based on 17 loci genotype data. As we discuss previously, it is evident that 10 MS loci data are not enough within thousands of individual to find proper similar example in model database, which can classify properly unknown individual. This is more evident on significantly worse result of IB1 and IB5 methods when they are used for allele frequency and allele lenght datasets as was discussed above.

Anyway, results of classification algorithms are comparable in horse and cattle, and does not depend highly on number of individuals (932 in horses, 3300 in cattle) as well as on number of loci included in dataset (17 in horses, 10 in cattle). It is quite important that results compared showed that the same set of algorithms is usable for two different animal species and does not depend on different breeding strategies in both of animal species as well as on data type (10 vs. 17 MS genotype data). Also, Bayes Net as one of best performed classification algorithm gave best and comparable results for both of species and datasets.

5.4.14 Comments on Diversity

| | Y100 | W100 | U100 | T100 | SM100 | Q100 | P100 | H100 | G100 | C100 |
|------------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| General dataset | | | | | | | | | | |
| <i>ZeroR</i> | 0.000 | 0.000 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| J48 | 0.277 | 0.439 | 0.518 | 0.711 | 0.610 | 0.110 | 0.072 | 0.527 | 0.634 | 0.245 |
| Jrip | 0.096 | 0.379 | 0.569 | 0.687 | 0.652 | 0.192 | 0.136 | 0.502 | 0.696 | 0.102 |
| Naive Bayes | 0.771 | 0.682 | 0.905 | 0.916 | 0.892 | 0.397 | 0.408 | 0.790 | 0.959 | 0.537 |
| Bayes Net | 0.809 | 0.758 | 0.920 | 0.915 | 0.870 | 0.562 | 0.480 | 0.802 | 0.960 | 0.579 |
| IB1 | 0.330 | 0.500 | 0.708 | 0.748 | 0.659 | 0.274 | 0.192 | 0.444 | 0.756 | 0.325 |
| IB5 | 0.287 | 0.288 | 0.766 | 0.901 | 0.838 | 0.014 | 0.048 | 0.428 | 0.847 | 0.132 |
| SMO | 0.723 | 0.742 | 0.854 | 0.887 | 0.811 | 0.562 | 0.456 | 0.654 | 0.920 | 0.521 |
| Vote | 0.787 | 0.727 | 0.912 | 0.915 | 0.881 | 0.479 | 0.456 | 0.790 | 0.959 | 0.551 |
| G-metric Slope | 0.830 | 0.652 | 0.961 | 0.455 | 0.686 | 0.370 | 0.080 | 0.535 | 0.714 | 0.022 |
| G-metric IB1 | 0.862 | 0.773 | 0.876 | 0.549 | 0.755 | 0.480 | 0.152 | 0.613 | 0.771 | 0.033 |

5 Results and Discussion

| | Y100 | W100 | U100 | T100 | SM100 | Q100 | P100 | H100 | G100 | C100 |
|---------------------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| Allele length dataset | | | | | | | | | | |
| J48 | 0.367 | 0.394 | 0.620 | 0.695 | 0.655 | 0.178 | 0.184 | 0.510 | 0.773 | 0.309 |
| Jrip | 0.399 | 0.470 | 0.730 | 0.681 | 0.708 | 0.425 | 0.328 | 0.638 | 0.743 | 0.287 |
| Naive Bayes | 0.431 | 0.470 | 0.788 | 0.715 | 0.536 | 0.274 | 0.160 | 0.539 | 0.679 | 0.298 |
| Bayes Net | 0.809 | 0.924 | 0.905 | 0.919 | 0.829 | 0.658 | 0.632 | 0.798 | 0.954 | 0.634 |
| IB1 | 0.346 | 0.152 | 0.672 | 0.603 | 0.508 | 0.096 | 0.136 | 0.317 | 0.611 | 0.264 |
| IB5 | 0.410 | 0.152 | 0.708 | 0.695 | 0.558 | 0.027 | 0.072 | 0.272 | 0.639 | 0.154 |
| SMO | 0.346 | 0.015 | 0.672 | 0.682 | 0.593 | 0.164 | 0.184 | 0.358 | 0.646 | 0.121 |
| Vote | 0.761 | 0.833 | 0.898 | 0.921 | 0.807 | 0.575 | 0.528 | 0.786 | 0.943 | 0.526 |
| Allele frequency dataset | | | | | | | | | | |
| J48 | 0.351 | 0.470 | 0.613 | 0.661 | 0.638 | 0.247 | 0.216 | 0.486 | 0.739 | 0.267 |
| Jrip | 0.356 | 0.485 | 0.664 | 0.691 | 0.712 | 0.384 | 0.288 | 0.584 | 0.766 | 0.314 |
| Naive Bayes | 0.399 | 0.348 | 0.774 | 0.746 | 0.614 | 0.288 | 0.208 | 0.535 | 0.700 | 0.273 |
| Bayes Net | 0.814 | 0.894 | 0.912 | 0.928 | 0.837 | 0.685 | 0.704 | 0.790 | 0.951 | 0.623 |
| IB1 | 0.191 | 0.227 | 0.650 | 0.550 | 0.460 | 0.027 | 0.144 | 0.263 | 0.611 | 0.226 |
| IB5 | 0.234 | 0.258 | 0.657 | 0.633 | 0.515 | 0.000 | 0.064 | 0.206 | 0.647 | 0.077 |
| SMO | 0.309 | 0.000 | 0.569 | 0.645 | 0.595 | 0.123 | 0.104 | 0.395 | 0.650 | 0.107 |
| Vote | 0.761 | 0.818 | 0.905 | 0.929 | 0.822 | 0.548 | 0.568 | 0.782 | 0.947 | 0.523 |
| Summary | | | | | | | | | | |
| General dataset | | | | | | | | | | |
| min | 0.096 | 0.288 | 0.518 | 0.455 | 0.610 | 0.014 | 0.048 | 0.428 | 0.634 | 0.022 |
| max | 0.862 | 0.773 | 0.961 | 0.916 | 0.892 | 0.562 | 0.480 | 0.802 | 0.960 | 0.579 |
| mean | 0.577 | 0.594 | 0.799 | 0.768 | 0.765 | 0.344 | 0.248 | 0.609 | 0.822 | 0.305 |
| Allele length dataset | | | | | | | | | | |
| min | 0.346 | 0.015 | 0.620 | 0.603 | 0.508 | 0.027 | 0.072 | 0.272 | 0.611 | 0.121 |
| max | 0.809 | 0.924 | 0.905 | 0.921 | 0.829 | 0.658 | 0.632 | 0.798 | 0.954 | 0.634 |
| mean | 0.484 | 0.426 | 0.749 | 0.739 | 0.649 | 0.300 | 0.278 | 0.527 | 0.749 | 0.324 |
| Allele frequency dataset | | | | | | | | | | |
| min | 0.191 | 0.000 | 0.569 | 0.550 | 0.460 | 0.000 | 0.064 | 0.206 | 0.611 | 0.077 |
| max | 0.814 | 0.894 | 0.912 | 0.929 | 0.837 | 0.685 | 0.704 | 0.790 | 0.951 | 0.623 |
| mean | 0.427 | 0.438 | 0.718 | 0.723 | 0.649 | 0.288 | 0.287 | 0.505 | 0.751 | 0.301 |
| Overall | | | | | | | | | | |
| min | 0.096 | 0.000 | 0.518 | 0.455 | 0.460 | 0.000 | 0.048 | 0.206 | 0.611 | 0.022 |
| max | 0.862 | 0.924 | 0.961 | 0.929 | 0.892 | 0.685 | 0.704 | 0.802 | 0.960 | 0.634 |
| mean | 0.502 | 0.494 | 0.759 | 0.745 | 0.694 | 0.313 | 0.269 | 0.552 | 0.778 | 0.310 |

Table 5.49 Summary of classes TP rates.

5 Results and Discussion

| | Y100 | W100 | U100 | T100 | SM100 | Q100 | P100 | H100 | G100 | C100 |
|---------------------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| General dataset | | | | | | | | | | |
| <i>ZeroR</i> | 0.000 | 0.000 | 0.000 | 0.000 | 0.219 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| J48 | 0.297 | 0.509 | 0.490 | 0.655 | 0.156 | 0.242 | 0.167 | 0.538 | 0.570 | 0.403 |
| Jrip | 0.367 | 0.610 | 0.639 | 0.679 | 0.349 | 0.500 | 0.531 | 0.689 | 0.694 | 0.363 |
| Naive Bayes | 0.797 | 0.918 | 0.655 | 0.672 | 0.736 | 0.879 | 0.879 | 0.914 | 0.919 | 0.654 |
| Bayes Net | 0.822 | 0.909 | 0.869 | 0.885 | 0.761 | 0.891 | 0.811 | 0.915 | 0.929 | 0.644 |
| IB1 | 0.488 | 0.647 | 0.719 | 0.634 | 0.554 | 0.455 | 0.358 | 0.568 | 0.718 | 0.423 |
| IB5 | 0.701 | 0.950 | 0.778 | 0.631 | 0.547 | 0.500 | 0.600 | 0.897 | 0.764 | 0.706 |
| SMO | 0.638 | 0.875 | 0.760 | 0.836 | 0.750 | 0.732 | 0.633 | 0.791 | 0.902 | 0.612 |
| Vote | 0.809 | 0.923 | 0.868 | 0.878 | 0.745 | 0.897 | 0.826 | 0.914 | 0.922 | 0.651 |
| G-metric Slope | 0.243 | 0.439 | 0.551 | 0.829 | 0.546 | 0.333 | 0.500 | 0.722 | 0.643 | 0.667 |
| G-metric IB1 | 0.544 | 0.539 | 0.682 | 0.561 | 0.491 | 0.406 | 0.312 | 0.556 | 0.630 | 0.301 |
| Allele length dataset | | | | | | | | | | |
| J48 | 0.367 | 0.481 | 0.680 | 0.679 | 0.568 | 0.245 | 0.258 | 0.603 | 0.713 | 0.386 |
| Jrip | 0.457 | 0.660 | 0.685 | 0.729 | 0.481 | 0.574 | 0.506 | 0.749 | 0.743 | 0.523 |
| Naive Bayes | 0.474 | 0.337 | 0.527 | 0.632 | 0.528 | 0.328 | 0.364 | 0.478 | 0.671 | 0.478 |
| Bayes Net | 0.792 | 0.897 | 0.867 | 0.886 | 0.784 | 0.716 | 0.782 | 0.886 | 0.952 | 0.687 |
| IB1 | 0.428 | 0.270 | 0.702 | 0.533 | 0.444 | 0.241 | 0.183 | 0.478 | 0.567 | 0.284 |
| IB5 | 0.385 | 0.370 | 0.713 | 0.493 | 0.442 | 0.286 | 0.300 | 0.606 | 0.574 | 0.438 |
| SMO | 0.442 | 1.000 | 0.657 | 0.529 | 0.445 | 0.750 | 0.479 | 0.621 | 0.524 | 0.473 |
| Vote | 0.781 | 0.859 | 0.759 | 0.851 | 0.748 | 0.737 | 0.776 | 0.799 | 0.922 | 0.697 |
| Allele frequency dataset | | | | | | | | | | |
| J48 | 0.357 | 0.470 | 0.604 | 0.653 | 0.572 | 0.265 | 0.262 | 0.551 | 0.711 | 0.324 |
| Jrip | 0.482 | 0.478 | 0.607 | 0.724 | 0.498 | 0.519 | 0.456 | 0.703 | 0.776 | 0.496 |
| Naive Bayes | 0.528 | 0.426 | 0.602 | 0.615 | 0.554 | 0.368 | 0.426 | 0.575 | 0.699 | 0.396 |
| Bayes Net | 0.814 | 0.908 | 0.893 | 0.883 | 0.805 | 0.769 | 0.815 | 0.893 | 0.941 | 0.663 |
| IB1 | 0.310 | 0.227 | 0.685 | 0.470 | 0.469 | 0.111 | 0.220 | 0.376 | 0.473 | 0.271 |
| IB5 | 0.314 | 0.266 | 0.573 | 0.437 | 0.475 | 0.000 | 0.216 | 0.568 | 0.482 | 0.318 |
| SMO | 0.464 | 0.000 | 0.645 | 0.495 | 0.473 | 0.450 | 0.406 | 0.623 | 0.492 | 0.333 |
| Vote | 0.808 | 0.947 | 0.827 | 0.841 | 0.749 | 0.851 | 0.845 | 0.848 | 0.912 | 0.669 |
| Summary | | | | | | | | | | |
| General dataset | | | | | | | | | | |
| min | 0.243 | 0.439 | 0.490 | 0.561 | 0.156 | 0.242 | 0.167 | 0.538 | 0.570 | 0.301 |
| max | 0.822 | 0.950 | 0.869 | 0.885 | 0.761 | 0.897 | 0.879 | 0.915 | 0.929 | 0.706 |
| mean | 0.571 | 0.732 | 0.701 | 0.726 | 0.563 | 0.584 | 0.562 | 0.750 | 0.769 | 0.542 |

5 Results and Discussion

| | Y100 | W100 | U100 | T100 | SM100 | Q100 | P100 | H100 | G100 | C100 |
|---------------------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| Allele length dataset | | | | | | | | | | |
| min | 0.367 | 0.270 | 0.527 | 0.493 | 0.442 | 0.241 | 0.183 | 0.478 | 0.524 | 0.284 |
| max | 0.792 | 1.000 | 0.867 | 0.886 | 0.784 | 0.750 | 0.782 | 0.886 | 0.952 | 0.697 |
| mean | 0.516 | 0.609 | 0.699 | 0.667 | 0.555 | 0.485 | 0.456 | 0.653 | 0.708 | 0.496 |
| Allele frequency dataset | | | | | | | | | | |
| min | 0.310 | 0.000 | 0.573 | 0.437 | 0.469 | 0.000 | 0.216 | 0.376 | 0.473 | 0.271 |
| max | 0.814 | 0.947 | 0.893 | 0.883 | 0.805 | 0.851 | 0.845 | 0.893 | 0.941 | 0.669 |
| mean | 0.510 | 0.465 | 0.680 | 0.640 | 0.574 | 0.417 | 0.456 | 0.642 | 0.686 | 0.434 |
| Overall | | | | | | | | | | |
| min | 0.243 | 0.000 | 0.490 | 0.437 | 0.156 | 0.000 | 0.167 | 0.376 | 0.473 | 0.271 |
| max | 0.822 | 1.000 | 0.893 | 0.886 | 0.805 | 0.897 | 0.879 | 0.915 | 0.952 | 0.706 |
| mean | 0.535 | 0.612 | 0.694 | 0.681 | 0.564 | 0.502 | 0.497 | 0.687 | 0.725 | 0.495 |

Table 5.50 Summary of classes precision values.

As classification method can define breed based on genotype data, we can assume that it is well defined by its genetic basis as well. From table 5.49 and table 5.50, it can be seen that TP rates and precision could be discussed as a parameters of genetic-well-definition of breeds within point of view of classification methods.

TP rate reflects portion of true positive classified individuals from true positive classified plus false negative classified number of individuals. So, if there are not a lot of individuals from actual class are classified as another breed and does not mean how many are classified as actual class (TP), TP rate has near 1 value. TP rate is then value which describes, how good is one class described by particular method. It does not reflect, another connections, like if and how many individuals from another classes (breeds) are classified into the actual one. As example - Czech Simmental breed has average TP Rate calculated across all of methods (except ZeroR), across all of datasets equals 0.694 and seems to be very well defined (as it is from this point of view), however big portion of individuals with actual class set to Czech Fleckvieh was classified as Czech Simmental as well across all of methods and datasets (average TP Rate of Czech Fleckvieh class equals 0.310). This really reflects the fact that quite a lot of crossbred individuals (SM100 × C100) were accepted in Czech Fleckvieh breed in past as purebreds and not so much in the opposite.

From this point of view, we can identify following order in TP Rates calculated across all of methods and datasets - Aberdeen Angus (0.778) > Hereford > Charolais > Czech Simmental > Holstein > Limousin > Galloway > Blonde d'Aquitaine > Czech Fleckvieh > Piedmontese (0.269). So, Aberdeen Angus breed can be identified in the best way without false positive results in average of all datasets and algorithms.

When we inspect results in table 5.49, we can see that among all of TP Rates obtained for particular classes, datasets and algorithms, the best result was reached for Aberdeen Angus breed as well in general dataset, within Bayes Net classifier (0.960).

Best results which can describe, how all of breeds can be classified in the best way under TP Rate assumption are: Limousin (G-metric IB1, general dataset, TP Rate=0.862), Galloway (Bayes Net, allele length dataset, TP Rate=0.924), Hereford (G-metric Slope, general dataset, TP Rate=0.961), Charolais (Vote classifier, allele frequency dataset, TP-Rate=0.929), Czech

5 Results and Discussion

Simmental (Naive Bayes, general dataset, TP-Rate=0.892), Blonde d'Aquitaine (Bayes Net, allele frequency dataset, TP-Rate=0.685), Piedmontese (Bayes Net, allele frequency dataset, TP-Rate=0.704), Holstein (Bayes Net, general dataset, TP-Rate=0.802), Aberdeen Angus (Bayes Net, general dataset, TP-Rate=0.960) and Czech Fleckvieh (Bayes Net, allele length dataset, TP-Rate=0.634). Results obtained for Limousin and Hereford breeds, when they were best classified by G-metric method shows, when we see how it is calculated, genetic similarities between individuals in each breed caused by close populations and not large imports and production of breeding animals instead of usage of AI in the past.

Table 5.50 summarizes results of precision obtained for each classes, each algorithm and each dataset. Precision is calculated as a portion of true positive classified instances on the sum of true positive plus false positive results of classification. It means it uses for calculation also individuals from another breeds than actual class which were classified as actual one. Another example can show how it works - as quite a big amount of individuals with actual class set to Czech Simmental are well indentified by classifications methods, there is still near the same amount of false positive identified individuals with another class indentified by the same algorithms, as can be seen when confusion matrices and connected results are analyzed. That is the reason, why only 0.564 precision was reached as the mean value across all sets and algorithms for Czech Simmental class. So, the precision value describes better how the particular breed is defined in comparison with the other breeds present in dataset, when we want to describe probability with which is individual classified as right breed and another breeds individuals are not classify as this breed at one time. In comparison with TP-Rate, which only describes, how good one breed is identify and is not classify as the other one for actual individual breed. So, precision seems to be better "purity" measure of breed, when we are talking about breed diversity.

Following order was obtained for breeds when average precision calculated across datasets and methods is used: Aberdeen Angus (0.725) > Hereford > Holstein > Charolais > Galloway > Czech Simmental > Limousin > Blonde d'Aquitaine > Piedmontese > Czech Fleckvieh (0.495).

Similar to TP-Rate, Aberdeen Angus breed has the best portion of true positive classified individuals and sum of true positive plus false positive individuals (0.725). Also, the other results reflect quite good situation in each breed and breed practice applied in the Czech Republic for all of them.

Following results were obtained as the best ones for precision for each breed, algorithm and dataset: Limousin (Bayes Net, general dataset, Precision=0.822), Galloway (SMO, allele length dataset, Precision=1.000), Hereford (Bayes Net, allele frequency dataset, Precision=0.893), Charolais (Bayes Net, allele length dataset, Precision=0.886), Czech Simmental (Bayes Net, allele frequency dataset, Precision=0.805), Blonde d'Aquitaine (Vote classifier, general dataset, Precision=0.897), Piedmontese (Naive Bayes, general dataset, Precision=0.879), Holstein (Bayes Net, general dataset, Precision=0.915), Aberdeen Angus (Bayes Net, allele length dataset, Precision=0.952) and Czech Fleckvieh (IB5, general dataset, Precision=0.706).

Results show, that all of breeds examined could be classified with very good precision or/and TP-Rate, however, these parameters are highly depend on algorithm and dataset used. For particular issues connected with breed classification of unknown sample, it can be recommended to build and combine classification models according results shown. As well, it is highly recommended to publish results of classification power of models used, so everybody can calculate with probabilities assign to results and can check reliability of results.

6 Summary

This thesis describes usage of microsatellite markers in cattle for specific tasks according to their usage for

- description of genetic diversity in cattle breeds sampled from subpopulations kept in Czech Republic,
- paternity testing - current state of usability within cattle population sampled by routine genetic laboratory,
- routine laboratory data handling by usage specific and new created software application what can resolve daily laboratory routines and specific issues connected,
- discrimination on breed level with usage of machine learning algorithms, with a new one created especially for microsatellite data.

Set of microsatellite markers recommended by ISAG/FAO routinely analyzed for proving of genetic type of breeding cattle is used in all of tasks mentioned above - *BM1824*, *BM2113*, *ETH3*, *ETH10*, *ETH225*, *INRA023*, *SPS115*, *TGLA122*, *TGLA126*, *TGLA227*. Daily routine testing in Lamgen accredited laboratory in years 2002-2009 created following data sets used in thesis: 730 individuals of Czech Simmental, 705 individuals of Charolais, 700 individuals of Aberdeen Angus breed, 363 individuals of Czech Fleckvieh, 243 Holsteins, 188 of Limousin, 137 individuals of Hereford, 125 of Piedmontese, 73 of Blonde d'Aquitaine and 66 individuals of Galloway breed. Also, set of 380 crossbreds randomly selected within the whole database (7776) to explore and characterize genetic diversity of crossbred cattle population in Czech republic and as a basis for comparison in particular thesis aims.

Regarding genetic diversity results calculated as average ones across all loci within breeds, Hereford with major allele frequency equals 0.500, observed and expected heterozygosity as 0.646, 0.637, inbreeding coefficient equals -0.011 and PIC=0.589 was detected as the most uniform breed in set. In contrary, Piedmontese and crossbred dataset were evaluated as the most divergent breeds in set.

When we analyse results of genetic diversity within breeds across loci examined, we can point *TGLA227* as the most divergent locus at all. Interesting results in beef breeds were explored as reduced variability in *ETH3* (Blonde d'Aquitaine), *ETH10* (Charolais, Galloway, Limousin) and *INRA023* (Blonde d'Aquitaine, Hereford) loci in intensive kept beef breeds. This fact reflects selection strategies of breeders as couple of authors mentioned loci above as ones connected with beef yield as genetic markers. In opposite, with results coming from 90's, we realized that intensive breeding breeds, previously alarming evaluated as uniform (or with reduced genetic diversity) like Holstein is, nowadays show very carefully performed breeding strategies in past years which has brought increase of genetic diversity in these breeds in Czech in comparison with older results.

Eight genetic distances (geometric, AIM and SMM based) were calculated between breeds and crossbred data sets. Both, UPGMA and NJ algorithm then were used to show results of these calculations visually. Hereford on the one side and Aberdeen Angus on the other side appeared as the most distinct breeds under assumptions of all genetic distance calculation methods. Each method clustered together Czech Fleckvieh with crossbred dataset what show large portion of Czech Fleckvieh breed used for producing of crossbred animals in both, dairy and beef cattle. As well, Czech Simmental seems to be very closely connected under assumptions of genetic drift, mutations and breeding strategies (which all are displayed compressed in genetic phylograms for particular distance method) to Czech Fleckvieh breed what is in accordance with real state as well.

Generally, based on results of genetic diversity obtained in this work, we can say that large-ly kept breeds in Czech or world-wide (like Czech Fleckvieh, Holstein, Limousin, Blonde d'Aquitaine are), genetically and evolutionary different kept in Czech (Piedmontese) and crossbreeds are more divergent in comparison with minor kept ones, however genetically different (Galloway), and more uniform beef populations like Hereford, Czech Simmental, Charolais and Abredeen Angus are. These results reflect completely breeding strategies for beef and dairy or dual purpose breeds as well as historical development of all breeds included. Additionally, it must be mentioned that by these facts, microsatellite markers are proved as a good tool for exploring and controlling of genetic variability in cattle breeds by selected methods.

All of three paternity exclusion scenarios were calculated across all examined loci and across all of breeds as combined exclusion probabilities as well as polymorphic information content values. As well they were calculated for the whole dataset used in work (3300 individuals of purebreds). For individual breeds, the worst values for all of probabilities calculated were reached for Hereford, the best for Piedmontese breed. This is in accordance with results previously mentioned for genetic variability. Anyway, we can only point on CEP1 calculated for Hereford with value equals 0.953371 what should be alarming when only of 95 percents of individuals in Czech Hereford subpopulation with given parents and known genotypes could be excluded properly. The same alarming results then was obtained for not so commonly tested CEP2 (0.948479). The other values of calculated CEPs as ~ 1.000 can be fully accepted for each breed as well as for the whole dataset. So, results proved that panel of microsatellite loci used for genotyping of dataset in this work fullfills recommendations on paternity exclusion as well as studies of genetic diversity of selected cattle breeds.

Models description, algorithms and interface of the software application created in Borland Delphi 2005 programming environment for routine handling of large microsatellite genotype datasets are proposed in thesis. Firstly, there are models describing typical usage of the system in the meaning of network environment, application design. Then key processes and users were identified and with usage of UML use case diagrams, they were modeled as the basis for implementation.

As the large data sets should be handled by application, SQL database was created for this purpose with design and relationships described by ERD diagrams. Then, there are presented key SQL queries used in software application e.g. for selecting, sorting, filtering of individuals. Queries for specific issues connected with microsatellite genotype data (e.g. calculation of allele or genotype frequencies in loci) also have to be created. This extends normal usage of SQL as database language handling normally with one value attributes.

Basic algorithms used for parsing SQL queries, calculation of allele frequencies and calculations of combined exclusion probabilities, paternity testing and sorting datasets according allele frequencies present in the whole dataset were also created and presented. Security issues are resolved within network architecture and specific algorithms created for logging user and activities in database as well.

Finally, we can see graphical user interface of application with basic description of its usage. Engine which producing protocols in routine laboratory operation is also presented as well as protocols used nowadays. Also, G-metric algorithm for classifying individuals into their breeds with two modifications of final decision metrics which was implemented in the application.

Software application is used by Mendel University accredited genetic laboratory (with updates) from 2009 till now in routine daily regime for automatic processes and dataflows, as well as for paternity testing, protocols issuing etc.

Ten of machine learning algorithms (J48, JRip, Naive Bayes, Bayes Net, IB1, IB5, SMO, Vote classifier, new created and implemented G-metric classifier with Slope and IB1 decision methods) were examined across three different microsatellite genotype datasets to show and discuss their classification power for breed discrimination of individuals in cattle by using microsatellite data.

First of all, results of each algorithm and its classification results are presented and discussed detailed for each dataset. Percentage of correctly classified instances, Kappa statistic, false positive rate, precision and F-Measure are used and discussed as values which can effectively describe classification power of mentioned methods. Also, confusion matrices and graphs of classes probabilities predicted for individuals on training set by each model are used for this purposes. Especially, graphical overview of classification on the whole dataset offer more information about genetic admixture between breeds and can effectively visualize it in comparison with dendrograms. As well, G-metric classifier is specially created for genotype data classification, so it is assuming SMM and can be used for genetic diversity purposes. The usage of these results for genetic diversity issues is described as well.

Results and discussion of usability of all of examined classification algorithms for breed discrimination with regard of datasets used represent following part of work done within thesis. This description shows and can be used in fact for optimal algorithm selection when this type of service should become as commercial one to choose best performing algorithm and data representation.

If we want to evaluate power of algorithms according to dataset used, we can note that algorithms based on frequency and probability principles (like Bayes Net, Vote classifier) seem to be robust and independent on genetic data type and can reach acceptable results on more detailed datasets (like allele frequency and allele lengths ones are). In opposite, classification power of lazy based (IB1 and IB5) or data space dividing based (SMO, J48, JRip) algorithms are highly dependent on amount of information to deal with (number of attributes) plus amount of individuals in training sets.

In average, Bayes Net seems to be the most useful method for classification of cattle breeds based on microsatellite genotype data - Bayes Net reached average percentage of correctly classified instances equals 84.274 across all of datasets, Kappa statistic=0.811, FP Rate=0.030, Precision=0.842 and F-Measure=0.841. These results are highly comparable with results previously presented in horses where subset of algorithms was used to show their classification abilities in horses. Comparison of results show, that they do not depend highly on number of individuals (932 in horses, 3300 in cattle) as well as on number of loci included in dataset (17 in horses, 10 in cattle). It is also important that results that the same set of algorithms is usable for two different animal species and does not depend on different breeding strategies in both of animal species as well as on data type (10 vs. 17 MS genotype data). Also, Bayes Net was evaluated as the best performed classification algorithm for both of species and datasets.

Finally, based on results above, it can be said that cattle breeds can be classified effectively by selected algorithms based on microsatellite genotype data sets used in this work and with selection of proper method and dataset, examined methods can be used for breed discrimination in cattle, otherwise detailed inspection of results for particular breeds is needed and recommended when this issue should become e.g. commercial service.

7 Souhrn

Práce popisuje využití mikrosatelitních markerů skotu v několika specifických úlohách:

- popis genetické diverzity vybraných subpopulací plemen skotu v České republice,
- testování paternity skotu a reálné možnosti jejich využití v současných podmínkách na základě dat získaných v servisní laboratoři,
- získávání genotypových dat a manipulace s nimi v podmínkách rutinního laboratorního provozu pomocí nově vytvořené softwarové aplikace a možnosti pro jejich využití ve specifických provozních podmínkách,
- diskriminace jedinců na základě genetického profilu mikrosatelitních markerů na úrovni plemene za pomoci algoritmů strojového učení.

Ve všech zmíněných úlohách je využita sada mikrosatelitních markerů doporučených ISAG/FAO pro určení tzv. genetického typu skotu. Jedná se o markery *BM1824*, *BM2113*, *ETH3*, *ETH10*, *ETH225*, *INRA023*, *SPS115*, *TGLA122*, *TGLA126*, *TGLA227*. Použitá data byla získána v akreditované laboratoři (Lamgen, ÚMFGŽ, Mendlova Univerzita) v letech 2002–2009. Data byla pořízena v rámci rutinního provozu laboratoře. Datový set tvoří 730 jedinců plemene Masný simentál, 705 jedinců plemene Charolais, 700 jedinců plemene Aberdeen Angus, 363 jedinců plemene České strakaté, 243 jedinců plemene Holstein, 188 jedinců plemene Limousin, 137 jedinců plemene Hereford, 125 jedinců plemene Piemontese, 73 jedinců plemene Blonde d'Aquitaine a 66 jedinců plemene Galloway. Dále bylo z celkové databáze 7776 jedinců náhodně vybráno 380 nepříbuzných kříženců. Tato podmnožina slouží k popisu populace kříženců skotu v podmínkách ČR a jako srovnávací báze pro diskusi některých výsledků.

Jako nejvíce uniformní plemeno z hlediska genetické diverzity bylo na základě výsledků v práci vyhodnoceno plemeno Hereford s následujícími průměrnými výsledky počítanými přes všechny lokusy – nejvyšší alelická frekvence 0,500, pozorovaná a očekávaná heterozygotnost 0,646, resp. 0,637, koeficient inbreedingu -0,011 a PIC=0,589. Naopak, plemeno Piemontese bylo v práci vyhodnoceno jako nejvíce divergentní z hlediska zmíněných parametrů.

Pokud se zaměříme na genetickou diverzitu pozorovanou v jednotlivých zkoumaných lokusech, můžeme jako nejvíce divergentní označit lokus *TGLA227*. Zajímavých výsledků bylo dosaženo z hlediska nízké (redukované) genetické variability v lokusech *ETH3* (Blonde d'Aquitaine), *ETH10* (Charolais, Galloway, Limousin) a *INRA023* (Blonde d'Aquitaine, Hereford) především u masných plemen skotu s intenzivním šlechtěním. Vzhledem k tomu, že několik autorů zmiňuje výše uvedené lokusy jako genetické markery spojené s masnou užitkovostí, pozorované výsledky reflektují šlechtitelský tlak v těchto populacích. V porovnání s výsledky uvedenými v pracích z 90. let minulého století, které uváděly, že intenzivně šlechtěná plemena vykazují alarmující výsledky zredukované genetické variability (např. Holstein), musíme na základě našich výsledků konstatovat, že se v současnosti díky pečlivě plánovaným šlechtitelským strategiím dokázala z tohoto stavu vrátit k akceptovatelné úrovni genetické variability.

Ke zjištění specifických vztahů a stavu sledovaných subpopulací skotu bylo v práci určeno 8 genetických distancí založených na různých teoretických modelech (geometrické, SMM, AIM). K vizualizaci zmíněných distancí bylo využito UPGMA a NJ algoritmů k budování genetických stromů. Plemena Hereford na jedné straně a Aberdeen Angus na straně druhé byla identifikována jako nejvíce rozdílná v souladu se všemi metodami kalkulace genetických distancí použitých v práci. Všechny metody také shodně určily do jed-

né podmnožiny (větve) stromů jedince plemene České strakaté a množinu kříženců. Tyto výsledky ukazují ve shodě s reálným stavem populace skotu široké využití plemene České strakaté pro produkci kříženců pro masnou i mléčnou užitkovost. Plemeno Masný simentál bylo v souladu s výsledky a předpoklady všech kalkulovaných metod vyhodnoceno z hlediska genetických vzdáleností a jejich zobrazení jako velice blízké plemenu České strakaté. Také tato skutečnost reflektuje reálný stav a historický vývoj obou plemen.

Obecně můžeme na základě dosažených výsledků konstatovat, že plemena chovaná v širokém měřítku v České republice i celosvětově (České strakaté, Holstein, Limousin, Blonde d'Aquitaine atd.) a plemena geneticky i vývojově odlišná (Piemontese) vykazují větší genetickou variabilitu v našich podmínkách ve srovnání s okrajově chovanými (přestože geneticky odlišnými) (Galloway) a více uniformními populacemi v rámci ČR jako Hereford, Masný simentál, Charolais a Aberdeen Angus. Tyto výsledky odrážejí jak chovatelské strategie (včetně importů genetického materiálu do ČR), tak historický vývoj plemen. Na základě těchto výsledků je možno konstatovat, že mikrosatelitní markery stále představují efektivní nástroj k popisu, sledování a kontrole genetické variability skotu pomocí použitých metod.

V práci byly pomocí kombinovaných pravděpodobností a hodnoty polymorfního informačního obsahu (PIC) vyhodnoceny možnosti využití sledované sady mikrosatelitů k úlohám vyloučení paternity u jednotlivých plemen i pro celou datovou sadu. Pro jednotlivá plemena bylo nejnižších hodnot u všech tří kombinovaných pravděpodobností dosaženo u plemene Hereford, nejvyšších pak u plemene Piemontese. Tyto výsledky odpovídají zjištěním týkajících se parametrů genetické variability u obou plemen. Ze všech dosažených výsledků uvádíme $CEP1=0,953371$ u plemene Hereford, které je možno označit za alarmující (pouze u ~ 95 % jedinců plemene je možno jednoznačně vyloučit rodiče při znalosti všech 3 genotypů). Obdobné výsledky pak byly vykalkulovány pro scénář $CEP2=0,948479$. Ostatní hodnoty kombinovaných pravděpodobností pro jednotlivé způsoby vyloučení paternity se blíží hodnotě 1,000 pro všechna sledovaná plemena. Je tedy možné konstatovat, že sada použitých mikrosatelitních markerů je použitelná pro úlohu vyloučení rodičovství u sledovaných plemen.

V práci dále uvádíme popis návrhu, algoritmů a implementace softwarové aplikace pro práci s rozsáhlými datovými sadami genotypových dat, zejména pro účely využití v rutinní genetické laboratoři. V souladu s pravidly designu software uvádíme nejdříve modely aplikace popisující její typické úlohy ve smyslu použití, uživatelů, síťových technologií, bezpečnosti a uživatelského prostředí. K popisu a návrhu software byl použit modelovací jazyk UML jako podklad pro následnou implementaci v objektovém prostředí Borland Delphi 2005.

Kvůli nutnosti práce s poměrně rozsáhlými datovými množinami byla jako úložiště zvolena SQL databáze a k jejímu návrhu využito ERD diagramů. Dále jsou v práci uvedeny SQL příkazy a fronty typické pro úlohy spojené (vyhledávání, třídění, filtrování) s genotypovými daty mikrosatelitních markerů. Specifická povaha genotypových dat vyžaduje některé modifikace SQL přístupu k jednoatributovým datům. Proto je v práci uvedeno, jak se s těmito omezeními vyrovnat – jsou uváděny SQL fronty pro kalkulaci alelických a genotypových frekvencí atd.

Dále uvádíme základní algoritmy pro tvorbu a práci s výsledky SQL front v úlohách s genotypovými daty – kalkulace genotypových a alelických frekvencí, kalkulace kombinovaných pravděpodobností vyloučení paternity, testování paternity a vyhledání potenciálních rodičů, manipulace s datovými sety na základě genotypových dat atd.

V práci je dále popsáno uživatelské rozhraní aplikace, která vznikla na základě popsaných

metod, a popsána její funkcionalita. Byly vytvořeny protokoly a subsystém pro jejich generování. Ve výsledné aplikaci je také implementován algoritmus G-metric, jehož návrh je rovněž součástí této práce. Výsledná softwarová aplikace je využívána akreditovanou laboratoří od jejích prvotních implementací v roce 2009.

Pro účely zjištění možností plemenné diskriminace na základě mikrosatelitních markerů skotu bylo použito 10 algoritmů metod strojového učení (J48, JRip, Naive Bayes, Bayes Net, IB1, IB5, SMO, Vote classifier, nově navržený a implementovaný algoritmus G-metric se dvěma metodami pravděpodobnostního rozhodování - Slope a IB1). Parametry klasifikace byly zkoumány na třech různých sadách genotypových dat. Datovou sadu tvořilo 3300 jedinců plemen skotu popsaných výše.

Pro každý druh datového setu a každý algoritmus jsou v práci popsány a diskutovány výsledky klasifikačních možností algoritmů z hlediska následujících parametrů: procento správně klasifikovaných instancí, Kappa statistika, procento falešně pozitivní klasifikace, přesnost a F-míra. Dále jsou prezentovány tvz. matice záměn a grafy pravděpodobnostních předpovědí pro celé datové sady. Především tyto grafy poskytují ve srovnání s klasickými dendrogramy efektivní nástroj pro vizuální prezentaci diverzity a příbuznosti jednotlivých plemen. V této souvislosti je třeba poznamenat, že algoritmus G-metric je speciálně vyvinut pro účely klasifikace genotypových dat a výsledky, které pomocí něj určíme mohou být využity právě v úlohách genetické diverzity, jak je v práci také prezentováno.

Další část práce je věnována výsledkům a diskuzi použitelnosti jednotlivých algoritmů a datových reprezentací k úloze plemenné diskriminace. Tato část práce poskytuje teoretický základ pro výběr vhodných algoritmů a datové reprezentace genotypových dat. Mohla by být využita i v praktických úlohách při tvorbě komerčních služeb laboratoří.

Z hlediska robustnosti a nezávislosti na datové reprezentaci jsme vyhodnotili pro danou úlohu jako nejvhodnější algoritmy založené na pravděpodobnostních principech (např. Bayesovské sítě, Vote classifier). Tyto algoritmy mohou i při využití datových sad s vyšším množstvím informace (datový set s genotypovou informací uspořádanou podle alelických frekvencí, případně délek) dosahovat velmi dobrých výsledků i v případech, že pro tvorbu modelů je použita trénovací množina s relativně nízkým počtem případů. Naproti tomu metody založené na tzv. "učení instancí" (IB1, IB5) nebo na dělení datového prostoru (SMO, J48, JRip) jsou vysoce závislé na množství informace, podle které následně klasifikují nové případy (množství atributů) a množství dat v trénovací množině.

Z hlediska využitelnosti pro úlohu plemenné diskriminace u skotu na základě genotypových dat mikrosatelitních markerů se jeví jako nejlepší algoritmus Bayesovských sítí (Bayes Net). V práci pro něj bylo dosaženo průměrných výsledků (přes všechny datové sady): 84,274 % správně klasifikovaných případů, Kappa=0,811, FP Rate=0,030, Přesnost=0,842, F-míra=0,841. Tyto výsledky jsou plně srovnatelné s předchozími pracemi, ve kterých byla podmnožina algoritmů aplikována na mikrosatelitní genotypová data u koní pro vyhodnocení stejné úlohy. Srovnání výsledků ukazuje, že robustnost algoritmu není příliš ovlivněna počtem jedinců použitých pro tvorbu modelů (3300 u skotu, 932 u koní) ani počtem použitých mikrosatelitních markerů (10 u skotu, 17 u koní). Je vhodné zmínit, že algoritmus se jeví jako robustní i v případě odlišných živočišných druhů a naprosto jiných podmínkách jejich chovu a šlechtění.

Na základě uvedených výsledků je možno konstatovat, že za pomoci metod popsaných v práci je možno efektivně predikovat plemena skotu na základě genotypových dat mikrosatelitních markerů.

8 Practical Usage of Results

First of all, practical results are coming from the part aimed to evaluation of genetic diversity. As used dataset of individuals was created within commercial accredited laboratory in three years, results reflect state of genetic diversity through Czech cattle subpopulations. For particular breeds, results can be compared with previous results and should be used by breeders associations in the Czech Republic as indicator of genetical changes in their breeds as well as control tool for evaluation of breeding strategies applied in Czech. It should be recommended to perform this evaluation periodically with regard to cattle populations evolution. So, 3 years periods seem to be suitable for this issue as well as results coming from commercial laboratories. Used methods and their results in this part of thesis showed as well that they can be used for these purposes and reflect real state of genetic diversity in cattle.

Results about microsatellite panel usability evaluation show if panel of loci can be used for commercial testing of paternity in cattle within and across tested breeds. Also, parameters and probabilities are shown for evaluated datasets. This offers to use probabilities as warning system as well when ever they are reduced significantly in particular scenario of paternity exclusion. Laboratories should use these type of results as was described, published them periodically, include them to yearly accreditation parameters and report them to authorities like ISAG/FAO is.

Software application created and described in thesis is practical result itself. Software application created based on proposed algorithms, models and methods is used by commercial laboratory nowadays. It offers to control periodically parameters of microsatellite set usability for genetic typing, paternity control and monitoring of genetic diversity. As data warehouse is built on SQL principles, all of evaluations could be done with usage of whole SQL power of selecting, sorting etc. Anyway, software application reduces human error factor in data processing as it offers automatic connection to sequencing instruments as well as semi-automated input of identification data and results. Software automates a lot of daily issues in normal laboratory operation as protocols creation, automated testing of paternity etc. It also deals with normal security and traceability conditions of operation done within network environment. As it is built as open system and all of development parts and algorithms are fully described, it represents good framework which can be extended by many functions requested in future as was thought during design.

The most recent practical outcome of thesis is represented by evaluated and proved possibilities of cattle breeds discrimination by using machine learning methods in microsatellite genotype datasets. When possibilities are proved, proper algorithms and datasets are selected, results of classification power are shown, then there is open space to extend results reached in commercial type service. This service can be used to identify unknown samples on breed level and could be used e.g. in farming, forensic issues, cattle diversity studies, traceability and security of beef cattle. Regarding to results, we can recommend to build classification models incrementally on yearly basis and to report classification power results. Theoretical base for these purposes is proposed in the thesis.

9 Theoretical Outcomes and Future work

Presented results of cattle subpopulations genetic diversity in the Czech Republic are fully comparable to studies done previously all over the world as common methods were used for this purpose. So, when results obtained are easily discussed with real state of breeding strategies applied in past, theoretical outcome which proves their usability for these purposes is result of the thesis. Also, when 3300 individuals were inspected in work, we can say that results of genetic diversity reached and presented give state-of-art of Czech cattle subpopulations genetic aspects nowadays.

As results of evaluation of paternity testing scenarios plus results of genetic diversity in Hereford subpopulation show alarming reduction of genetic variability, we can recommend more detailed inspection in Czech Hereford population.

Regarding design of software application and its implementation, a lot of principles of theoretical computation science were applied. As large datasets are operated and thanks to different basis of genotype data, a lot of algorithms could be applied or developed especially in proposed tasks. Proposed design of software application gives a framework which can be extended by future work in the following topics:

- searching database to find possible pairs of parents,
- creating interface independent on laboratory technology used,
- handling with genotype data in more effective ways,
- incorporating of clustering algorithms to find family relationships within databases,
- implementing of machine learning classifiers for breed discrimination,
- extension of database providing yield traits data and algorithms for breeding strategies selection based on them,
- incorporating with national authorities systems for central evidence,
- functional genotypes storage and computations,
- direct web-based communication with breeders,
- fully automated solutions for data and samples flow,
- etc.

Main theoretical outcome obtained by thesis in cattle breed discrimination task is that cattle breeds can be classified on breed level properly by machine learning methods based on microsatellite data. Then combination of principles and datasets with different data representations were examined to show how task can be resolved and what expected results should be from many point of views (usage for genetic diversity exploration, choosing of datasets representation, robustness of methods with different datasets, classification power of algorithms themselves, specific results reached within each breed etc.). This offers a large opportunities for future work. Also, these results offer manual for future work as well as for creating commercial applications based on used theory and obtained results.

Implementation of G-metric classifier and its theoretical base also opens a lot of future issues:

- usage of G-metric based classifiers in different species,
- inspection and extension of theoretical base in which G-metric classifier is de-

9 Theoretical Outcomes and Future work

signed, as it is designed for usage with genotype data in comparison with common classification algorithms using one attribute principle,

- design of metric functions suitable for usage with lazy based classifiers and their theory in accordance with population like effects,
- new implementations and design of voting functions in lazy based classifiers (and extension of regression principle designed and proposed in this work),
- design of new algorithms suitable for breed classification task.

Concept of breed discrimination task itself opens interesting questions connected with theoretical base of machine learning and genetics together. Lets mention couple of the most interesting:

- to extend results coming from this work and to inspect classification abilities of algorithms on more levels of classification like subpopulations, large families etc.,
- to inspect if and how precisely portions of breed for individual can be predicted,

General concept proposed in this work - usage of machine learning classification methods for classification of genetic data can be extended in future mainly by:

- development and application of new algorithms for classification of genotype data,
- application of normally used classification methods on genotype data for another classification purposes (like evaluation or prediction of haplotypes, prediction based on genotype multi loci data, etc.),
- creation of commercial based services using these methods and their implementation in routinely used SW.

Finally, we need to mention that with usage of routine genotyping in commercial laboratory, material for large study can be extended easily in comparison with research projects and grants. In this case, general problem of funding in Czech research and no existence of relationships between R&D and commercial sphere must be mentioned. Results show, that, thanks to cooperation, useful results with positive outcome for both of spheres can be concluded.

10 References

1. Achilli, A., Olivieri, A., Pellecchia, M., Uboldi, C., Colli, L., Al-Zahery, N., Accetturo, M., Pala, M., Kashani, B. H., Perego, U. A., Battaglia, V., Fornarino, S., Kalamati, J., Houshmand, M., Negrini, R., Semino, O., Richards, M., Macaulay, V., Ferretti, L., Bandelt, H. J., Jmone-Marsan, P. & Torroni, A. (2008). Mitochondrial genomes of extinct aurochs survive in domestic cattle. *Current Biology*, Volume 18, Issue 4, pp R157-R158.
2. Aha, D. W., Kibler, D. & Albert, M. K. (1991). Instance-based learning algorithms. *Machine learning*, Volume 6, Issue 1, pp 37-66.
3. American Angus Association, available on: <http://www.angus.org/>, on date: 1.3.2012.
4. American International Charolais Association, available on: <http://www.charolaisusa.com/history.html>, on date: 1.3.2012.
5. Andrieu, C., De Freitas, N., Doucet, A. & Jordan, M. I. (2003). An introduction to MCMC for machine learning. *Machine learning*, Volume 50, Issue 1, pp 5-43.
6. Arranz, J. J., Bayon, Y. & SanPrimitivo, F. (1996). Genetic variation at five microsatellite loci in four breeds of cattle. *Journal of Agricultural Science*, Volume 127, pp 533-538.
7. Baker, C. M. A. & Manwell, C. (1980). Chemical Classification of Cattle Breed Groups. *Animal Blood Groups and Biochemical Genetics*, Volume 11, Issue 3, pp 127-150.
8. Barker, J. S. F. (1999). Conservation of livestock breed diversity. *Animal Genetic Resources Information*, Volume 25, pp 33-44.
9. Beaumont, M. A. & Rannala, B. (2004). The Bayesian revolution in genetics. *Nature Reviews Genetics*, Volume 5, Issue 4, pp 251-261.
10. Beja-Pereira, A., Alexandrino, P., Bessa, I., Carretero, Y., Dunner, S., Ferrand, N., Jordana, J., Laloe, D., Moazami-Goudarzi, K., Sanchez, A. & Canon, J. (2003). Genetic characterization of southwestern European bovine breeds: A historical and biogeographical reassessment with a set of 16 microsatellites. *J Hered*, Volume 94, Issue 3, pp 243-250.
11. Berka, P. (2001). *Dobývání znalostí z databází*. Academia, Prague.
12. Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer New York.
13. Bjornstad, G. & Roed, K. H. (2002). Evaluation of factors affecting individual assignment precision using microsatellite data from horse breeds and simulated breed crosses. *Animal Genetics*, Volume 33, Issue 4, pp 264-270.
14. Blott, S. C., Williams, J. L. & Haley, C. S. (1998a). Genetic variation within the Hereford breed of cattle. *Animal Genetics*, Volume 29, Issue 3, pp 202-211.
15. Blott, S. C., Williams, J. L. & Haley, C. S. (1998b). Genetic relationships among European cattle breeds. *Animal Genetics*, Volume 29, Issue 4, pp 273-282.
16. Bonadonna, T. (1959). *Le razze bovine, bufali, cattali, zeb*, *Telqsforo Bonadonna*. Progresso zootecnico. Milano. IT.
17. Borland Inc. (2007). Borland Delphi 2005. Available on: www.borland.com
18. Botstein, D., White, R. L., Skolnick, M. & Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics*, Volume 32, Issue 3, pp 314.
19. Bradley, D. G., Loftus, R. T., Cunningham, P. & Machugh, D. E. (1998). Genetics

- and domestic cattle origins. *Evolutionary Anthropology*, Volume 6, Issue 3, pp 79-86.
20. Bradley, D. G., Machugh, D. E., Cunningham, P. & Loftus, R. T. (1996). Mitochondrial diversity and the origins of African and European cattle. *Proceedings of the National Academy of Sciences of the United States of America*, Volume 93, Issue 10, pp 5131-5135.
 21. Buchanan, F. C., Adams, L. J., Littlejohn, R. P., Maddox, J. F. & Crawford, A. M. (1994). Determination of Evolutionary Relationships Among Sheep Breeds Using Microsatellites. *Genomics*, Volume 22, Issue 2, pp 397-403.
 22. Burócziová, M. & Říha, J. (2009). Horse breed discrimination using machine learning methods. *JOURNAL OF APPLIED GENETICS*, Volume 50 (4), Issue 4, pp 375-377.
 23. Canadian Simmental Association, available on: <http://www.simmental.com/>, on date: 1.3.2012.
 24. Canon, J., Alexandrino, P., Bessa, I., Carleos, C., Carretero, Y., Dunner, S., Ferran, N., Garcia, D., Jordana, J., Laloe, D., Pereira, A., Sanchez, A. & Moazami-Goudarzi, K. (2001). Genetic diversity measures of local European beef cattle breeds for conservation purposes. *Genetics Selection Evolution*, Volume 33, Issue 3, pp 311-332.
 25. Canon, J., Checa, M. L., Carleos, C., Vega-Pla, J. L., Vallejo, M. & Dunner, S. (2000). The genetic structure of Spanish Celtic horse breeds inferred from microsatellite data. *Animal Genetics*, Volume 31, Issue 1, pp 39-48.
 26. Cavalli-Sforza, L. L. & Edwards, A. W. F. (1967). Phylogenetic analysis. Models and estimation procedures. *American Journal of Human Genetics*, Volume 19, Issue 3 Pt 1, pp 233.
 27. Chakraborty, R. & Jin, L. (1993). Determination of relatedness between individuals using DNA fingerprinting. *Hum. Biol.*, Volume 65, Issue 6, pp 875-895.
 28. Choi, J. H., Jung, H. Y., Kim, H. S. & Cho, H. G. (2000). PhyloDraw: a phylogenetic tree drawing system. *Bioinformatics*, Volume 16, Issue 11, pp 1056-1058.
 29. Choroszy, B., Janik, A., Choroszy, Z. & Zabek, T. (2006). Polymorphism of selected microsatellite DNA sequences in Simmental cattle chosen for identification of QTLs for meat traits. *Animal Science Papers and Reports*, Volume 24, Issue 2, pp 71-77.
 30. Ciampolini, R., Mazzanti, E. & Cianci, D. (2002). DNA microsatellites associated with morphological traits in beef cattle. *Annali della Facoltà di Medicina veterinaria*, Volume 55, pp 205-221.
 31. Ciampolini, R., Moazamigoudarzi, K., Vaiman, D., Dillmann, C., Mazzanti, E., Foulley, J. L., Leveziel, H. & Cianci, D. (1995). Individual Multilocus Genotypes Using Microsatellite Polymorphisms to Permit the Analysis of the Genetic-Variability Within and Between Italian Beef-Cattle Breeds. *Journal of Animal Science*, Volume 73, Issue 11, pp 3259-3268.
 32. Čítek, J. & Řehout, V. (2001). Evaluation of the genetic diversity in cattle using microsatellites and protein markers. *Czech Journal of Animal Science*, Volume 46, Issue 9, pp 393-400.
 33. Clave, P. (2003). Past and future activities to harmonize farm animal biodiversity studies on a global scale. *Archivos de zootecnia*, Volume 52, Issue 198, pp 193-199.
 34. Cohen, W. W. (1995). Fast effective rule induction. *Proceedings of the Twelfth International Conference on Machine Learning*, pp 115-123.

35. Cornuet, J. M., Piry, S., Luikart, G., Estoup, A. & Solignac, M. (1999). New methods employing multilocus genotypes to select or exclude populations as origins of individuals. *Genetics*, Volume 153, Issue 4, pp 1989-2000.
36. Czech Beef Breeders Association, available on: <http://www.cschms.cz/english/index.php>, on date: 1.3.2012.
37. Czech Fleckvieh Breeders Association, available on: <http://www.cestr.info/>, on date: 1.3.2012.
38. Czernekova, V., Kott, T., Dudkova, G., Sztankoova, Z. & Soldat, J. (2006). Genetic diversity between seven Central European cattle breeds as revealed by microsatellite analysis. *Czech Journal of Animal Science*, Volume 51, Issue 1, pp 1-7.
39. D'Andrea, M., Pariset, L., Matassino, D., Valentini, A., Lenstra, J. A., Maiorano, G. & Pilla, F. (2011). Genetic characterization and structure of the Italian Podolian cattle breed and its relationship with some major European breeds. *Italian Journal of Animal Science*, Volume 10, Issue 4, pp 54.
40. Dawson, K. J. & Belkhir, K. (2001). A Bayesian approach to the identification of panmictic populations and the assignment of individuals. *Genetical Research*, Volume 78, Issue 01, pp 59-77.
41. De la Vega, F. M., Dailey, D., Ziegler, J., Williams, J., Madden, D. & Gilbert, D. A. (2002). New generation pharmacogenomic tools: A SNP linkage disequilibrium map, validated SNP assay resource, and high-throughput instrumentation system for large-scale genetic studies. *Biotechniques*, pp 48-+.
42. DeAtley, K. L., Rincon, G., Farber, C. R., Medrano, J. F., Enns, R. M., Silver, G. A. & Thomas, M. G. (2008). Association of microsatellite ETH10 genotypes with growth and carcass trait levels in Brangus cattle. *Proceedings, Western Section, American Society of Animal Science. Journal of Animal Science*, Volume 86.
43. DeAtley, K. L., Rincon, G., Farber, C. R., Medrano, J. F., Luna-Nevarez, P., Enns, R. M., VanLeeuwen, D. M., Silver, G. A. & Thomas, M. G. (2011). Genetic analyses involving microsatellite ETH10 genotypes on bovine chromosome 5 and performance trait measures in Angus-and Brahman-influenced cattle. *Journal of Animal Science*, Volume 89, Issue 7, pp 2031-2041.
44. Del Bol, L., Polli, M., Longeri, M., Ceriotti, G., Looft, C., Barre-Dirie, A., Dolf, G. & Zanotti, M. (2001). Genetic diversity among some cattle breeds in the Alpine area. *Journal of Animal Breeding and Genetics*, Volume 118, Issue 5, pp 317-325.
45. Estoup, A., Garnery, L., Solignac, M. & Cornuet, J. M. (1995). Microsatellite variation in honey bee (*Apis mellifera* L.) populations: hierarchical genetic structure and test of the infinite allele and stepwise mutation models. *Genetics*, Volume 140, Issue 2, pp 679.
46. Eugenio, B. D. & Glass, M. (2004). The kappa statistic: A second look. *Computational linguistics*, Volume 30, Issue 1, pp 95-101.
47. Fan, B., Wang, Z. G., Li, Y. J., Zhao, X. L., Liu, B., Zhao, S. H., Yu, M., Li, M. H., Chen, S. L., Xiong, T. A. & Li, K. (2002). Genetic variation analysis within and among Chinese indigenous swine populations using microsatellite markers. *Animal Genetics*, Volume 33, Issue 6, pp 422-427.
48. FAO - Measurements of Domestic Animal Diversity, available on: <http://dad.fao.org>, on date: 1.3.2012.
49. Farnir, F., Coppieters, W., Arranz, J. J., Berzi, P., Cambisano, N., Grisart, B., Karim, L., Marcq, F., Moreau, L., Mni, M., Nezer, C., Simon, P., Vanmanshoven, P., Wagenaar, D. & Georges, M. (2000). Extensive genome-wide linkage disequilibrium

- rium in cattle. *Genome Research*, Volume 10, Issue 2, pp 220-227.
50. Fast Reports Inc. (2007). FreeReports 2.33. Available on: www.fast-reports.com; <http://freereport.sf.net>.
 51. Felius, M. (1995). *Cattle breeds: an encyclopedia*. C Misset bv.
 52. Felius, M., Koolmees, P. A., Theunissen, B. & Lenstra, J. A. (2011). On the Breeds of CattleGÇöHistoric and Current Classifications. *Diversity*, Volume 3, Issue 4, pp 660-692.
 53. Firebird Database Project. (2008). Firebird 2.0. Available on: www.firebirdsql.com
 54. Freeman, A. R., Bradley, D. G., Nagda, S., Gibson, J. P. & Hanotte, O. (2006). Combination of multiple microsatellite data sets to investigate genetic diversity and admixture of domestic cattle. *Animal Genetics*, Volume 37, Issue 1, pp 1-9.
 55. Freeman, A. R., Meghen, C. M., Machugh, D. E., Loftus, R. T., Achukwi, M. D., Bado, A., Sauveroche, B. & Bradley, D. G. (2004). Admixture and diversity in West African cattle populations. *Molecular Ecology*, Volume 13, Issue 11, pp 3477-3487.
 56. Freund, Y. & Schapire, R. E. (1996). Experiments with a new boosting algorithm. *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE*, pp 148-156.
 57. Garza, J. C., Slatkin, M. & Freimer, N. B. (1995). Microsatellite allele frequencies in humans and chimpanzees, with implications for constraints on allele size. *Molecular Biology and Evolution*, Volume 12, Issue 4, pp 594-603.
 58. Gautier, M., Faraut, T., Moazami-Goudarzi, K., Navratil, V., Fogfio, M., Grohs, C., Boland, A., Garnier, J. G., Boichard, D., Lathrop, G. M., Gut, I. G. & Eggen, A. (2007). Genetic and haplotypic structure in 14 European and African cattle breeds. *Genetics*, Volume 177, Issue 2, pp 1059-1070.
 59. Glowatzki-Mullis, M. L., Muntwyler, J., Pfister, W., Marti, E., Rieder, S., Poncet, P. A. & Gaillard, C. (2006). Genetic diversity among horse populations with a special focus on the Franches-Montagnes breed. *Animal Genetics*, Volume 37, Issue 1, pp 33-39.
 60. Goldstein, D. B., Linares, A. R., Cavalli-Sforza, L. L. & Feldman, M. W. (1995a). An Evaluation of Genetic Distances for Use With Microsatellite Loci. *Genetics*, Volume 139, Issue 1, pp 463-471.
 61. Goldstein, D. B., Ruiz Linares, A., Cavalli-Sforza, L. L. & Feldman, M. W. (1995b). Genetic absolute dating based on microsatellites and the origin of modern humans. *Proceedings of the National Academy of Sciences*, Volume 92, Issue 15, pp 6723.
 62. Guinand, B., Topchy, A., Page, K. S., Burnham-Curtis, M. K., Punch, W. F. & Scribner, K. T. (2002). Comparisons of Likelihood and Machine Learning Methods of Individual Classification. *J Hered*, Volume 93, Issue 4.
 63. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). Waikato Environment for Knowledge Analysis - Weka 3.6.6. Available on: <http://www.cs.waikato.ac.nz/~ml/weka/>
 64. Hall, S. J. G. (2004). *Livestock biodiversity: genetic resources for the farming of the future*. Blackwell Pub.
 65. Han, M. V. & Zmasek, C. M. (2009). phyloXML: XML for evolutionary biology and comparative genomics. *BMC bioinformatics*, Volume 10, Issue 1, pp 356.
 66. Hancock, J. M., Goldstein, D. B. & Schlotterer, C. (1999). *Microsatellites: Evolution and Applications*. Oxford University Press.

67. Hanotte, O. & Jianlin, H. (2005). 8. Genetic characterization of livestock populations and its use in conservation decision-making. *The role of biotechnology in exploring and protecting agricultural genetic resources*, pp 89.
68. Hanotte, O., Bradley, D. G., Ochieng, J. W., Verjee, Y., Hill, E. W. & Rege, J. E. O. (2002). African pastoralism: Genetic imprints of origins and migrations. *Science*, Volume **296**, Issue 5566, pp 336-339.
69. Hanslik, S., Harr, B., Brem, G. & Schlotterer, C. (2000). Microsatellite analysis reveals substantial genetic differentiation between contemporary new world and old world Holstein Friesian populations. *Animal Genetics*, Volume **31**, Issue 1, pp 31-38.
70. Hayes, B. J., Bowman, P. J., Chamberlain, A. J. & Goddard, M. E. (2009). Invited review: Genomic selection in dairy cattle: Progress and challenges. *Journal of Dairy Science*, Volume **92**, Issue 2, pp 433-443.
71. Hedrick, P. W. (2011). *Genetics of populations*. Jones & Bartlett Learning.
72. Helmer, D., Gourichon, L., Monchot, H., Peters, J. & Sana Segui, M. (2005). Identifying early domestic cattle from Pre-Pottery Neolithic sites on the Middle Euphrates using sexual dimorphism. *The first steps of animal domestication: New archaeological approaches*. Great Britain: Oxbow Books, pp 86-95.
73. Hill, W. G. (1981). Estimation of Effective Population-Size from Data on Linkage Disequilibrium. *Genetical Research*, Volume **38**, Issue 3, pp 209-216.
74. Holstein Association USA, available on: <http://www.holsteinusa.com/>, on date: 1.3.2012.
75. International Society for Animal Genetics, available on: <http://www.isag.us>, on date: 1.3.2012.
76. Jamieson, A. & Taylor, S. C. S. (1997). Comparisons of three probability formulae for parentage exclusion. *Animal Genetics*, Volume **28**, Issue 6, pp 397-400.
77. Kantanen, J., Olsaker, I., Holm, L. E., Lien, S., Vilkki, J., Brusgaard, K., Eythorsdottir, E., Danell, B. & Adalsteinsson, S. (2000). Genetic diversity and population structure of 20 North European cattle breeds. *J Hered*, Volume **91**, Issue 6, pp 446-457.
78. Kitada, S., Hayashi, T. & Kishino, H. (2000). Empirical Bayes Procedure for Estimating Genetic Distance Between Populations and Effective Population Size. *Genetics*, Volume **156**, Issue 4, pp 2063-2079.
79. Kohavi, R. (1996). Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pp 202-207.
80. Koskinen, M. T. (2003). Individual assignment using microsatellite DNA reveals unambiguous breed identification in the domestic dog. *Animal Genetics*, Volume **34**, Issue 4, pp 297-301.
81. Kruger, K., Gaillard, C., Stranzinger, G. & Rieder, S. (2005). Phylogenetic analysis and species allocation of individual equids using microsatellite data. *Journal of Animal Breeding and Genetics*, Volume **122**, Issue Suppl. 1, pp 78-86.
82. Kuhn, C., Leveziel, H., Renand, G., Goldammer, T., Schwerin, M. & Williams, J. (2005). Genetic markers for beef quality. *Indicators of milk and beef quality, EAAP Publ*, Volume **112**, pp 23-32.
83. Kullback, S. (1987). The kullback-leibler distance. *The American Statistician*, Volume **41**, pp 340-341.
84. Kumar, P., Freeman, A. R., Loftus, R. T., Gaillard, C., Fuller, D. Q. & Bradley, D.

- G. (2003). Admixture analysis of South Asian cattle. *Heredity*, Volume **91**, Issue 1, pp 43-50.
85. Lenstra, J. A. (2006). Marker-assisted conservation of European cattle breeds: an evaluation. *Animal Genetics*, Volume **37**, Issue 5, pp 475-481.
86. Li, M. H., Tapio, I., Vilkki, J., Ivanova, Z., Kiselyova, T., Marzanov, N., Cinkulov, M., Stojanovic, S., Ammosov, I., Popov, R. & Kantanen, J. (2007). The genetic structure of cattle populations (*Bos taurus*) in northern Eurasia and the neighbouring Near Eastern regions: implications for breeding strategies and conservation. *Molecular Ecology*, Volume **16**, Issue 18, pp 3839-3853.
87. Liu, J. (2006). PowerMarker 3.25. Available on: <http://statgen.ncsu.edu/power-marker/>
88. Liu, K. & Muse, S. V. (2005). PowerMarker: an integrated analysis environment for genetic marker analysis. *Bioinformatics*, Volume **21**, Issue 9, pp 2128.
89. Loftus, R. T., Machugh, D. E., Bradley, D. G., Sharp, P. M. & Cunningham, P. (1994). Evidence for two independent domestications of cattle. *Proc. Natl. Acad. Sci. USA*, Volume **91**, pp 2757-2761.
90. Machado, M. A., Schuster, I., Martinez, M. L. & Carnpos, A. L. (2003). Genetic diversity of four cattle breeds using microsatellite markers. *Revista Brasileira de Zootecnia-Brazilian Journal of Animal Science*, Volume **32**, Issue 1, pp 93-98.
91. Machugh, D. E., Loftus, R. T., Cunningham, P. & Bradley, D. G. (1998). Genetic structure of seven European cattle breeds assessed using 20 microsatellite markers. *Animal Genetics*, Volume **29**, Issue 5, pp 333-340.
92. Machugh, D. E., Shriver, M. D., Loftus, R. T., Cunningham, P. & Bradley, D. G. (1997). Microsatellite DNA variation and the evolution, domestication and phylogeography of taurine and zebu cattle (*Bos taurus* and *Bos indicus*). *Genetics*, Volume **146**, Issue 3, pp 1071.
93. Manel, S., Berthier, P. & Luikart, G. (2002). Detecting Wildlife Poaching: Identifying the Origin of Individuals with Bayesian Assignment Tests and Multilocus Genotypes. *Conservation Biology*, Volume **16**, Issue 3, pp 650-659.
94. Mannen, H., Kohno, M., Nagata, Y., Tsuji, S., Bradley, D. G., Yeo, J. S., Nyamsamba, D., Zagdsuren, Y., Yokohama, M., Nomura, K. & Amano, T. (2004). Independent mitochondrial origin and historical genetic differentiation in North Eastern Asian cattle. *Molecular Phylogenetics and Evolution*, Volume **32**, Issue 2, pp 539-544.
95. Mannen, H., Tsuji, S., Loftus, R. T. & Bradley, D. G. (1998). Mitochondrial DNA variation and evolution of Japanese black cattle (*Bos taurus*). *Genetics*, Volume **150**, Issue 3, pp 1169-1175.
96. Masuda, M. & Pella, J. (2004). Identification of source populations of mixture individuals from their genotypes. *North Pacific Anadromous Fish Commission Technical Report*, Volume **5**, pp 103-105.
97. Maudet, C., Luikart, G. & Taberlet, P. (2002). Genetic diversity and assignment tests among seven French cattle breeds based on microsatellite DNA analysis. *Journal of Animal Science*, Volume **80**, Issue 4, pp 942-950.
98. Mc Kay, S. D., Schnabel, R. D., Murdoch, B. M., Matukumalli, L. K., Aerts, J., Coppieters, W., Crews, D., Dias, E., Gill, C. A., Gao, C., Mannen, H., Wang, Z. Q., Van Tassell, C. P., Williams, J. L., Taylor, J. F. & Moore, S. S. (2008). An assessment of population structure in eight breeds of cattle using a whole genome SNP panel. *Bmc Genetics*, Volume **9**.

99. Meirelles, S. L., Gouveia, G. V., Gasparin, G., Alencar, M. M., Gouveia, J. J. S. & Regitano, L. C. A. (2011a). Candidate gene region for control of rib eye area in Canchim beef cattle. *Genetics and Molecular Research*, Volume 10, Issue 2, pp 1220-1226.
100. Michalakis, Y. & Excoffier, L. (1996). A Generic Estimation of Population Subdivision Using Distances Between Alleles With Special Reference for Microsatellite Loci. *Genetics*, Volume 142, Issue 3, pp 1061-1064.
101. Michie, D., Spiegelhalter, D. J., Taylor, C. C. & Campbell, J. (1994). Machine learning, neural and statistical classification.
102. Moazamigoudarzi, K., Vaiman, D., Mercier, D., Grohs, C., Furet, J. P., Leveziel, H. & Martin, P. (1994). Analysis of Genetic Diversity in French Cattle Breeds by the Use of Microsatellites - Preliminary-Results. *Genetics Selection Evolution*, Volume 26, pp S155-S165.
103. Moioli, B., Napolitano, F. & Catillo, G. (2004). Genetic diversity between Piedmontese, Maremmana, and Podolica cattle breeds. *J Hered*, Volume 95, Issue 3, pp 250-256.
104. Moore, S. S. & Hansen, C. (2003). Genomics: delivering added value to the beef industry? *Outlook on Agriculture*, Volume 32, Issue 4, pp 247-252.
105. National Association of Piemontese Cattle Breeders, available on: <http://www.anaborapi.it/>, on date: 1.3.2012.
106. Negrini, R., Nijman, I. J., Milanese, E., MoazamiGoudarzi, K., Williams, J. L., Erhardt, G., Dunner, S., Rodellar, C., Valentini, A. & Bradley, D. G. (2007). Differentiation of European cattle by AFLP fingerprinting. *Animal Genetics*, Volume 38, Issue 1, pp 60-66.
107. Nei, M. (1972). Genetic Distance Between Populations. *American Naturalist*, Volume 106, Issue 949, pp 283-&.
108. Nei, M. (1973). The theory and estimation of genetic distance. *Genetic structure of populations*. University of Hawaii Press, Honolulu, pp 45-54.
109. Nei, M., Tajima, F. & Tateno, Y. (1983). Accuracy of estimated phylogenetic trees from molecular data. *Journal of Molecular Evolution*, Volume 19, Issue 2, pp 153-170.
110. Nomura, T., Honda, T. & Mukai, F. (2001). Inbreeding and effective population size of Japanese Black cattle. *Journal of Animal Science*, Volume 79, Issue 2, pp 366-370.
111. Odani, M., Narita, A., Watanabe, T., Yokouchi, K., Sugimoto, Y., Fujita, T., Oguni, T., Matsumoto, M. & Sasaki, Y. (2006). Genome-wide linkage disequilibrium in two Japanese beef cattle breeds. *Animal Genetics*, Volume 37, Issue 2, pp 139-144.
112. Paetkau, D., Waits, L. P., Clarkson, P. L., Craighead, L., Vyse, E., Ward, R. & Strobeck, C. (1998). Variation in genetic diversity across the range of North American brown bears. *Conservation Biology*, Volume 12, Issue 2, pp 418-429.
113. Payne, W. J. A. & Hodges, J. (1997). Tropical cattle: origins, breeds and breeding policies.
114. Pearse, D. E. & Crandall, K. A. (2004). Beyond F ST: Analysis of population genetic data for conservation. *Conservation Genetics*, Volume 5, Issue 5, pp 585-602.
115. Phillips, W. (1961). World distribution of the major types of cattle. *J. Hered*, Volume 52, pp 207-213.
116. Prevosti, A., Ocana, J. & Alonso, G. (1975). Distances between populations of *Drosophila subobscura*, based on chromosome arrangement frequencies. *TAG*

- Theoretical and Applied Genetics*, Volume 45, Issue 6, pp 231-241.
117. Pritchard, J. K. & Przeworski, M. (2001). Linkage disequilibrium in humans: Models and data. *American Journal of Human Genetics*, Volume 69, Issue 1, pp 1-14.
 118. Pritchard, J. K., Stephens, M. & Donnelly, P. (2000). Inference of Population Structure Using Multilocus Genotype Data. *Genetics*, Volume 155, Issue 2, pp 945-959.
 119. Purdy, H. R., Dawes, R. J. & Hough, R. (2008). *Breeds of Cattle, 2nd Edition*. TRS Publishing Corp. Springfield, MI.
 120. Putnova, L., Vrtkova, I., Srubarova, P. & Stehlik, L. (2011). Utilization of a 17 Microsatellites Set For Bovine Traceability in Czech Cattle Populations. *Iranian Journal of Applied Animal Science*, Volume 1, Issue 1, pp 31-37.
 121. Qian, H. (2001). Relative entropy: free energy associated with equilibrium fluctuations and nonequilibrium deviations. *Phys. Rev. E. Stat. Nonlin. Soft. Matter Phys.*, Volume 63, Issue 4 Pt 1, pp 042103.
 122. Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, Volume 1, Issue 1, pp 81-106.
 123. Quinlan, J. R. (1992). Learning with continuous classes. *5th Australian Joint Conference on Artificial Intelligence*, pp 343-348.
 124. Quinlan, J. R. (1993). *C4. 5: programs for machine learning*. Morgan Kaufmann Publishers Inc. San Francisco, CA, USA.
 125. Radko, A. (2010). Application of a complementary set of 10 microsatellite DNA markers for parentage verification in Polish Red Cattle. *Annal. Anim. Sci*, Volume 10, Issue 1, pp 9-15.
 126. Rannala, B. & Mountain, J. L. (1997). Detecting immigration by using multilocus genotypes. *Proceedings of the National Academy of Sciences of the United States of America*, Volume 94, Issue 17, pp 9197-9201.
 127. Reynolds, J., Weir, B. S. & Cockerham, C. C. (1983). Estimation of the coancestry coefficient: basis for a short-term genetic distance. *Genetics*, Volume 105, Issue 3, pp 767.
 128. Roeder, K., Escobar, M., Kadane, J. B. & Balasz, I. (1998). Measuring heterogeneity in forensic databases using hierarchical Bayes models. *Biometrika*, Volume 85, Issue 2, pp 269.
 129. Rogers, J. S. (1972). Measures of genetic similarity and genetic distance. *Studies in Genetics VII. University of Texas Publication*, Volume 7213, pp 145-153.
 130. Russell, N. D., Rios, J., Erosa, G., Remmenga, M. D. & Hawkins, D. E. (2000). Genetic differentiation among geographically isolated populations of Criollo cattle and their divergence from other *Bos taurus* breeds. *Journal of Animal Science*, Volume 78, Issue 9, pp 2314-2322.
 131. Russell, W. C., Brinks, J. S. & Richardson, G. V. (1984). Changes in Genetic Variances with Increased Inbreeding of Beef-Cattle. *J Hered*, Volume 75, Issue 1, pp 8-10.
 132. Saitbekova, N., Gaillard, C., Obexer-Ruff, G. & Dolf, G. (1999). Genetic diversity in Swiss goat breeds based on microsatellite analysis. *Animal Genetics*, Volume 30, Issue 1, pp 36-41.
 133. Saitou, N. & Nei, M. (1987). The Neighbor-Joining Method - A New Method for Reconstructing Phylogenetic Trees. *Molecular Biology and Evolution*, Volume 4, Issue 4, pp 406-425.
 134. Schmid, M., Saitbekova, N., Gaillard, C. & Dolf, G. (1999). Genetic diversity in

- Swiss cattle breeds. *Journal of Animal Breeding and Genetics-Zeitschrift fur Tierzucht und Zuchtungsbiologie*, Volume **116**, Issue 1, pp 1-8.
135. Schork, N. J. (2002). Power calculations for genetic association studies using estimated probability distributions. *American Journal of Human Genetics*, Volume **70**, Issue 6, pp 1480-1489.
136. Shriver, M. D., Jin, L., Boerwinkle, E., Deka, R., Ferrell, R. E. & Chakraborty, R. (1995). A novel measure of genetic distance for highly polymorphic tandem repeat loci. *Molecular Biology and Evolution*, Volume **12**, Issue 5, pp 914-920.
137. Slatkin, M. (1995). A Measure of Population Subdivision Based on Microsatellite Allele Frequencies. *Genetics*, Volume **139**, Issue 1, pp 457-462.
138. Sorensen, A. C., Sorensen, M. K. & Berg, P. (2005). Inbreeding in Danish dairy cattle breeds. *Journal of Dairy Science*, Volume **88**, Issue 5, pp 1865-1872.
139. Stevanovic, J., Stanimirovic, Z., Dimitrijevic, V. & Maletic, M. (2010). Evaluation of 11 microsatellite loci for their use in paternity testing in Yugoslav Pied cattle (YU Simmental cattle). *Czech J. Anim. Sci.*, Volume **55**, Issue 6, pp 221-226.
140. Stonaker, H. H. (1951). A Unique Herd of Hereford Cattle - the Gietz,R.G. Herd Has An Average Inter-Se Relationship Greater Than That Between Full Sibs. *J Hered*, Volume **42**, Issue 4, pp 207-209.
141. Tautz, D. (1993). Notes on the definition and nomenclature of tandemly repetitive DNA sequences. *EXS*, Volume **67**, pp 21-28.
142. Tenesa, A., Knott, S. A., Ward, D., Smith, D., Williams, J. L. & Visscher, P. M. (2003). Estimation of linkage disequilibrium in a sample of the United Kingdom dairy cattle population using unphased genotypes. *Journal of Animal Science*, Volume **81**, Issue 3, pp 617-623.
143. Tenesa, A., Navarro, P., Hayes, B. J., Duffy, D. L., Clarke, G. M., Goddard, M. E. & Visscher, P. M. (2007). Recent human effective population size estimated from linkage disequilibrium. *Genome Research*, Volume **17**, Issue 4, pp 520-526.
144. The Delphi Inspiration. (2008). DIContainers, version 3.0.0. Available on: www.yunqa.de
145. Troy, C. S., Machugh, D. E., Bailey, J. F., Magee, D. A., Loftus, R. T., Cunningham, P., Chamberlain, A. T., Sykes, B. C. & Bradley, D. G. (2001). Genetic evidence for Near-Eastern origins of European cattle. *Nature*, Volume **410**, Issue 6832, pp 1088-1091.
146. Wang, W. Y. S., Barratt, B. J., Clayton, D. G. & Todd, J. A. (2005). Genome-wide association studies: Theoretical and practical concerns. *Nature Reviews Genetics*, Volume **6**, Issue 2, pp 109-118.
147. Weir, B. S. & Hill, W. G. (2002). Estimating F-statistics. *Annual Review of Genetics*, Volume **36**, Issue 1, pp 721-750.
148. Weir, B. S. (1996). *Genetic data analysis II: methods for discrete population genetic data*. Sinauer Associates.
149. Willhalm, O. S. (1937). A Genetic History Of Hereford Cattle In The United States. *J Hered*, Volume **28**, Issue 8, pp 283-294.
150. Witten, I. H., Frank, E. & Hall, M. A. (2011). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.
151. Witten, I. H., Frank, E., Trigg, L. E., Hall, M. A., Holmes, G. & Cunningham, S. J. (1999). Weka: Practical machine learning tools and techniques with Java implementations.
152. World Simmental Fleckvieh Federation, available on: <http://www.wsff.info/>, on

- date: 1.3.2012.
153. Wright, S. (1922). Coefficients of inbreeding and relationship. *The American Naturalist*, Volume **56**, Issue 645, pp 330.
 154. Wright, S. (1965). The Interpretation of Population Structure by F-Statistics with Special Regard to Systems of Mating. *Evolution*, Volume **19**, Issue 3, pp 395-420.
 155. Zmasek, C. M. & Eddy, S. R. (2001). ATV: display and manipulation of annotated phylogenetic trees. *Bioinformatics*, Volume **17**, Issue 4, pp 383-384.

11 Curriculum Vitae

Jan Říha
Maletín 4
Zábřeh na Moravě
789 01
Czech Republic

EDUCATION

1. 10. 2008 - present

University of South Bohemia, Faculty of Agriculture, Department of Genetics, Breeding and Nutrition. Ph.D. studies in Biotechnologies, Agricultural Biotechnologies.

1. 8. 2005 - 1. 8. 2010

Mendel University, Faculty of Agriculture, Department of Animal Morphology, Physiology and Genetics. Ph.D. studies in Animal Molecular Biology and Genetics.

1. 9. 2003 - 18. 5. 2005

Masaryk University in Brno, Faculty of Informatics. Master studies in Applied Informatics. Master thesis: Image Analysis and Machine Learning.

1. 9. 1999 - 30. 6. 2003

Masaryk University in Brno, Faculty of Informatics. Bachelor studies in Informatics. Bachelor thesis: Selection of Usable Text Segments for Native Voice Synthesis.

1. 9. 1995 - 30. 6. 1999

Gymnázium Šumperk, Secondary Grammar School education.

EMPLOYMENT

1. 10. 2010 - present

Dairy Research Institute Ltd. Research and Development in Life Sciences.

16. 8. 2010 - present

Bentley Czech Ltd. Executive manager.

1. 8. 2004 - 30. 10. 2011

Agriresearch Rapotín Ltd. Research and Development in Life Sciences.

1. 1. 2000 - 30. 10. 2011

Research Institute of Cattle Breeding, Ltd. Research and Development in Life Sciences.

R&D PROJECTS PARTICIPATION

NAZV - National Agency for Agricultural Research, Ministry of Agriculture of the Czech Republic

QF3024 Utilizing of biotechnological methods and genetics for effective rearing and breeding of beef cattle breeds. *Member of project team.*

QF3020 Utilisation of biodiversity between breeds of dairy and beef cattle and assignation

effects of hybridisation for increasing of performance of their crossbreds. *Member of project team.*

QC1083 Apparative methods of evaluation and classification objectivisation of carcass of cattle in SEUROP system. *Author of project proposal, member of project team.*

1G58073 The research and pilot validation of genomic methods applicable to the selection of quality and commercial utilisation of farm animals and their products. *Member of project team.*

QJ1210301 Research, new products and services to create the centre of mastitis prevention, detection and support of its treatment. *Author of project proposal, responsible project team member.*

Ministry of Education of the Czech Republic

2B06107 Standardisation of evaluation cattle carcass bodies by aparative methods according to SEUROP system. *Author of project proposal, responsible project coordinator.*

2B08037 Biotechnological methods for innovations of manufacturing and consumerist quality evaluation of beef as source of an animal proteins. *Author of project proposal, responsible project coordinator.*

MSM2678846201 Exercise of European model of multifunctional agriculture in less favoured areas (LFA) of the Czech Republic. *Member of project team.*

Czech Science Foundation

GD523/03/H076 Enhancing of Methodological Level and Theoretical Education of Students of the Ph. D. Program 4103V - Animal Husbandry - a Perspective Field General Animal Husbandary. *Member of project team.*

FRVŠ - University Development Fund

FRVŠ MŠMT 2280/2007 (hlavní řešitel) Optimalization of laboratory methods of DNA analysis for genetic diversity evaluation purposes in Czech horse breeds, interpretation of data. *Responsible project coordinator.*

MVČR - Ministry of the Interior of the Czech Republic

VG20102015023 Rapid Decision Systems for Food Safety. *Author of project proposal author, member of project team.*

TACR - Technology Agency of the Czech Republic

TA01010153 Technologies for routine measurement of meat quality parameters and controlled process of meat maturation to increasing of product added value and new products in meat industry. *Author of project proposal, responsible project coordinator.*

12 List of Publications

1. Říha, J. (2004). Agriculture in Czech Republic. In: Proceeding book of the 11th International Congress on Biotechnology in Animal Reproduction. 11th International Congress on Biotechnology in Animal Reproduction. Rapotín. ISSN 0139-7265.
Conference Proceeding
2. Říha, J. (2004). Software pro hodnocení jatečných těl podle normy SEUROP. In: Genetické základy šlechtění na kvalitu jatečných těl a hovězího masa. Genetické základy šlechtění na kvalitu jatečných těl a hovězího masa. Rapotín. ISBN 80-903143-6-8.
Conference Proceeding
3. Dvořák, J. and Gazdová, V. Říha J. (2005). Využití genetických markeru a molekulárne genetických metod pro šlechtění specializovaných masných plemen skotu. In: Využití genetických metod ve šlechtění skotu na masnou užitkovost a její ovlivnění faktory prostředí. Využití genetických metod ve šlechtění skotu na masnou užitkovost a její ovlivnění faktory prostředí. ISBN 80-903143-7-6.
Conference Proceeding
4. Dvořák, J. and Říha, J. (2005). Genetické markery, jejich současné a budoucí využití v chovu skotu. In: Využití genetických metod ve šlechtění skotu na masnou užitkovost a její ovlivnění faktory prostředí. Využití genetických metod ve šlechtění skotu na masnou užitkovost a její ovlivnění faktory prostředí. Rapotín.
Conference Proceeding
5. Bjelka, M., Šubrt, J., Říha, J., Homola, M., Filipčík, R., and Havlíčková, M. (2006). Vliv porážkové hmotnosti na parametry kvality JUT. In: Aktuální otázky produkce jatečných zvířat. Aktuální otázky produkce jatečných zvířat. ISBN 80-7157-976-9.
Conference Proceeding
6. Burócziová, M., Říha, J., and Humpolíček, P. (2006). Genetic diversity of the Czech and Slovak Thoroughbreds using 17 microsatellite loci. In: Book of Abstracts of the 57th Annual Meeting of the European Association for Animal Production. 57th Annual Meeting of the European Association for Animal Production. Wageningen. ISBN 90-8686-003-6.
Conference Proceeding
7. Burócziová, M., Říha, J., and Vrtková, I. (2006). Molekulárno-genetická charakteristika plemien Český teplokrvník a Slovenský teplokrvník. In: Sborník zemědělské fakulty JU. Biotechnologie 2006. České Budějovice. ISBN 80-8564-553-X.
Conference Proceeding
8. Burócziová, M. & Říha, J. (2006). The molecular-genetic characterization of Czech and Slovak Warmblood horse breeds. *Scripta medica*, Volume 79, Issue 6, pp 1211.
Journal
9. Cíváňová, K., Putnová, L., and Říha, J. (2006). Molekulárně-genetická charakteristika genového zdroje prasat v České republice. In: Sborník zemědělské fakulty JU. Biotechnologie 2006. České Budějovice. ISBN 80-8564-553-X.
Conference Proceeding
10. Hanuš, O., Vaculíková, J., Říha, J., Koza, M., Jedelská, R. & Suchá, S. (2006).

- Účinnost postgraduálního odborného vzdělávání v problematice chovu skotu a kvality jeho produktů -- zdraví dojnic a kvalita produkce mléka. *Výzkum v chovu skotu/Cattle Research*.
Journal
11. Humpolíček, P. and Říha, J. (2006). Aplikace molekulárně-genetických dat ve šlechtění hospodářských zvířat - seminář. Tauferova posluchárna, A31, budova A, 3. patro, AF, MZLU v Brně, Zemědělská 1, 613 00 Brno (CZ).
Conference
 12. Manga, I., Říha, J. & Dvořák, J. (2006). Porovnanie vplyvu markerov CSN3 a CSN2 na parametre mliečnej úžitkovosti českého strakatého a holštajnského dobytka testovaného pri prvej, piatej a vyšších laktáciách. *Acta Fytotechnica et Zootechnica*, Volume 9, Issue supplement, pp 13-15.
Journal
 13. Putnová, L., Vrtková, I., Říha, J., Burócziová, M., and Dvořák, J. (2006). The genetic structure of different horse breeds in the Czech Republic inferred from microsatellite markers. In: Proceedings of the 30th International Conference on Animal Genetics. 30th International Conference on Animal Genetics. Porto Seguro, Brazil: Belo Horizonte, Brazil. ISBN 85-85584-03-3.
Conference Proceeding
 14. Říha, J. and Hanuš, O. (2006). Návrh systému pro implementaci algoritmu pro grafické vyhodnocování výsledků složení a kvality bazénových vzorků mléka. In: Vliv výrobních faktorů a welfare na zdraví a plodnost dojnic a kvalitu a bezpečnost mléka jako potravinové suroviny. Vliv výrobních faktorů a welfare na zdraví a plodnost dojnic a kvalitu a bezpečnost mléka jako potravinové suroviny. Rapotín. ISBN 80-903142-6-0.
Conference Proceeding
 15. Říha, J. (2006). Software pro hodnocení JUT v rámci SEUROP, version [1.00]. Available on: Research Institute for Cattle Breeding. Rapotín.
Authorized software
 16. Project no. 2B06107 - SEUROP - Standardizace hodnocení jatečných těl skotu aparativními metodami v systému SEUROP/Standardisation of evaluation cattle carcass bodies by aparative methods according to SEUROP system. Říha, J. (2006-2011) Status: in progress. Budget value: 10 385 ths CZK. Funding agency: MŠMT ČR . Received by Research Institute for Cattle Breeding, Ltd.
Abstract: Proposing project deepens obtained knowledge and outputs with aim to maximum objective of classification carcass body processing (JUT) with regard to inter-breed and inter-sexual variances by carcass cattles. Exploitation acquired results Csoftwar) at the clasificcation system carcass body processing of cattle (JUT) and the control machanism applied at work ÚKZÚZ. Motivation of the proposal is unified and increase the precision of clasificcation with minimum influence human factor on final evaluaiton ofcattle carcass body as a raw material for processing industry.
Grant
 17. Říha, J., Frelich, J. & Šlachta, M. (2006). Vliv paternálního efektu na vybrané znaky masné užitkovosti skotu. *Acta Fytotechnica et Zootechnica*, Volume 9, Issue supplement, pp 150-154.
Journal
 18. Bezdíček, J., Říha, J. & Bjelka, M. (2007). Comparison of milk production of

- Czech Fleckvieh cows - sired by bulls in artificial insemination (AI) and sired by bulls in natural service (NS). *Výzkum v chovu skotu/Cattle Research*, Volume **II**, Issue 4, pp 34-40.
Journal
19. Bjelka, M., Šubrt, J., Simeonovová, J., Homola, M., Filipčík, R., Říha, J., and Dufek, A. (2007). Posudzovanie JUT býkov a volov metódou SEUROP a kvality ich mäsa metódami WB a TPA. In: Bezpečnosť a kontrola potravín. Bezpečnosť a kontrola potravín.
Conference Proceeding
 20. Bjelka, M., Homola, M., Říha, J., and Dufek, A. (2007). Vliv výživy a zraní na křehkost masa volů. In: DEN MASA 2007 „Masná užitkovost, skotu, koz a ovcí“. DEN MASA 2007 „Masná užitkovost, skotu, koz a ovcí“. Praha. ISBN 978-80-213-1645-4.
Conference Proceeding
 21. Burócziová, M., Říha, J., Židek, R., Trandžík, J., and Jakabová, D. (2007). Genetic structure of nine horse populations. In: Risk Factors of Food Chain. Risk Factors of Food Chain. Nitra. ISBN 978-80-8069-948-2.
Conference Proceeding
 22. Burócziová, M. and Říha, J. (2007). Microsatellite diversity of horse breeds in Czech republic. MendelNet'07 Agro. ISBN 978-80-7375-119-7.
Conference Proceeding
 23. Civaňová, K., Říha, J., and Dvořák, J. (2007). Horned and Polled Simmental Cattle in the Czech Republic. In: Book of proceedings from 2nd International Conference on Agricultural and Rural Development. Nitra. ISBN 978-80-8069-962-8.
Conference Proceeding
 24. Hadvová, S., Hanuš, O., Třináctý, J., Říha, J., Koza, M., Vaculíková, J. & Suchá, S. (2007). Dopad profesního vzdělávání v oboru chovu skotu – výživa dojníc a kvalita mléka (ekologické, zdravotní a hygienické faktory kvality a bezpečnosti mléka jako suroviny a potraviny). *Výzkum v chovu skotu/Cattle Research*, Volume **XLIX**, Issue 3, pp 54-62.
Journal
 25. Manga, I., Putnová, L., Říha, J., Vrtková, I., and Dvořák, J. (2007). Genetic characterization of the Czech Spotted cattle breed using panel of 10 microsatellite markers.
Generic
 26. Putnová, L., Vrtková, I., Hořín, P., Říha, J., and Dvořák, J. (2007). Genetic diversity in the genetic resource of Old Kladruber Horse using microsatellite DNA markers.
Generic
 27. Říha, J. & Burócziová, M. (2007). *Evaluation of Genetic Diversity*. pp. 1-90. DAMFG Mendel University, Brno.
E-Book
 28. Říha, J., Homola, M., and Burócziová, M. (2007). Parametry bourání jatečných těl skotu v rámci tříd zmasilosti SEUROP. In: Risk Factors of Food Chain. Risk Factors of Food Chain. Nitra. ISBN 978-80-8069-948-2.
Conference Proceeding
 29. Slezáková, M., Hegedušová, Z. & Říha, J. (2007). The influence of sires on the results of superovulation and embryo transfer in beef cattle. *Reproduction in Do-*

- mestic Animals*, Volume 42, pp 100.
Journal
30. Burócziová, M., Říha, J., Židek, R., Trandžík, J. & Jakabová, D. (2008). Genetic structure of nine horse populations in Czech and Slovak republic. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*, Volume LVI, Issue 2, pp 57-60.
Journal
 31. Civaňová, K., Kaplanová, K., and Říha, J. (2008). Genetic polledness in Simmental cattle in the Czech Republic. In: *Biotechnology 2008*. Biotechnology 2008. České Budějovice. ISBN 80-85645-58-0.
Conference Proceeding
 32. Civaňová, K., Říha, J., Sojková, K., and Dvořák, J. (2008). Molecular characterisation of Czech Red Spotted Cattle. In: *Biotechnology 2008*. Biotechnology 2008. České Budějovice. ISBN 80-85645-58-0.
Conference Proceeding
 33. Dufek, A., Bjelka, M., Šubrt, J., Simeonovová, J., Homola, M., Filipčík, R. & Říha, J. (2008). Effect of different feeding conditions and aging on meat tenderness in bulls. *Archiva zootechnica*, Volume 11, Issue 1, pp 6 pages.
Journal
 34. Hdrová, S., Hanuš, O., Bjelka, M., Třináctý, J., Říha, J., Pozdíšek, J., Koza, M., Suchá, S. & Kopecký, J. (2008). Dopad profesního vzdělávání v oboru chovu skotu – významné faktory kvality hovězího masa a jeho zpracování. *Výzkum v chovu skotu/Cattle Research*, Volume L, Issue 2, pp 11 pages.
Journal
 35. Hanuš, O., Genčurová, V., Říha, J., Vyletělová, M., Jedelská, R., and Kopecký, J. (2008). Specifika referenčních materiálů a výkonnostního testování způsobilosti výsledků uzákladních mlékařských analýz. In: *Referenční materiály a mezilaboratorní porovnávání zkoušek. III. Reference materials and interlaboratory investigation comparison*. Medlov. ISBN 978-80-86380-46-9, 53-78.
Conference Proceeding
 36. Hering, P., Hanuš, O., Říha, J., Klímová, Z., Sojková, K., Jedelská, R. & Kopecký, J. (2008). Test věrohodnosti stanovení počtu somatických buněk ze vzorků mléka pro kontrolu užitkovosti. Reliability test of somatic cell count determination in the samples for milk recording. *Výzkum v chovu skotu/Cattle Research*, Volume L, Issue 184 (4), pp 28-37.
Journal
 37. Manga, I., Říha, J. & Vrtková, I. (2008). Polymorphism of CSN3, Pit-1, LGB and its impact on milk performance trait at the Czech spotted and the Czech Holstein breed. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*, Volume LVI, Issue 1, pp 6 pages.
Journal
 38. Říha, J., Hanuš, O., Ledvina, D., Genčurová, V., Sojková, K., Jedelská, R. & Kopecký, J. (2008). Autorizovaný software AS 1 – MSM 2678846201, SomaRing, www.vuchs.cz/software/somaring; informace ve *Výzkum v chovu skotu*. *Výzkum v chovu skotu/Cattle Research*, Volume L, Issue 3 (183), pp 67.
Journal
 39. Project no. 2B08037 - AGROGEN - Biotechnologické metody pro inovace hodno-

cení zpracovatelské a spotřebitelské kvality hovězího masa jako potravinového zdroje živočišných proteinů/Biotechnological methods for innovations of manufacturing and consumerist quality evaluation of beef as source of animal proteins. Říha, J. (2008-2011) Status: in progress. Budget value: 11194 ths CZK. Funding agency: MŠMT ČR. Received by Research Institute for Cattle Breeding, Ltd. **Abstract:** Main goal of the project is to develop and validate recent biotechnological methods for innovations in prediction and estimation of beef quality with usage of genomical markers, breeding programs and feeding facilities. We want to develop recent methods for estimation of parameters of beef quality as CLA, marbling, photometric methods etc. In connection with DNA analysis of genomic markers of selected QTLs we want to develop new methods for prediction of beef consumer's, producer's and processor's quality what can be used just in neonatal period of calves' life. The part of project is also concerned to construction of MSs and SNPs panel what can be useful for identification and traceability of beef from bioproduction farms in CR.

Grant

40. Říha, J. and Bjelka, M. (2008). Marbling IA, version [1.00]. Available on: www.vuchs.cz/software/marblingia. Research Institute for Cattle Breeding, Ltd. Rapotín.

Authorized software

41. Říha, J. and Vrtková, I. (2008). Markery ve šlechtění skotu na maso. In BJELKA, M. Šetrné čerpání přírodních zdrojů a údržba krajiny pomocí chovu krav bez tržní produkce mléka. In: Šetrné čerpání přírodních zdrojů a údržba krajiny pomocí chovu krav bez tržní produkce mléka. Šetrné čerpání přírodních zdrojů a údržba krajiny pomocí chovu krav bez tržní produkce mléka. Šumperk. ISBN 978-80-87144-04-6.

Conference Proceeding

42. Říha, J. (2008). Software pro vizuální hodnocení intramuskulárního protučnění a marblingu hovězího masa pomocí analýzy obrazu – Marbling IA, MŠMT 2B06107. *Výzkum v chovu skotu/Cattle Research*, Volume 50, Issue 4, pp 2 pages.

Journal

43. Sojková, K., Říha, J. & Hanuš, O. (2008). Analysis of nitrogen fraction to amino acids profile in raw cow milk. *Výzkum v chovu skotu/Cattle Research*, Volume 11, Issue 3.

Journal

44. Sojková, K., Říha, J., Hanuš, O., Macek, A., and Jedelská, R. (2008). Analysis of relationships between health state indicators of dairy cows and profile of amino acids. In: Proteiny 2008. Proteiny 2008. Zlín. ISBN 978-80-7318-706-4.

Conference Proceeding

45. Sojková, K., Říha, J., Hanuš, O., Jedelská, R., and Kopecký, J. (2008). Analysis of relationships between somatic cell count and other quality indicators in goat and sheep milk. In: Sborník 11.mezinárodní konference Den mléka 2008. *Den mléka 2008*. Praha. 978-80-213-1822-9.

Conference Proceeding

46. Sojková, K., Říha, J., Hanuš, O., Třináctý, J., Hadrová, S., and Genčurová, V. (2008). Estimation of significant linear relationships with variability effects between amino acid profile and health state indicators in dairy cows. In: *Výživa dojnic*. Výživa dojnic. Pohořelice. ISBN 978-80-87144-02-2.

- Conference Proceeding*
47. Sojková, K., Říha, J., Manga, I., and Dvořák, J. (2008). Vliv polymorfismu genu CSN2 na počet somatických buněk českého strakatého a holštýnského mléčného skotu. In: Mléko a sýry 2008. Mléko a sýry 2008. Prague. ISBN978-80-7080-695-1.
- Conference Proceeding*
48. Voříšková, J., Šlachta, M., Říha, J., and Frelich, J. (2008). Rozhodovací a predikční modul chovu masných plemen skotu v horských oblastech - BEEFCAT1.0, version [1.0]. Available on: Asociace chovatelů masných plemen. ZF JČU. České Budějovice.
- Authorized software*
49. Vyletělová, M. and Říha, J. (2008). Analýza a identifikace mastitidních patogenů v ovčím mléce a určení jejich vztahu k ostatním ukazatelům. In: Sborník referátů z konference. *Den mléka 2008*. Praha.
- Conference Proceeding*
50. Vyletělová, M., Hanuš, O., and Říha, J. (2008). Mastitis occurrence in sheep milk and its relationship to other milk parameters. In: Abstract books. Symposium ISME: The 12 International symposium on microbial ecology. Cairns, Australia.
- Conference Proceeding*
51. Vyletělová, M., Hanuš, O., and Říha, J. (2008). The influence of milking type on hygienic indicators within dairy cows. In: Abstract books. XI. International congress IUMS 2008. Istanbul.
- Conference Proceeding*
52. Bezdíček, J., Říha, J. & Šubrt, J. (2009). American Dairy Association a výroční konference ADSA/CSAS/ASAS. *Výzkum v chovu skotu/Cattle Research*, Volume **LI**, Issue 3, pp 91.
With support of: MŠMT MSM 2678846201, MŠMT INGO LA09033, NAZV 1B44035.
- Journal*
53. Bezdíček, J., Šubrt, J., Filipčík, R. & Říha, J. (2009). Breeding Values of Fat and Protein Content in Inbred and Outbred Cows. *Journal of Animal Science/Journal of Dairy Science*, Volume **87**, Issue E-Suppl. 2/J, pp 207.
- Journal*
54. Bezdíček, J., Šubrt, J., Filipčík, R. & Říha, J. (2009). Evaluation of milkfat and milkprotein production in inbred and outbred Holstein cows. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*, Volume **LVII**, Issue 5, pp 19-25.
- Journal*
55. Bezdíček, J., Říha, J. & Bjelka, M. (2009). PastureCalc 1.0 - software pro aproximační kalkulaci nákladů oplocení pastevních areálů v různě definovaných podmínkách. *Výzkum v chovu skotu/Cattle Research*, Volume **LI**, Issue 4, pp 81-82.
- Journal*
56. Bezdíček, J. and Říha, J. (2009). Skot na Valašsku v LFA oblastech a jeho zpeněžení v systému SEURO. Zemědělské odborné dny pod záštitou Zlínského kraje, ve dnech 19. a 20. března 2009.
- Audiovisual Material*
57. Bjelka, M., Hanuš, O., Říha, J., Bezdíček, J., Pozdíšek, J., Suchá, S. & Koza, M. (2009). Dopad profesního vzdělávání v oboru chovu skotu - 6. část - šetrné čerpání přírodních zdrojů a údržba krajiny pomocí chovu krav bez tržní

- produkce mléka. *Výzkum v chovu skotu/Cattle Research*, Volume **LI**, Issue 2, pp 54-65.
Journal
58. Burócziová, M. & Říha, J. (2009). Horse breed discrimination using machine learning methods. *JOURNAL OF APPLIED GENETICS*, Volume **50 (4)**, Issue 4, pp 375-377.
This senior authors contributed equally to this work.
Journal
59. Dufek, A., Šubrt, J., Simeonovová, J., Říha, J., and Homola, M. (2009). Hodnocení efektu extenzivního výkrmu na ztrátu vody hovězího masa během zrání. In: Sborník příspěvků z konference Den masa 2009. Den masa 2009. Prague, CZ. 978-80-213-2005-5.
Conference Proceeding
60. Hanuš, O., Říha, J., Suchá, S., Kopecký, J., Koza, M. & Jedelská, R. (2009). Dopad profesního vzdělávání v oboru chovu skotu – výrobní zemědělská praxe a potravinářské biotechnologické úpravy pro zvýraznění pozitivních zdravotních vlivů mléka a mléčných výrobků. *Výzkum v chovu skotu/Cattle Research*, Volume **LI**, Issue 1, pp 49-62.
Journal
61. Hegedušová, Z., Říha, J., Bjelka, M., and Slezáková, M. (2009). Content of Acetone and Urea in Cervical Mucus in Cows.
Generic
62. Homola, M., Bezdíček, J., Říha, J. & Vacátko, E. (2009). Does the SEUROP trading class reflect portion of cutting parts in Czech Fleckvieh bulls? *Výzkum v chovu skotu/Cattle Research*, Volume **LI**, Issue 3, pp 11-18.
Journal
63. Kadlec, R. and Říha, J. (11-11-2009). Utility design no. 20429: Zařízení pro měření bioimpedance v syrovém mase/Device for measuring of bioimpedance in raw meat. Application no. 2009-21918. Agriresearch Rapotín Ltd. Czech Republic. Filed in Věstník úřadu průmyslového vlastnictví 3-2010 CZ, 20. 1. 2010.
Patent
64. Manga, I. & Říha J. (2009). Analýza spojitého efektu SNPs génů CSN3, CSN2, DGAT1 a Pit-1 na parametre mliečnej úžitkovosti holštýnskeho skotu. *Výzkum v chovu skotu/Cattle Research*, Volume **LI**, Issue 4, pp 12-21.
Journal
65. Růžička F., Mahelová M., Holá V., Kadlec R. & Říha J. (2009). Výskyt *Candida dubliniensis* v gastrointestiálním traktu a možnosti její identifikace. *Výzkum v chovu skotu/Cattle Research*, Volume **LI**, Issue 4, pp 69-73.
Journal
66. Říha, J., Manga, I., and Vyleťelová, M. (2009). ColonyCounter 1.0, version [1.00]. Available on: <http://www.vuchs.cz/software/ColonyCounter/>. Agriresearch Rapotín Ltd. Rapotín.
Authorized software
67. Říha, J., Manga, I., Bezdíček, J. & Šubrt, J. (2009). Effect of CSN2 gene polymorphism on somatic cell count in Czech Fleckvieh. *Journal of Animal Science/Journal of Dairy Science*, Volume **87**, pp 396-397.
Journal
68. Říha, J., Vrtková, I., and Dvořák, J. (2009). Effect of TG5 gene polymorphism on

- a basic chemical composition of beef. In: Book of Abstracts of the 60th Annual Meeting of the EAAP. 60th Annual Meeting of the EAAP. Netherlands. 978-90-8686121-7.
- Conference Proceeding*
69. Říha, J., Kadlec, R., Vondra, V. & Bezdíček, J. (2009). Experimental verification of the possibility to estimate sensoric and quality parameters of beef with use of bioimpedance. *Výzkum v chovu skotu/Cattle Research*, Volume **LI**, Issue 4, pp 38-49.
- Journal*
70. Říha, J. & Hanuš, O. (2009). GRADIM 0.1 - betaverze systému implementujícího algoritmus pro grafické vyhodnocování výsledků složení a kvality bazénových vzorků mléka. *Výzkum v chovu skotu/Cattle Research*, Volume **LI**, Issue 1, pp 63-67.
- Journal*
71. Říha, J., Bezdíček, J., Čítek, J., Voříšková, J. & Řehout, V. (2009). Interaction of genotypes of GH, Pit-1, CAPN1 genes and their influence on shear force of grilled beef in Czech Fleckvieh bulls during the period of maturation. *Výzkum v chovu skotu/Cattle Research*, Volume **LI**, Issue 4, pp 22-29.
- Journal*
72. Říha, J., Homola, M., and Hegedušová, Z. (2009). Marbling IA: software for fat content estimation on digital image of cutted beef. In: Book of Abstracts of the 60th Annual Meeting of the EAAP. 60th Annual Meeting of the EAAP. Netherlands. 978-90-8686121-7.
- Conference Proceeding*
73. Říha, J., Bezdíček, J., and Bjelka, M. (2009). PastureCalc 1.0, version [1.0]. Available on: <http://www.vuchs.cz/software/PastureCalc/>. Agriresearch Rapotín Ltd. Rapotín.
- Authorized software*
74. Říha, J., Hanuš, O., Ledvina, D., Genčurová, V., Sojková, K., Jedelská, R., and Kopecký, J. (2009). SomaRing, version [1.0]. Available on: www.vuchs.cz. Agriresearch Rapotín Ltd. Rapotín.
- Authorized software*
75. Říha, J., Bezdíček, J., Čítek, J., Voříšková, J., and Řehout, V. (2009). Vliv interakcí genotypů GH, Pit-1 a CAPN1 na sřížnou sílu grilovaného hovězího masa v 28 dnech zrání. In: Sborník příspěvků z konference Den masa 2009. Den masa 2009. Prague, CZ. 978-80-213-2005-5.
- Conference Proceeding*
76. Sojková, K., Říha, J., Hanuš, O., Jedelská, R. & Kopecký, J. (2009). Analýza vztahu mezi počtem somatických buněk a technologickými ukazateli kvality v kozím mléce. *Výzkum v chovu skotu/Cattle Research*, Volume **LI**, Issue 2, pp 24-28.
- Journal*
77. Sztankoová, Z., Mátlová, V., Kyselová, J., Jandurová, O., Říha, J. & Senece, C. (2009). Polymorphism of casein cluster genes in Czech local goat breeds. *Journal of Dairy Science*, Volume **92**, Issue 12.
- Journal*
78. Štýbnarová, M., Pozdíšek, J., Kohoutek, A., Říha, J. & Němcová, P. (2009). Srovnání stravitelnosti organické hmoty píce stanovované v laboratoři Výzkumného ústavu pro chov skotu, s.r.o. v Rapotíně a v laboratoři LFZ Raumberg-Gumpenstein, Rakousko. *Výzkum v chovu skotu/Cattle Research*, Volume **LI**, Issue

- 3, pp 77-83.
Journal
79. Třináctý, J., Křížová, L., Richter, M., Černý, V. & Říha, J. (2009). Effect of rumen-protected methionine, lysine or both on milk production and plasma amino acids of high-yielding dairy cows. *Czech Journal of Animal Science*, Volume **54**, Issue 6, pp 239-248.
Journal
80. Vyletělová, M., Říha, J. & Hanuš, O. (2009). Occurrence of mastitis infections in goat milk and their relation to other milk parameters. *Výzkum v chovu skotu/Cattle Research*, Volume **LI**, Issue 2, pp 14-18.
Journal
81. Židek, R., Jakobová, D., Trandžík, J., Gralak, B., Burócziová, M., Buleca, J., Massányi, P., Dvořák, J., Říha, J. & Zöldág, L. (2009). Analysis Of Hucul Horse Population Based On Molecular Genetic Data. *Magyar Állatorvosok Lápja*, Volume **131**, Issue 11, pp 685-691.
Journal
82. Bezdíček, J., Říha, J., Šubrt, J. & Vacátko, E. (2010). Comparison of the cutting parts portions of carcasses in beef and combined cattle. *Výzkum v chovu skotu/Cattle Research*, Volume **LII**, Issue 3, pp 8 pages.
Journal
83. Bezdíček, J. and Říha, J. (2010). Evaluation of the effect of inbreeding on age at first calving in holstein cows. In: ADSA Proceedings. American Dairy Science Association (ADSA) 2010. Toronto, CA.
Conference Proceeding
84. Bezdíček, J., Říha, J., Kučera, J., Dufek, A., Bjelka, M. & Šubrt, J. (2010). Relationship of breeding values and cutting parts of progeny in Czech Fleckvieh bulls. *Archiv Tierzucht*, Volume **2010**, Issue 53, pp 415-425.
Journal
85. Bezdíček, J. & Říha, J. (2010). The influence of multiple births on subsequent production and reproduction traits in Holstein cattle. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*, Volume **58**, Issue 4, pp 43-48.
Journal
86. Hegedušová, Z., Louda, F., Říha, J., and Kubica, J. (2010). Detekce říje v chovech skotu - cesta ke zlepšení úrovně reprodukce/Detection of estrus in cattle - a way to improve levels of reproduction. Application no. J03/10.
Certified method
87. Kaplanová, K., Říha, J., Vrtková, I. & Dvořák, J. (2010). Association of 5 candidate genes potentially affected beef quality with carcass traits and cutting parts in cross-breed cattle. *Výzkum v chovu skotu/Cattle Research*, Volume **LII**, Issue 1, pp 26-34.
Journal
88. Project no. VG20102015023 - SRRBP - Systémy rychlého rozhodování pro bezpečnost potravin/Rapid Decision Systems for Food Safety. Říha, J. (2010-2015) Status: in progress. Budget value: 18 821 ths CZK. Funding agency: MV ČR. Received by Dairy Research Institute, Ltd.
Abstract: The main objective of the project is to develop and pilot plant to verify the results of research systems to support rapid decision making when assessing the security risks of food stored in substandard conditions of crisis. Developed

systems will work as a tool for crisis management or other entities and enables fast and qualified decision. The systems can also be used as support tools for forensic analysis or optimization of food storage. DC1 automated technological processes in microbiological and analytical methods DC2 predictive quality systems and food security using the automated technological processes DC3 Adaptation and verification of the use of electromigration methods for rapid detection of microbial contamination of food with pathogenic microorganisms.

Grant

89. Říha, J., Bezdíček, J., Homola, M., Vacátko, E. & Šubrt, J. (2010). The influence of cattle breed on the portion of highly valued cutting parts of carcasses. *Journal of Animal Science/Journal of Dairy Science*, Volume **88**, Issue E-Suppl. 2/J, pp 360-361.
Journal
90. Sojková, K., Hanuš, O., Říha, J., Yong, T., Hulová, I., Vyletělová, M., Jedelská, R. & Kopecký, J. (2010). A comparison of lactation physiology effects at high and lower yield on components, properties and health state indicators of milk in Czech Fleckvieh. *Scientia Agriculturae Bohemica*, Volume **2**, Issue 41, pp 84-91.
Journal
91. Sojková, K., Hanuš, O., Říha, J., Yong, T., Hulová, I., Vyletělová, M., Jedelská, R. & Kopecký, J. (2010). A comparison of lactation physiology effects at high and lower yield on components, properties and health state indicators of milk in Czech Fleckvieh. *Scientia Agriculturae Bohemica*, Volume **41**, Issue 2.
Journal
92. Sojková, K., Hanuš, O., Říha, J., Genčurová, V., Hulová, I., Jedelská, R. & Kopecký, J. (2010). Impacts of lactation physiology at higher and average yield on composition, properties and health indicators of milk in Holstein breed. *Scientia Agriculturae Bohemica*, Volume **41**, Issue 1, pp 8 pages.
Journal
93. Šarovská, L., Dufek, A., Homola, M., Vacátko, E., Kubica, J., Bjelka, M., Bezdíček, J., Říha, J., Kousalová, L., Procházková, M., Voříšková, J. & Šubrt, J. (2010). Složení jatečných těl a křehkost masa. Rozbor jatečně upravených těl potomků testovaných býků českého strakatého skotu. *Chov skotu*, Volume **7**, Issue 6, pp 18-19.
Journal
94. Šubrt, J., Bjelka, M., Filipčík, R., Bezdíček, J., Dračková, E. & Říha, J. (2010). Variabilita celkového obsahu N a hydroxyprolinu v hovězím mase po dobu jeho zrání a v závislosti na základních chovatelských faktorech. *Výzkum v chovu skotu/Cattle Research*, Volume **LII**, Issue 4, pp 9 pages.
Journal
95. Watzková, J., Říha, J., Křížová, L. & Třináctý, J. (2010). Průzkum spotřebitelských postojů k mléku a mléčným výrobkům. *Mlékařské listy*, Volume **2010**, Issue 121, pp 7 pages.
Journal
96. Bezdíček, J. & Říha, J. (2011). Correlation analysis of production and reproduction traits in twins living under identical conditions - sent 12/2011. *Acta Fytotechnica et Zootechnica*.
Journal
97. Bezdíček, J. & Říha, J. (2011). Differences in the production and reproduction traits of embryo transfer full siblings living under different and identical condi-

- tions. *Journal of Animal Science/Journal of Dairy Science*, Volume 89, Issue E-Suppl. 1/J.
- Journal*
98. Bezdíček, J. & Říha, J. (2011). Genové a genotypové frekvence významných produkčních genů u českého strakatého skotu - sent 11/2011. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*.
- Journal*
99. Bezdíček, J. & Říha, J. (2011). Relationship between kappa-casein genotype in inseminated bulls and the milk composition of their daughters. *Journal of Animal Science/Journal of Dairy Science*, Volume 89, Issue E-Suppl. 1/J, pp 279-280.
- Journal*
100. Bezdíček, J. & Říha, J. (2011). Změny krevního obrazu laboratorních potkanů při zkrmování monodiety masa - sent 11/2011. *Animal Science*.
- Journal*
101. Filipčík, R. and Říha, J. (2011). Vliv hmotnosti jatečně upraveného těla býků na jatečnou hodnotu. In: *Animal Breeding*. Animal Breeding 2011. ISBN 978-80-7375-446-4.
- Conference Proceeding*
102. Kadlec, R. and Říha, J. (28-3-2011). Utility design no. 22000: Device for sterile measuring of biopedancy in raw meat. Application no. 2010-23723. Agri-research Rapotín Ltd. Czech Republic.
- Patent*
103. Křížová, L., Richter, M., Třináctý, J., Říha, J. & Kumprechtová, D. (2011). The effect of feeding live yeast cultures on ruminal pH and redox potential in dry cows as continuously measured by a new wireless device. *Czech Journal of Animal Science*, Volume 56, Issue 1, pp 37-45.
- Journal*
104. Manga, I. & Říha, J. (2011). The DGAT1 gene K232A mutation is associated with milk fat content, milk yield and milk somatic cell count in cattle. *Archiv Tierzucht*, Volume 54, Issue 3, pp 257-263.
- Journal*
105. Richter, M., Třináctý, J., Říha, J., Hegedušová, Z., and Pozdíšek, J. (17-1-2011). Utility design no. 21641: Zařízení pro kontinuální měření koncentrace amoniakálních iontů v bachoru přežvýkavců/Device for continuous measurement of ammonium concentration in rumen of ruminants. Application no. 2009-21641. Agri-research Rapotín Ltd. Czech Republic.
- Patent*
106. Říha, J. & Bezdíček, J. (2011). Breeding values and their relationship within cutting parts of progeny in beef bulls - sent 08/11. *Archiv Tierzucht*.
- Journal*
107. Project no. TA01010153 - beeftech - Technologie pro rutinní měření kvalitativních parametrů výsekového masa a pro kontrolovaný proces jeho zrání za účelem zvýšení přidané hodnoty výrobků a definice nových produktů v masném průmyslu/Technologies for routine measurement of meat quality parameters and controlled process of meat maturation to increasing of product added value and new products in meat industry. Říha, J. (2011-2014) Status: in progress. Budget

value: 9 159 ths CZK. Funding agency: TA ČR. Received by Agriresearch, Ltd.

Abstract: On the base of applied research, the aim of the project is to develop and prove technologies for routine controlling of chemical, nutritional, sensoric and technological quality of meat and technologies for optimal maturation of meat. Technologies will be applied in project participant production, they will use for new products definition for relevant marketing sectors.

Grant

108. Voříšková, J., Dufek, A. , Homola, M., Šubrt, J., Šarovská, L., Vacátko, E., Bjelka, M., Kubica, J., and Říha, J. (2011). Příručka se znaky JUT potomků testovaných býků českého strakatého skotu pro posuzování jejich masné užitkovosti.

Certified method

109. Project no. QJ1210301 - NAMC - Výzkum, nové produkty a služby pro vytvoření centra prevence, detekce a podpory léčby mastitid/Research, new products and services to create the centre of mastitis prevention, detection and support of its treatment. Říha, J. (2012-2016) Status: in progress. Budget value: 29 618 ths CZK. Funding agency: NAZV ČR. Received by Bentley Czech, Ltd.

Abstract: The aim of proposed project is usage of agriculture reseach and development to develop methods, products and procedures for detection, prevention and treatment support of mastitis in dairy cows under the conditions of the Czech Republic. Results of project will be created regarding to their routine usage in the national centre what will offer products and services to milk producers. All of project results will be proved regarding to their functionality as well as economical rentability of their application. In the project, we will use recent methods of molecular genetics, microbiology, pharmacy and agricultural economy. Users of results will be directly connected to project resolving.

Grant