TESTING ASCERTAINMENT BIAS AND RNA SECONDARY STRUCTURE MORPHOLOGY FOR PHYLOGENETIC SIGNALS

MARIENE, Grace Mukiri (BSc. Biochem & Mol. Biology, Jomo Kenyatta)

(Registration No: I56/71795/2008)

Thesis submitted in partial fulfillment for the award of Master of Science degree in Bioinformatics, University of Nairobi.

Centre for Biotechnology and Bioinformatics

October 2013

DECLARATION AND APPROVAL

I declare that this thesis is my original work, and has not been presented for an award at any other University. Sources of secondary information have been duly referenced.

`

Signature _____ Date _____

Grace Mukiri Mariene, Candidate

APPROVAL

This thesis has been submitted for examination with our approval as University supervisors

Signature _____ Date _____

Joel W. Ochieng, PhD

Center for Biotechnology & Bioinformatics, University of Nairobi

Signature _____ Date _____

Anne Wang'ombe, PhD

School of Mathematics, University of Nairobi

Signature _____ Date _____

Prof. James O. Ochanda

Centre for Biotechnology & Bioinformatics, University of Nairobi

ABSTRACT

Phylogenetic reconstruction is essential to many decisions in the medical and agricultural sciences. However, a common drawback in many of these studies is the failure of different datasets to recover the same phylogeny, using the same individuals. Such incongruence result mainly from factors inherent in the evolutionary process itself such as homoplasy or evolutionary 'noise', not adequately treated in many analysis programs available. The current study evaluated the usefulness of ascertainment bias (increase in microsatellite allele size range with evolutionary distance from focal taxon) as well as RNA secondary structure morphology in reconstructing accurate phylogenetic relationships. Two domesticated animal systems, one with an unresolved and often controversial evolutionary history, (the camel) and another with a well resolved phylogeny at the species level (cattle), were used to test the reliability of the two methods, and as a spinoff, to revisit the camel's unresolved history. Published camel and cattle microsatellite genotype data were used to test the utility of ascertainment bias, while cattle mitochondrial cytochrome *b* sequence data were obtained from a public repository at the National Centre for Biotechnology Information (NCBI). Allele frequency statistics, number of alleles and the allelic size ranges were estimated for each taxonomic group using Microsat toolkit. The means of the number of alleles and size ranges were determined, treating populations separately. The average of means, which is the mean of the means generated, was computed and compared with the mean of all, when the populations were combined. Secondary structures were predicted using MFOLD version 3.5, both at the default temperature (37^{o}C) and at 25^{o}C. The degree of congruence between predicted structures in different taxonomic groups were

compared, based on shapes, sizes (in bases) and positioning of hairpins, and lengths of helices. The predicted secondary structure morphologies compared in a manner reflecting evolutionary distances of major Bovine lineages. Whereas individuals within species were the most congruent followed by those between species within the genus, the most distant ones also differed the most, reinforcing their usefulness in resolving enigmatic phylogenies. However, in both test systems used in this study (Camelini and Bovini), ascertainment bias did not exhibit the uniformity required of a good phylogenetic probe. In many cases and for many loci, the principle (reduced allele size range proportional to evolutionary distance from the focal taxon) was not obeyed especially in the Bovini. This confirms that ascertainment bias may reflect phylogenetic trends in some systems but not others. The results of this study contradicted two major evolutionary, migration and domestication theories. The data suggested that first, unlike the current tenet that cattle (*Bos taurus* and *Bos indicus*) descended from the Auroch in Eurasia and then *B. indicus* migrated into Africa, and that Bos *javanicus* (banteng) and the Auroch shared a common ancestor, it is evident that *B. indicus* may have evolved independently from the Auroch in North Africa, making the indicine-taurine clade paraphyletic with respect to banteng. Second, this study suggested that the one humped and two humped camels did not simultaneously radiate from their common ancestor (*Paracamelus*) in western Asia, rather, it showed the dromedary to recently emerge from the Bactrian. These are interesting paleontological questions needing further examination from whole genome scans, as the current study relied on single genes.

DEDICATION

To my loving parents John and Charity Mariene, my sister Leah Wanja and nephew Jayden Amani.

ACKNOWLEDGEMENTS

I am forever grateful to the Author of life, for His mercies, love, and care throughout this study. Many thanks are due to my parents John and Charity for their support during the entire course and to my only sister Leah for being a role model to me.

My gratitude is due to my supervisors J.W. Ochieng, A. Wang'ombe and J.O. Ochanda for their patience and guidance on all aspects of this study. I am particularly thankful for their constructive comments that stimulated me to think broadly as I undertook the work.

I also owe much appreciation to my friends and classmates for their continued moral support that urged me on. Thank you all for your love, prayers, encouragement and presence. God bless you abundantly.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

1.1 General introduction

Most studies in agriculture, medicine, conservation and related disciplines require knowledge of genealogical relationships between taxa. In agriculture and specifically in conventional breeding, closely related species that possess unique desired genetic characteristics, are often used to transfer the traits to a target species through hybridization. However, not any species can hybridize with another. Two species can only hybridize if they do not have pre-zygotic (such as geographic isolation, behavioral isolation, and mechanical isolation) and post-zygotic reproductive isolation (such as hybrid depression). In determining whether species can hybridize, it is common practice to ascertain the genealogy as closely related species are expected to hybridize more often than the more distantly related ones. Phylogenetic analysis is often applied in this determination. Similarly, herbicide use depends on relationships among angiosperms as species closely related to weeds may be affected during spraying. In medicine, phylogenetic analysis can help in the prediction of drug response, when drugs are to be used across related taxonomic groups. Further, phylogenetic analysis is often used in drug discovery and efficacy trials using model organisms.

Despite the critical need in agriculture, medicine, and industrial undertakings, a phylogeny is simply a hypothesis on the evolutionary relationships among taxa, as such it does not always reflect the true relationships. Some of the reasons why the hypothesized tree can be incongruent with the true tree include those inherent in the biology of the organisms as well as those resulting from workmanship. Causes of

incongruence due to the latter include inadequate and non-judicious sampling and choice of non-optimal computational methods. These can easily be resolved. Incongruence due to biological reasons are more stubborn and include historical hybridization or admixture, incomplete lineage sorting into descendant and now extant taxa, presence of homoplastic characters, which is perhaps the greatest nuisance in phylogenetic analysis under a deep divergence scenario. Methods have been proposed that take into account and correct some of these factors during phylogenetic reconstruction. However, some of the problems specifically the ones that have their origin in the genealogy require methods that can deal with both historical and biological confounds such as those that avoid the bifurcating trees as well as those that can minimize the effect of natural selection on loci.

1.2 Problem statement

Computer based programs have been developed for the reconstruction of phylogenetic relationships among taxa. However, existing methods are reliable only when the divergence among taxa is moderate. In the extreme cases where divergences are especially deep, they have proved unreliable. This infidelity has misled phylogenetic inferences for many life forms, leading to mismanagement with varying degrees of severity, from taxonomic lumping in biological conservation to misdiagnosis in human medicine. Especially in medicine where human health is concerned and hence accurate phylogenies are essential to understand pathogen lineages; accurate algorithms whose fidelity transcends hierarchical levels of genetic divergences are required.

1.3 Research Objectives

Overall objective

To find a more accurate and reliable biological signal/character for phylogenetic reconstruction in order to improve the determination of species relationships

Specific objectives:

1. To test the diagnostic value of Ascertainment bias as a phylogenetic probe
2. To test whether RNA secondary structure morphology is a valuable phylogenetic signal in taxa with deep separations
3. Resolve enigmatic phylogenies of the Bovini and Camelini tribes

1.4 Hypothesis

I hypothesize that morphology variations on RNA secondary structures among taxa is proportional to their evolutionary distances, and ascertainment bias (absolute allele sizes in focal species being often greater than that found in related species) reflects the relative phylogenetic distance from the focal taxa.

CHAPTER 2: LITERATURE REVIEW

2.1 Background

Identifying close relationships among taxa in many life forms, is particularly important in decisions such as in conservation biology to outline different taxonomic groups, in agriculture where closely related groups are hybridized for improvement, in the manufacture and application of antimicrobials that are used cautiously across taxonomic groups. In medical applications, reconstruction of phylogenies has enabled the prediction of the evolution of antibiotic resistant genes (Hall, 2004), and thus helping in the defense of rapidly mutating viruses and test new models of evolution. Vaccines developed for a species, in the pharmaceutical industry, are advantageously applied to closely related taxa. Further, knowledge of historical relationships is important in both paleontology and archaeology, and in understanding evolution.

Phylogenies reflect the evolutionary path of speciation, leading to current relationships among extant taxa. During speciation, certain characters leave marks on the path of evolution. Thus fossils, morphological characters and molecular data can be used to study evolutionary relationships. This is possible as evolutionary histories of genes mark functional demands to which they have been subjected and hence can elucidate functional relationships within living cells (Gu, 2001; Zhu et al., 2000). This informs the increasing use of phylogenetics by pharmaceutical companies to make functional predictions like in drug discovery, where natural ligands have to be predicted for cell surface receptors that are usually targeted (Chambers et al., 2000).

Fossil record is one of the earliest character states used to root phylogenetic trees. However, fossil record is in many cases incomplete or fragmented. Other characters in use today include quantitative, ecological (such as behavior, host commonality), biochemical data (such as blood group, tissue oil content, toxicity), morphological features (such as presence or absence of fins, number of legs, lengths of legs, etc) and molecular data. Molecular characters are the most commonly used phylogenetic signals today. Molecular data has several advantages over other data systems: (1) It is more abundant. (2) There exists varying evolutionary rates across nucleotides or amino acid residues enabling the analysis of distantly related species (3)molecular patterns are well established - four nucleotide bases and 20 amino acids (4) Non-invasive sampling (such as fecal samples) involving cryptic and endangered species, or where human disturbance would have a negative impact is possible with molecular data (5) damage to diagnostic morphological characters can limit its use (6) late onset morphological characters such as those diagnosable only at adult stage in long live species is another hindrance.

Reservations have been raised regarding the use of morphology due to the fear that its characters are subjective and could vary in the way they are measured or interpreted (such as pubescence or colour, texture). Further, some of the characters used may show phenotypic plasticity (Pigliucci et al., 2006) but this may be difficult to ascertain. Indeed, some characters vary according to the geographic locations or micro niche variations (Martin and Pitocchelli, 1991) as a result of either plasticity or adaptive divergence. However, they have often provided reliable phylogenetic signals, even

where DNA based phylogenies are misleading (Wiens et al., 2010). Despite significant chorus condemning the use of morphology in phylogenetics, there are a significant number of cases where molecular data has failed to resolve organismal phylogenies (e.g. Russo et al., 1996; Parra-O et al., 2006; Ochieng et al., 2007a). Some of the reasons why DNA data may fail to resolve organismal phylogeny include: (1) homoplasy, (2) paralogy (3) hybridization (4) Incomplete lineage sorting (5) inadequate phylogenetic characters (6) Selection at gene loci. Other factors include sub-optimal tree reconstruction methods and non-judicial sampling. However, homoplasy seems the greater nuisance to phylogenetic analysis.

Homoplasy is the similarity in state among sequences but whose basis is not their common ancestry; rather, it arises as a result of convergent or parallel evolution, or by stochastic processes. Within a technological and analysis framework, it occurs as a result of insertion and deletions in the flanking region making alleles similar in state but not necessarily by descent. The possibility of harboring such similarity is higher in older populations that have had time to accumulate them; thus homoplasy is expected to increase with evolutionary distances, posing a greater nuisance in lineages that have diverged for millions of generations. This is why for deep divergences, homoplasy increases to obscure phylogenetic signals (Ochieng et al., 2007a).

Reconstructed phylogenetic trees using the characters available at the time is a hypothesis of the evolutionary lineage of taxa. This hypothesis can sometimes vary from the true history. Incongruence between phylogenetic trees that arise as a result of

computational and sampling problems can be resolved by choosing appropriate tree reconstruction methods and judicious representation of major clades in the analysis, and by using adequate number of characters. However, incongruence that arises from biological reasons such as homoplasies cannot be resolved simply by computational manipulations or by sampling. These cases require methods that deal with elements that are retained during evolution and can hence eliminate 'noise' characters. Given that evolution is a heritable change in genetic makeup of populations brought about by forces such as selection and gene flow, retained elements would thus include functional constraints in the RNA secondary structures and directional evolution that has the potential to affect the size of gene fragments following divergences, such as in ascertainment bias.

2.2 RNA secondary structure

While DNA occurs as a double helix, RNA is usually single stranded, but can fold back onto itself to form helices and hairpins usually referred to as secondary structures. The number of intramolecular hydrogen bonds between complementary bases, G-C, versus U-A or G-U determines the amount of free energy used or released in forming these base-pairs and hence affecting the structure's stability. The RNA tertiary structure is stabilized by the intramolecular interactions between the existing secondary structural motifs (Wu and TinocoJr, 1998). Positive free energy uses up the energy while negative free energy releases the energy stored leading to the likely formation of a secondary structure. Most prediction algorithms thus make use of base pair configurations that have the least possible free energy to predict secondary structures of RNA.

Transcription occurs in the helices while the hairpins act to terminate transcription processes. This is why variations within sequences that form the helix are seldom tolerated and in many cases accompanied by compensatory mutations in the complementary bases to maintain the secondary structure stability. This explains why most variations among individuals within a species are found within the hairpins and not the helices (Ochieng et al., 2007a). Different techniques can be used to predict RNA secondary structure. First using graphical representations, all possibilities can be assessed. Second, the laws of thermodynamics can be used to compute a conformation of minimum free-energy. Third, the phylogeny approach can be used if the sequences for functionally identical molecules have been determined for several organisms or organelles (Zuker and Sancoff, 1984). A common secondary structure can be assigned to species with similar functions or closely identical structures. It thus appears that the integrity and hence shape of RNA secondary structures can be an indicator of evolutionary distances among taxa.

2.3 Ascertainment bias

Ascertainment bias describes the observation that the range in allele size of microsatellite DNA (length of repeats) is often smaller in related species than in the focal taxon from which they were isolated; and this difference increases with evolutionary distance from that focal taxon (Ellegren et al., 1995; Rubinstein et al., 1995). This systematic variation in size range has been suggested to result from directional evolution (Rubinstein et al., 1995) or a bias during microsatellite isolation for population analysis (Ellegren et al., 1995). This phenomenon has not been utilized in

phylogenetic analysis; however, variation in microsatellite allele size range has been shown to display a pattern congruent to evolutionary divergence of species within a family (Ochieng et al., 2007b).

CHAPTER 3: ASCERTAINMENT BIAS AND ALLELE DISTRIBUTION PATTERNS SHOW DROMEDARY TO ORIGINATE FROM THE BACTRIAN

3.1 INTRODUCTION

Phylogenetic relationships, origin and domestication of camels remain unresolved. The family Camelidae comprises two genera: Camelus and Llama. The division between Llama and Camelus dates about 30 MYA according to paleontological evidences (Wilson, 1984) or 11 MYA following mitochondrial DNA studies (Stanley et al., 1994). Fossil camels of the genus Camelus descend from Plio-Pleistocene forms of the genus paracamelus, recorded in north-eastern China, north-western Mongolia, Tadzhikistaan and Kazakhstan [Young, 1932; Haveson, 1954 (cited in Peters and Von Driesch, 1997)]. It has been proposed that the domestic Bactrian and dromedary each descend from a wild species (Peters and Von Driesch, 1997), and that the wild camel (Camelus ferus) did not share a common ancestor with the domestic Bactrian (Ji et al, 2008). The separation between Bactrian and dromedary apparently occurred recently, as cross breeding between them produce fertile offspring that exhibit hybrid vigour (Peters and Von Driesch, 1997).

The evolutionary history of the camel remains unresolved perhaps due to deep level of divergence, migration patterns and multiple domestications. Of the three species, the wild Bactrian exhibits highest levels of genetic diversity, followed by the domestic Bactrian (Jianlin et al, 2000). The comparatively low diversity in the dromedary might suggest that this group were the last to emerge. Past research on proteins (Penedo et

al., 1988) and satellite DNA (Vidal-Rioja et al., 1987) has provided little information on the evolution of the family. Thus this group provides an excellent model for evaluating the diagnostic value of other phylogenetic probes.

Ascertainment bias describes the observation that the range in allele size of microsatellite DNA (length of repeats) is often smaller in related species than in the focal taxon from which they were isolated; and this difference increases with evolutionary distance from that focal taxon (Ellegren et al., 1995; Rubinstein et al., 1995). This systematic variation in size range has been suggested to result from directional evolution (Rubinstein et al., 1995) or a bias during microsatellite isolation for population analysis (Ellegren et al., 1995). This phenomenon has not been utilized in phylogenetic analysis; however, variation in microsatellite allele size range has been shown to display a pattern congruent to evolutionary divergence of species within a family (Ochieng et al., 2007b).

This study sought to evaluate whether ascertainment bias and allele distribution patterns at nuclear microsatellite loci would provide diagnostic phylogenetic evidence of relationships among the three camel species without the use of tree methods. I wanted to ascertain whether the two domestic species diverged independently from a common ancestor as suggested in phylogenetic trees, and whether the wild camel is indeed the ancestor of the domestic Bactrian (I am aware that in *sensu stricto*, no living taxon can be 'ancestor' of another living species).

3.2 MATERIALS AND METHODS

3.2.1 Source of data for analysis

This study utilized published camel microsatellite genotype data (Jianlin et al, 2000; Ochieng, 2002; Mburu et al, 2003) to test the utility of ascertainment bias as a reliable phylogenetic signal in the genus with ambiguous evolutionary history that is still under discussion. The two datasets used in this study were credibly generated for different kinds of study, and the results of which have been published: Science (Hanotte et al., 2002) for the cattle SSR dataset; Jianlin et al, 2000; Mburu et al, 2003; and dataset for publication under preparation availed by my supervisor JW Ochieng).

3.2.2 SSR statistical analyses

Input files based on the primers and phylogroups used to generate different SSR data loci was prepared on the MS excel sheet and imported into the Microsatellite (Microsat) toolkit for Microsoft excel (Stephen D.E Park) available at the Trinity College, Department of Genetics, in a two column data format. All the populations were included. The allele frequencies and statistics were determined, firstly by treating populations separately and by combining them. Allelic counts, showing the number of alleles and the allelic size ranges, which is the difference between the largest and smallest allele, were estimated for each taxonomic group. Variance in allele size for each locus per taxon was computed from MS Excel spreadsheet. Cumulative variance was considered as the sum of single locus variances, taking allele sizes (in bp) as values. The means of the number of alleles and size ranges while populations were treated separately were

determined. The average of means, which is the mean of the means generated above was computed and compared with the mean of all, when the populations were combined (Ochieng et al., 2007b).

As a control analysis, genetic distances and conventional phylogenetic analysis were implemented using allele size variations. Genetic distances based on allele size variation are modeled on the premise that when a mutation occurs, the new mutant is related to the allele from which it was derived. In this case, the difference in length between alleles contains phylogenetic information (Goldstein et al., 1995). Two different measures were employed to estimate the between individual genetic distance: the average square distance (*D1*) of Goldstein et al.(1995), and Nei's (1972) standard genetic distance (*D*). The average square distance accounts for size homoplasy, and is suitable for reconstructing trees that include more distantly related taxa. Both distances were computed using the MICROSAT program available from the Human Population Genetics Laboratory (HPGL), Stanford University, with the option of either exhaustive or 100 bootstrap replicates. The allele sizes analyzed are nucleotide counts rather than repeat scores, using the option that allows for repeat lengths = 2. Duration of linearity was calculated for each locus and averaged over loci. The primer error (size of the region flanking the SSR), when found, was entered and corrected for, by assuming a default of no error (*i.e.*, 0 nucleotides).

3.3 RESULTS

3.3.1 Ascertainment bias at camel SSR

The 12 markers used were variable, having a total of 173 unique alleles in 503 samples representing three species. The most variable locus was CV1 with 37 unique alleles while the least variable locus was V32 with 5 unique alleles (Table 3.1). Domestic Bactrian (DB) showed the greatest intragroup diversity, by cumulative variance and the mean number of alleles (cumulative variance=49.38; MNA = 10.4) as compared to Wild Bactrian (WIL) (cumulative variance = 39.79; MNA = 4.8) or Dromedary (DR; cumulative variance = 37.62; MNA = 9.1). There was variation between the average of means, which is the mean of the means generated from each species, **21 (8.1)** and the mean of all, when the populations were combined, **34.8 (14.4)**. The wild Bactrian had a higher cumulative variance (39.79) despite its smaller sample size (15) compared to dromedary (37.62) with a larger sample size (332). A comparison between dromedaries and bactrian for loci isolated in dromedary genomes (VOLP 8, VOLP10, VOLP 32 and CVRL 1, CVRL 2, CVRL 5, CVRL 6) showed allele size ranges to increase in bactrians compared to the focal taxon (Table 3.1). However, allele size ranges observed in the dromedary decreases in the bactrians for a locus isolated from the Llama genome (LC66; Table 3.1).

Table 3. 1 Allele size ranges, number of alleles per locus (in parentheses) and cumulative variances in three camel species. The table shows allele size ranges to consistently increase in bactrians for loci isolated in dromedaries, suggesting negative ascertainment bias – dromedaries seemingly a recent offshoot from the Bactrian.

Marker	Size range and number of alleles (in parentheses)			
	W.BACT	**D.BACT**	**DROM**	**ALL**
V8	28 (6)	38 (16)	06 (05)	38 (20)
V10	28 (6)	28 (13)	18 (09)	36 (16)
V32	06 (3)	54 (06)	02 (02)	54 (05)
Y8	10 (5)	22 (11)	40 (17)	40 (19)
Y38	12 (5)	12 (06)	11 (06)	13 (09)
Y44	14 (4)	16 (08)	22 (05)	26 (10)
CV1	26 (8)	65 (30)	40 (20)	65 (37)
CV2	14 (4)	10 (06)	06 (04)	14 (07)
CV5	08 (4)	26 (10)	24 (13)	30 (15)
CV6	16 (6)	20 (07)	06 (04)	20 (07)
CV7	08 (4)	08 (05)	44 (16)	50 (17)
LC66	12 (3)	28 (07)	30 (08)	32 (11)
Mean	15.1 (4.8)	27.3 (10.4)	20.6 (9.1)	34.8 (14.4)
Average of means	**21(8.1)**	NA	NA	**34.8 (14.4)**
Sample size	15	156	332	503
Cum Variance	39.79	49.38	37.62	71.45
Total no. of alleles	58	125	109	173

Allele distribution patterns at SSR locus VOLP-08

Figure 3.1 gives the frequencies of the different alleles present in two populations of Wild camels, *Camelus bactrianus ferus* (n = 12), seven populations of domestic Bactrian, *Camelus bactrianus* (n = 156), two Arabian dromedary, *Camelus dromedaries* (n = 32), and four Kenyan dromedary breeds *Camelus dromedarius* (n = 265) at microsatellite locus (*VOLP-08*). At this locus, the two main camel groups, dromedary and Bactrian, do not share any alleles as shown in the figure. This locus was isolated from the genome of the dromedary.

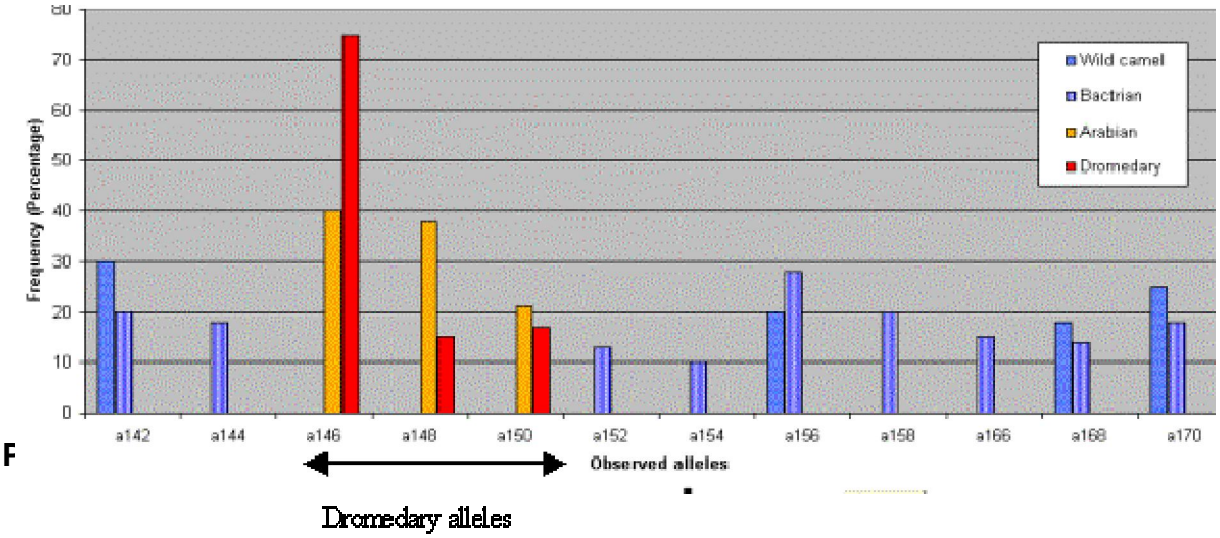


Figure 3. 1 Allele sizes and frequencies at VOLP-08 locus in wild bactrian domestic bactrian and dromedary. Alleles in the two bactrians 'avoid' areas occupied by dromedary alleles, and showing dromedary alleles to be a subset (recent offshoot) from the bactrian.

Phylogeny of camel species based on tree methods

A phylogeny of the three camel species (wild Bactrian, domestic Bactrian and dromedary) reconstructed from SSR loci did not resolve their evolutionary relationships. Instead, the cladogram appeared polytomic – where the three clades appeared to diverge from a common ancestor at the same time (Figure 3.2). This was the same observation regardless of the distance method and tree model.

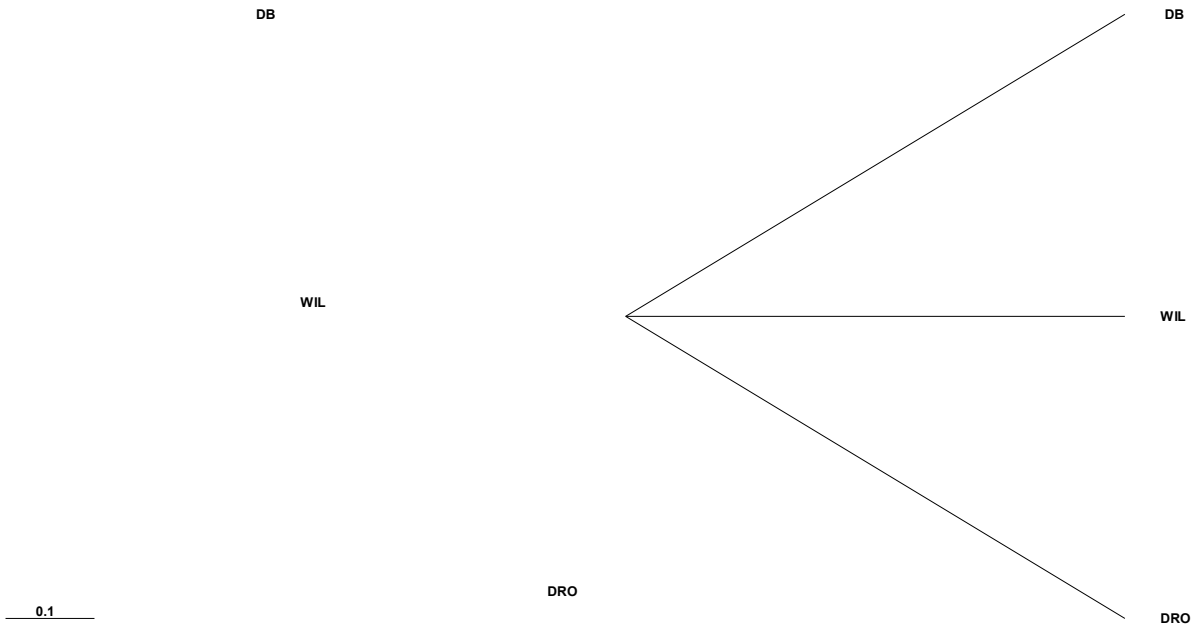


Figure 3. 2 (A and B) Camelini phylogeny. Relationships between the three camel species are equivocal and polytomic in a phylogenetic framework with the three clades branching out at the same time. Camels therefore show an unresolved phylogeny.

3.3.2 Ascertainment bias in Bovini is unreliable

The 15 markers used were variable, having a total of 174 unique alleles in 1286 samples representing 4 species. The most variable locus was ILSTS 036 with 23 unique alleles while the least variable locus was TGLA 48 with 4 unique alleles (Table 3.2). *Bos taurus* (TAU) had the greatest intragroup diversity with the cumulative variance =29.09 and mean number of alleles (MNA) = 10.1 compared to the *Bos indicus*, IND, (cumulative variance=27.69; MNA=10.2), the admixture group, ADM (cumulative variance=26.55; MNA=9.4), and the European *taurus*, TAUEURO (cumulative variance=26.34; MNA=7.5). Analysis showed variation between the average of means (mean of the means generated from each species) = **21.97(9.3)**) and the mean of all, (when the populations were combined) = **26.4 (11.6)** (Table 3.2). For loci TGLA 126,

TGLA 227, TGLA 48 and TGLA 122, isolated from the *Bos taurus* genome, there was either a fixed or a slight increase in the allele size range from the *taurus* to *indicus*. For loci denoted ILST, isolated from the *indicus* genome, the allele size ranges either showed an increase or a decrease across the four species (Table 3.2).

Table 3. 2 Allele size ranges, mean number of alleles and cumulative variances in four bovid species. The table shows allele size ranges to remain fixed or either increase or decrease for the three loci isolated from the *taurus* or *indicus*, suggesting that for the Bovidae, ascertainment bias is unreliable.

Marker	**Size range and number of alleles (in parentheses)**				
	ADM	**IND**	**TAU**	**TAUEURO**	**ALL**
TGLA 126	14 (08)	14 (08)	14 (08)	13 (07)	14 (08)
AGLA 293	30 (15)	32 (15)	26 (13)	28 (11)	32 (16)
TGLA 227	22 (04)	22 (04)	22 (04)	22 (05)	22 (05)
ILSTS 005	12 (06)	12 (06)	12 (06)	02 (02)	12 (06)
MGTG 4b	24 (12)	24 (13)	24 (13)	24 (12)	24 (13)
ILSTS 006	24 (12)	20 (11)	20 (11)	16 (09)	24 (13)
ILSTS 103	22 (11)	20 (10)	24 (12)	14 (07)	44 (13)
ILSTS 023	16 (07)	20 (09)	24 (10)	18 (07)	24 (11)
TGLA 48	04 (03)	04 (03)	06 (04)	04 (03)	06 (04)
ILSTS 036	48 (15)	42 (17)	38 (19)	20 (10)	48 (23)
ILSTS 50	22 (11)	34 (13)	34 (13)	16 (08)	34 (14)
TGLA 122	36 (17)	46 (20)	38 (17)	44 (16)	46 (22)
ILSTS 008	12 (07)	10 (06)	10 (05)	05 (03)	12 (07)
ILSTS 028	28 (09)	30 (14)	28 (12)	30 (08)	30 (14)
ILSTS 033	24 (04)	24 (05)	24 (05)	24 (05)	24 (05)
Mean	22.5(9.4)	23.8(10.2)	22.9(10.1)	18.7(7.5)	26.4(11.6)
Average of means	**21.97(9.3)**	NA	NA	NA	**26.4(11.6)**
Sample size	210	554	384	138	1286
Cum Variance	26.55	27.69	29.09	26.34	28.8
Total no. of alleles	141	154	152	113	174

Tree phylogeny show *Bos taurus* and *Bos indicus* to be sister species

Phylogenetic analysis using SSR loci grouped the admixed with the *indicus*, showing a closer relationship with the latter compared to *taurus* (Figure 3.3). The cladogram also showed sister relationship between the *taurus*, *European taurus* and the *indicus*. This kind of relationship is expected since *Bos taurus* and *Bos indicus* are considered sister species within the bovini family.

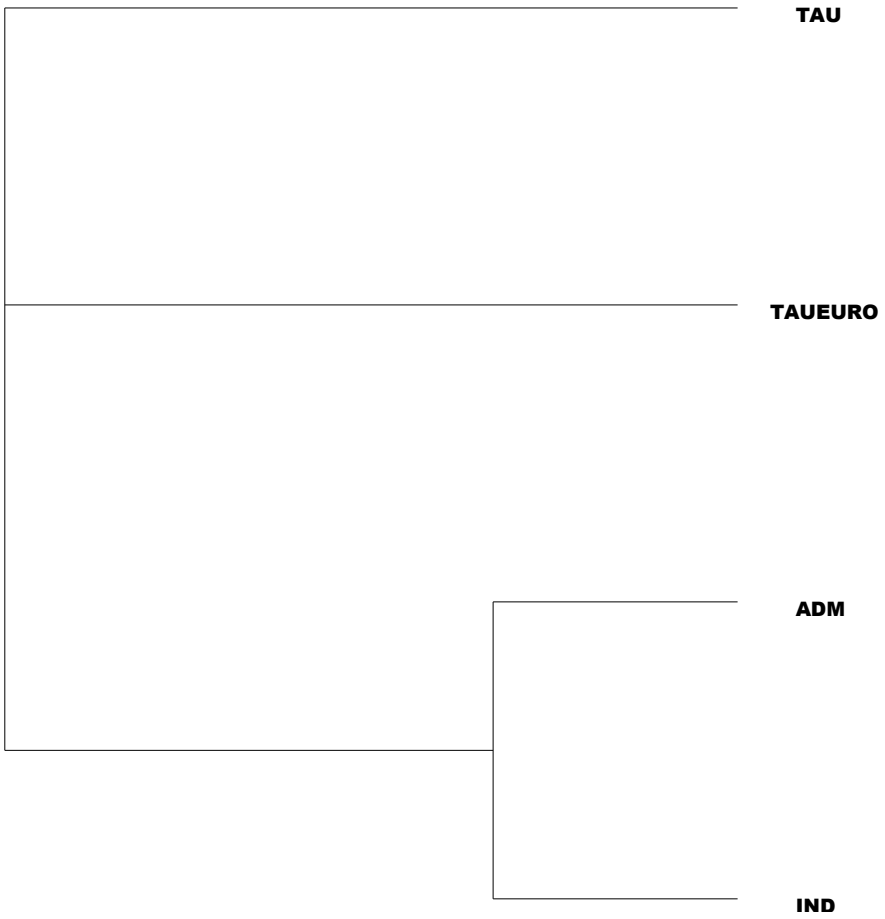


Figure 3. 3 Bovine phylogeny at SSR loci. The admixed group are grouped with the indicine, while the *taurus* and *indicus* are sister species. Taurine samples of Europe are polytomic with respect to other taurine and indicine groups.

3.4 DISCUSSION

This study sought to determine the usefulness of ascertainment bias in inferring evolutionary relationships among taxa, using a camel model. Several hypotheses have been proposed to explain the origin and divergence of extant camels: wild Bactrian, domestic Bactrian and dromedary. While some scholars propose that Bactrian and dromedary diverged simultaneously from a common ancestor (Peters and Von Driesch, 1997), others have recognized the wild Bactrian to be the ancestor of the domestic groups. Phylogenetic analyses using proteins and DNA variations have not provided unequivocal relationships among these taxa. Such a system provided a model to test the reliability of ascertainment bias (explained earlier) in phylogenetic inference. Microsatellite allele ascertainment bias as well as distribution patterns showed an unusual relationship, suggesting that the dromedaries 'descended' from the bactrians. This new hypothesis is tenable considering both direct and circumstantial evidence.

3.4.1 Camel SSR show dromedary to be a more recent lineage

Variation between the average of means (mean of the means generated from each species) and the mean of all (when populations were combined) in the camel SSR data (Table 3.1) suggested the occurrence of ascertainment bias. The first indication that dromedaries could have 'descended' from the bactrian was evidenced by observation of negative ascertainment bias. Compared to the focal taxon, allele size ranges increased in bactrians for markers isolated in dromedary genomes. However, size ranges observed in the dromedary decreased in the bactrian for a locus isolated from the Llama

genome (LC66; Table 3.1). An ancestral population or species is expected to be more diverse compared to emergent taxa. The bactrian showed a high level of diversity such that even with a smaller sample size, the mean nears that of the dromedary.

3.4.2 Allele distribution patterns show dromedary to emerge from Bactrian

Allele distribution patterns was analysed at 12 SSR loci, with one isolated in the dromedary (VOLP-08) showing the two main camel groups, dromedary and Bactrian, not to share any alleles. At this locus, the more widespread alleles in the wild and domestic Bactrian clustered on either side, 'avoiding' areas occupied by dromedary alleles. This pattern suggested that the two species did not descend simultaneously from common ancestor as proposed from paleontological evidence; rather, it suggests that dromedary camels recently descended from the Bactrian. Under this scheme, the ancestral camel (Paracamelus) owns all the alleles in the allele distribution range observed (Figure 3.1). Later, selective phenotypic divergence and taxonomic reclassification occurs, and the owners of a subset of the alleles are reclassified into dromedary. In this case, Bactrian alleles will appear to 'avoid' areas occupied by dromedary alleles. This scenario can be interpreted to mean that the dromedary emerged ('descended') from the Bactrian rather than simultaneously as proposed previously. This hypothesis is consistent with theoretical expectations where the emergent taxon (dromedary) would show lower diversity when compared with the 'parental' taxa (wild and domestic Bactrian) as seen in this study.

3.4.3 Camel tree-based phylogeny remained unresolved

Phylogenetic inference using tree building methods showed all the three camel species to simultaneously diverge from a common ancestor, regardless of the tree and distance methods used. More recent studies suggested that the wild bactrian is a separate lineage and evolved independently from a wild progenitor separate from the domestic bactrian. This is still plausible considering the equivocal nature of their tree-based phylogeny.

3.4.4 Ascertainment bias in bovini is unreliable.

There was inconsistent change in allele size range for non-focal cattle species: For loci isolated from the *indicus* genome, the allele size ranges either showed an increase or a decrease across the four species, while there was either a fixed or a slight increase in the allele size range from *taurus* to *indicus* for loci isolated from the *Bos taurus* genome. This showed that ascertainment bias was weak in Bovidae. The inconsistency in allele size ranges, remaining fixed or increasing/ decreasing for loci isolated from the *taurus* or *indicus*, suggested that for Bovidae, ascertainment bias is unreliable. Perhaps the complex migration patterns and multiple centres of domestication confound the analysis of evolutionary relationships in this group.

3.5 CONCLUSION

Ascertainment bias and allelic patterns were tested for utility as phylogenetic signals in reconstructing relationships among camel species. Results showed that ascertainment bias is a useful probe for reconstructing evolutionary relationships among camels, and revealed a hypothesis not yet proposed in genetic analyses: that the dromedary might have descended from the bactrian. This hypothesis refutes the paleontological version of simultaneous divergence of the two from a common ancestor.

CHAPTER 4: RNA SECONDARY STRUCTURE

4.1 INTRODUCTION

Phylogenies reflect the evolutionary path of speciation, leading to current relationships among extant taxa. Knowledge of this historical relationship is important in both paleontology and archaeology, and in understanding evolution. Character states used in phylogenetic reconstruction include molecular data, biochemical data (such as blood group, tissue oil content, toxicity), quantitative, ecological (such as behavior, host commonality), morphological features (such as presence or absence of fins, number of legs, lengths of legs, etc) and fossil record. Of these, molecular data is the most commonly used today. However, there are a significant number of cases where molecular data has failed to resolve organismal phylogenies (e.g. Russo et al., 1996; Ochieng et al., 2007a; Parra-O et al., 2006) due to reasons that include homoplasy and paralogy. These cases require methods that deal with elements that are retained during evolution and can hence eliminate 'noise' characters. Such retained elements would include functional constraints in the RNA secondary structures. A detailed review on RNA secondary structures is given in Chapter 2.

Different techniques (explained in Chapter 2) can be used to predict RNA secondary structure, upon which a common secondary structure can be assigned to species with similar functions or closely identical structures. It thus appears that the integrity and hence shape of RNA secondary structures can be an indicator of evolutionary distances among taxa. This study sought to test the diagnostic value of RNA secondary structure

morphology as a phylogenetic signal. A suitable model for testing this was the phylogenetic relationship among species and genera in Bovidae, one of the most difficult groups to classify, and whose higher order classification remains unresolved. Different datasets (morphological, fossil data and DNA evidence) in Bovidae often conflict in their phylogenetic resolve (Miyamoto and Goodman, 1986; Kraus and Miyamoto, 1991; Gatesy et al., 1992; MacHugh, 1996), such as the morphological data that disagrees with DNA based evidence on the placement of Bison species (Groves, 1981; Geraads, 1992). I hypothesized that morphometric variation on RNA secondary structures among bovini species would be proportional to their evolutionary distances and thus would help resolve the enigmatic question of their evolutionary relationship. If such predictions are made based on non-functional regions of the genome, this can mislead because RNA structures may differ over timescales. To assure consistency to enable unbiased comparisons of secondary structures such as in studies of speciation, a gene whose function will ensure repeatability is preferred. I used Bovidae mitochondrial cytochrome *b* gene, whose sequences are available in a public repository (Genbank; NCBI). Cyto b contains redox centres involved in electron transfer (Hatefi, 1985), transferring electrons from one molecule to another in various pathways that form new molecules. Because of this critical transport role, this gene can be used to infer speciation and evolution as it is conserved and only evolutionarily significant variations are expected. Species targeted included both resolved, and those whose placement has often raised controversies, such as placement of *Bison*, the descendants of the *Auroch* (*Bos taurus*, *B. indicus*), and *Banteng* cattle.

4.2 MATERIALS AND METHODS

4.2.1 Source of data for analysis

This study utilized sequence data to test the utility of RNA secondary structures as a reliable phylogenetic signal. Bovine cytochrome *b* sequences from the following species were obtained from Genbank (NCBI) and assembled for analysis: *Bos taurus*- accession nos. HM045018, AF492351, GQ129208, AY526085, GQ129207, NC_006853; *Bos indicus*- AF492350, AY126697, NC_005971; *Bos javanicus*- NC_012706, FJ997262; *European bison*- HQ223450, NC_014044, HM045017; *American bison*- EU177871, NC_012346; *Bos primigenius*- NC_013996.

4.2.2 Prediction of RNA secondary structure and morphological analysis

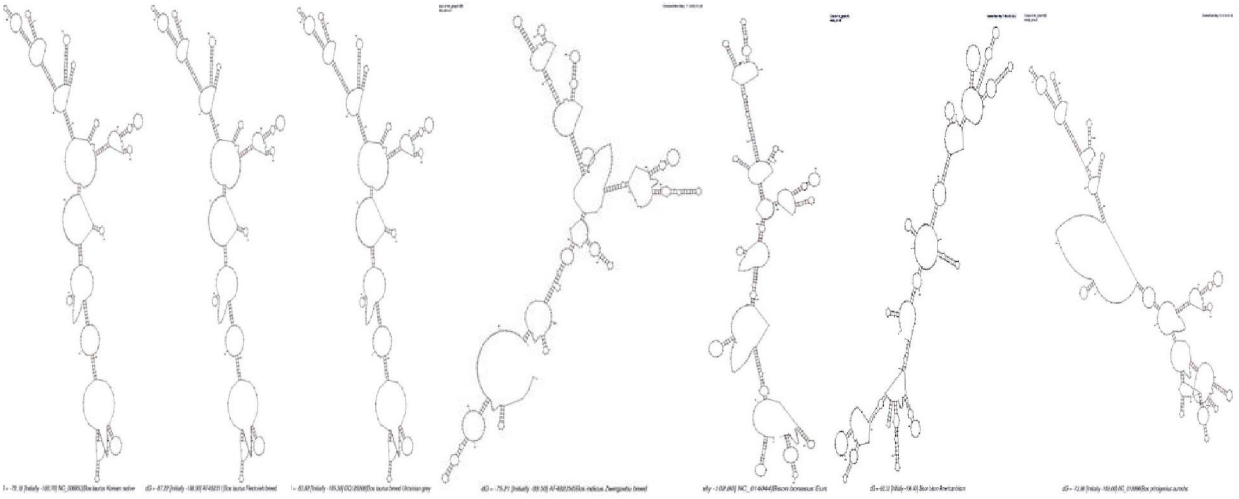
The data was formatted and imported into the BioEdit sequence alignment editor for Windows (Version 7.1.3.0). Aligned sequences were used to reconstruct a phylogeny using standard methods as described in Ochieng et al., (2007a) as control (see Appendix 1 for this tree). Secondary structures were predicted using MFOLD version 3.5, a web server for nucleic acid folding prediction (Zuker, 2003). This server is a harmonized version of closely related software applications available on the web for the prediction of the secondary structure of single stranded nucleic acids. Folding for each DNA sequence was conducted both at the default temperature (37^{o}C) and at 25^{o}C. MFOLD was also used to compute the CG content and Free Energy (thermodynamic stability) for each sequence. Figures representing the predicted RNA secondary structures were used to infer the degree of congruence between predicted structures in

different taxonomic groups. This was done by comparing the shapes and sizes of hairpins (in bases) and their positioning relative to helices. The lengths of helices were also compared. Number of paired bases was not compared as this required enormous physical labour, but would be a useful character in such analyses.

4.3 RESULTS

4.3.1 RNA secondary structure morphology distinguished species

Morphologically, the RNA secondary structure comparison was able to discern species, with a considerable difference between taurine cattle and an indicine. There was however a high-level consistency in structural morphology among the taurines irrespective of the country or region of origin (Figure 4.1). Further, differences between bovine species and others were apparent. Corresponding phylogeny based on DNA sequences appears in Appendix 1 and is consistent with RNA structures.



taurine (Korean) | taurine (Fleckvieh) | taurine (Ukranian) | indicine (*Bos indicus)* | *European bison* | *American bison* | *Bos primigenius*

Figure 4. 1 Morphological layout of predicted secondary structure for three different breeds of *Bos taurus* and a *Bos indicus* cattle: Korean native, taurine, Fleckvieh breed, taurine, Ukranian grey, taurine, *Bos indicus*, *European bison*, *American bison* and *Bos primigenius.* Legible versions appear in appendix 3. There is consistency in morphology in the taurine regardless of the country of origin while the morphology of the indicine is incongruent (Only the general morphology is shown for comparison). *B. taurus/indicus* divergence is approx 500,000 YA; *indicus/javanicus* is 2MYA; Bos/Bison is 4 MYA (MacHugh, 1996; Lenstra and Bradley, 1999)

4.3.2 RNA structure deviates from traditional phylogeny of banteng

RNA secondary structure showed that *B. taurus* and *B. indicus* are not the closest relatives. *Bos javanicus* (*banteng*) and *B. taurus* were the most similar, differing only in the region marked out in the rhomboid shape (Figure 4.2A) quite distinct from the *Bos indicus*. This observation defies the traditional phylogenetic clustering, which groups the *B. taurus* and *B. indicus* together, different from *B. javanicus* (Figure 4.2B).

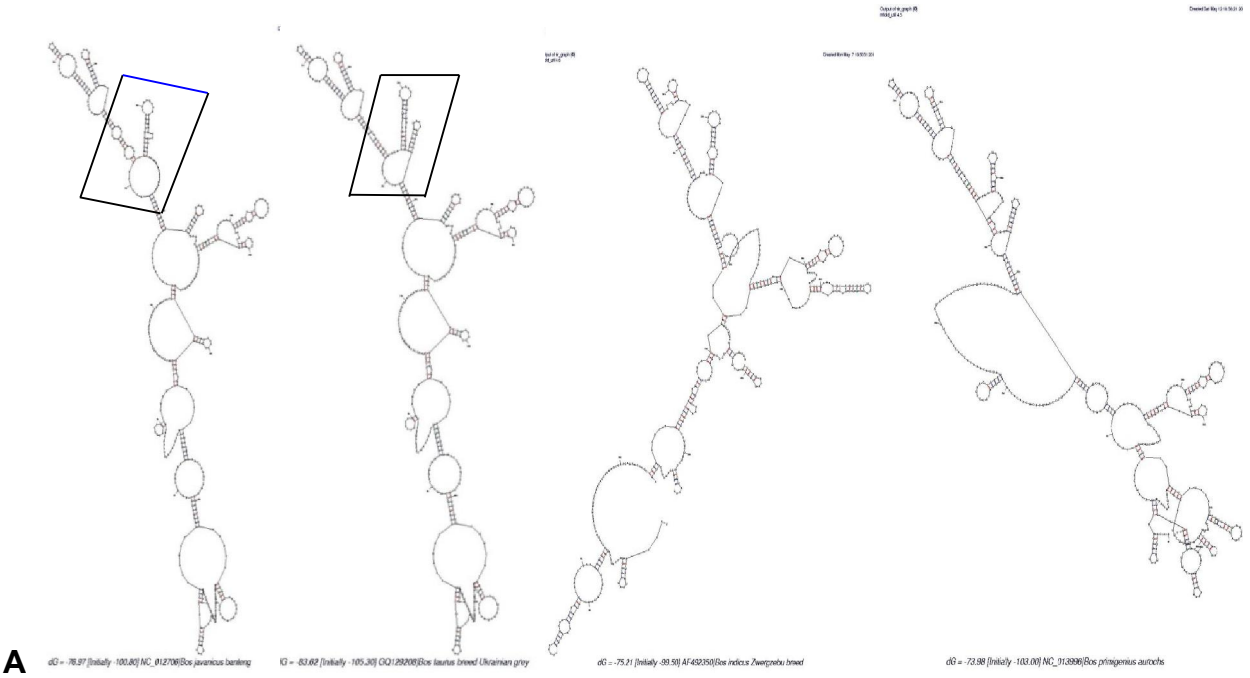


Figure 4. 2 Morphological layout of predicted secondary structures of three cattle species: *Bos javanicus*, *Bos taurus* and *Bos indicus* (A).The figure shows *B. javanicus* to be closer to the *B.taurus,* against current understanding that *B. taurus* and *B. indicus* are the closest relatives. *Bos javanicus* differs from the *taurus* only at the region marked out in rhomboid shapes, but the pair are generally different from the *B. indicus*. (B): Phylogenetic relationships as expected from traditional taxonomy (left) and relationship as reflected by RNA secondary structure in this study (right). *B. taurus/indicus* divergence is approx 500,000 YA; *indicus/javanicus* is 2MYA.

4.4 DISCUSSION

4.4.1 RNA structures show *Banteng* cattle to originate from the *Auroch*

Despite existing controversies regarding the taxonomy and phylogeny of bovidae (see for example (Hassanin and Douzery,1999) their relationship among *B. taurus , B. indicus* and *B. javanicus* (*banteng*) has not been part of this controversy. It has been universally accepted that *B. taurus* and *B. indicus* are the closest living relatives of each other, and that *B. javanicus* is distant from the two, based on both morphological and molecular datasets (MacHugh, 1996; Lenstra and Bradley, 1999). Indeed, it has been proposed that *B. taurus* and *B. indicus* are different breeds of the same species (Hiendleder et.al, 2008). Despite this clarity, the RNA secondary structure morphologies suggest otherwise: that *B. javanicus* is indeed closer to the taurine cattle, making the taurine-indicine clade paraphyletic. The two closest living relatives, *B. taurus* and *B. indicus* are known to have evolved from a most recent common ancestor *B. primigenius* only 500,000 years ago (MacHugh, 1996; Lenstra and Bradley, 1999) hence they are expected to be closer to each other than to any other species. However, it is still credible to believe the version presented by the current dataset. Given the phylogenetic consistency shown by the RNA secondary structure morphological comparisons, it is possible that banteng is one of the descendants of the *Auroch*. And based on the observed alliance, it would be *taurine* and this can be authenticated using the Y-chromosome specific marker and mtDNA haplotyping. Analysis of RNA structures of other gene loci may provide a more comprehensive picture.

4.4.2 Did *B. indicus* evolve from African auroch?

Phylogeny revealed in this study showing the known descendants of the *Auroch* (*B. taurus* and *B. indicus*) to be paraphyletic with respect to the wild *Banteng* (*Bos javanicus*) may be surprising based on previously accepted phylogeny and dating (*B. taurus/indicus* divergence is approx 500,000 YA, *indicus/javanicus* is 2 Million YA, Bos/Bison is 4 Million YA; MacHugh, 1996; Lenstra and Bradley, 1999). The most immediate way to explain this would be the variation in multiple structures predicted from the same sequence. Only one representative sequence was 'folded' for the analysis. However, it is still possible that the unusual relationship reflects the true phylogeny. This would call for a re-examination of origin and migration theories for domesticated cattle. Until now, it is believed that the ancestor of domestic cattle, the *Auroch*, lived both in North Africa (*Bos primigenius opisthonomous*) and Eurasia (*Bos primigenius namadicus* and *Bos primigenius primigenius*), became extinct everywhere else before finally going extinct in Europe, and that *Bos indicus* and *B. taurus* then moved into Africa from two different centres of domestication in Eurasia. However, the present study suggests that *B. indicus* may have evolved independently from *Bos primigenius opisthonomous* in North Africa, and hence would be distant from the *B. taurus* (which evolved from Eurasian *auroch*) and *B. javanicus*. Another plausible explanation for these results is that *B. taurus* has possibly retained ancestral polymorphisms, making it similar to *B. javanicus* which has been considered a sister to their common ancestor. Finally, it is also reasonable to believe that hybridization between banteng and taurine cattle could confound the true evolutionary history of these species.

4.5 CONCLUSION

This study has revealed the potential of RNA structures to resolve organismal phylogenies with a significant consistency within species while differing across species and genera. However, the test model used was a case of unresolved and quite often controversial phylogeny of the bovini. Despite the long held tenet, the data show the *Bos taurus-indicus* alliance to be paraphyletic with respect to *Bos javanicus* (banteng cattle). Despite there being several plausible explanations to this unusual outcome, it is highly likely that the indicine cattle evolved independently from the *Auroch* in northern Africa, or that banteng cattle evolved from the *Auroch* in Eurasia.

CHAPTER 5: GENERAL DISCUSSION

Phylogenetic reconstruction has become routine in many organismal studies in the medical and agricultural sciences. However, a common drawback in many of these studies is the failure of different datasets using the same individuals to recover the same phylogeny. Many explanations have been advanced to account for this discrepancy, chiefly the biological processes inherent in the evolutionary process itself such as homoplasy or evolutionary noise, not adequately treated in analysis programs available. It is against this background that this study was designed to evaluate the usefulness of two biological phenomena capable of resolving phylogenetic relationships among taxa: ascertainment bias and incongruence in RNA secondary structure morphology.

RNA secondary structures are useful in phylogenetic inference.

Morphological congruence of the predicted secondary structures compared in a manner reflecting evolutionary distances of major Bovine lineages. Whereas individuals within species were the most congruent followed by those between species within the genus, the most distant ones also differed the most. This outcome reinforces the usefulness of secondary structures in resolving enigmatic phylogenies. However, in both test systems used in this study (Camelini and Bovini), ascertainment bias did not exhibit the uniformity required to meet the threshold of a good phylogenetic probe. In many cases and for many loci, the principle (reduced allele size range proportional to evolutionary distance from the focal taxon from which they were developed) was not obeyed

especially in the Bovini. This confirms that ascertainment bias may reflect phylogenetic trends in some systems but not others.

Origins and migration theories of two major domestic animals

Apart from providing reliable alternative phylogenetic probes, this study has produced results that contradict two major evolutionary, migration and domestication theories. The data obtained in this study contradicts the following theories: (1) *Bos taurus* and *Bos indicus* descended from the *Auroch* in Eurasia and then the *B.indicus* migrated into Africa (2) *Bos javanicus* and the *Auroch* shared a common ancestor, hence indicine-taurine clade is monophyletic (3) One humped and two humped camels simultaneously radiated from their common ancestor (Paracamelus), then the dromedary moved into Africa from Asia. Currently, it is held that the dromedary, or its ancestors, separated from the Bactrian in western Asia and spread across Arabia (Clutton-Brock, 1999) and into North Africa. It is also held that the ancestor of domestic cattle, the auroch, lived in Eurasia (*Bos primigenius namadicus*) and *Bos primigenius primigenius*), and that *Bos indicus* and *B.taurus* then moved into Africa from two different centres of domestication in Eurasia.

The present study suggests that *B.indicus* may have evolved independently from *Bos primigenius opisthonomous* in North Africa, and *B.taurus* (which evolved from Eurasian auroch) and *B.javanicus* might share a most recent common ancestor (MRCA). Similarly, the data suggested that the dromedary (*Camelus dromedarius*) emerged more recently from the Bactrian, rather than radiating simultaneously from a common

ancestor (Paracamelus). It is interesting that the two theories concern a similar pattern and involve widespread domestic animal species. Both Camelidae and Bovidae species found in Africa (dromedary and zebu cattle) are postulated to have their origin in Asia, and are considered to have radiated simultaneously from a common ancestor with their sister species before moving into Africa. Coincidentally, the results of this study refute the sister species hypothesis in both cases.

Despite the possibility of multiple structures being predicted from the same sequence, the gene used in this study (Cytochrome b) is functionally constrained from variations that have no adaptive significance, and often causes lethal diseases. This suggests that RNA secondary structures predicted for this gene will be mostly conservative and reliable. Whole genome analysis could provide a more comprehensive picture on the history of these taxa, however, current versions of programs used in this study such as MFOLD, cannot fold whole genomes due to the large genomic size. Further, RNA secondary structures can only be predicted for transcribed genomic regions.

RECOMMENDATION

Following careful consideration of the research problem that necessitated this work, the solution pathways taken, results obtained and interpretations therein, I make the following recommendations:

- RNA secondary structure morphological congruence based on Cytochrome b gene in mammals is a reliable tool for inferring phylogenies, especially in cases where evolutionary noise is a likely nuisance, and may be way better than DNA sequences.
- Ascertainment bias should not be relied on as a universal phylogenetic signal. However, it may reflect evolutionary distances in some systems but not others.
- More data should be accumulated to revisit two independent hypotheses, one that the dromedary camel found in Africa have their origin in Asia, having simultaneously radiated from their common ancestor with Bactrian, and the other that zebu cattle (*Bos indicus*) entered Africa from Eurasia, having evolved from the Auroch.
    - This study suggests that dromedary is a recent offshoot from Bactrian
    - The study showed zebu cattle to evolve independently from African Auroch.
- With recent advances in bioinformatics, it should be possible to resolve enigmatic phylogenies from whole genome analyses

REFERENCES

Chambers, J.K., and McDonald, L.E. (2000). A G protein-coupled receptor for UDP-glucose. Journal of Biological Chemistry **275,** 10767-10771.

Ellegren, H., Primmer, C.R., and Sheldon, B.C. (1995). Microsatellite evolution: Directionality or bias in locus selection. Nature Genetics **11,** 360-362.

Gatesy, J., Yelon, D., DeSalle, R., and Vrba, E.S. (1992). Phylogeny of the Bovidae (Artiodactyla, Mammalia), based on Mitochondrial Ribosomal DNA sequences. Molecular Biology and Evolution **9,** 433-446.

Geraads, D. (1992). Phylogenetic analysis of the tribe Bovini (Mammalia: Artiodactyla). Zoological Journal of the Linnean Society **104,** 193-207.

Clutton-Brock, J. (1999). A Natural History of Domesticated Mammals (2nd edn, Cambridge University Press, Cambridge), 238 pp

Goldstein, D.B., Ruiz, L.A., Cavalli-Sforza, L.L., and Feldman, M.W. (1995). An evaluation of genetic distances for use with microsatellite loci. Genetics **139,** 463-471.

Groves, C.P. (1981). Systematic relationships in the Bovini (Artiodactyla, Bovidae). Journal of Zoological Systematics and Evolutionary Research **19,** 264-278.

Gu, X. (2001). Maximum-likelihood approach for gene family evolution under functional divergence. Molecular Biology and Evolution **18,** 453-464.

Hall, B.G. (2004). Predicting the evolution of antibiotic resistance genes. Nature reviews Microbiology **2,** 430-435.

Hanotte, O., Bradley, D.G., Ochieng, J.W., Verjee, Y., Hill, E.W., and Rege, J.E.O. (2002). African Pastoralism: Genetic Imprints of Origins and Migrations. Science **296,** 336-339.

Hassanin, A., and Douzery, E.J.P. (1999). The tribal radiation of the family Bovidae (Artiodactyla) and the evolution of the mitochondrial cytochrome b gene. Molecular Phylogenetics and Evolution **13,** 227-243.

Hatefi, Y. (1985). The mitochondrial electron transport and oxidative phosphorylation system. Annual Review of Biochemistry **54,** 1015-1069.

Haveson, Y. (1954). Tertiary camels of the Eastern Hemisphere (genus *paracamelus*). Proceedings of Paleontological Institute **67,** 100-161.

Hiendleder, S., Lewalski, H., and Janke, A. (2008). Complete mitochondrial genomes of *Bos taurus* and *Bos indicus* provide new insights into intraspecies variation, taxonomy and domestication. Cytogenetic and genome research **120,** 150-156.

Ji, R., Cui, P., Ding, F., Geng, J., Gao, H., Zhang, H., Yu, J., Hu, S., and Meng, H. (2009). Monophyletic origin of domestic bactrian camel (*Camelus bactrianus*) and its evolutionary relationship with the extant wild camel (*Camelus bactrianus ferus*). Animal Genetics **40,** 377-382.

Jianlin, H., Mburu, D., Ochieng, J.W., Rege, J.E.O., and Hanotte, O. (2000). Usefulness of New World Camelidae microsatellite primers for amplification of polymorphic loci in Old World Camelids. Animal Genetics **31,** 404-406.

Kraus, F., and Miyamoto, M.M. (1991). Rapid cladogenesis among the pecoran ruminants: evidence from mitochondrial DNA sequences. Systematic Zoology **40,** 117-130.

Lenstra, J.A., and Bradley, D.G. (1999). Systematics and Phylogeny of cattle. In The Genetics of Cattle, R. Fries, Ruvinsky, A., ed (CABI Publishing), pp. 1-14.

MacHugh, D.E. (1996). Molecular Biogeography and Genetic Structure of Domesticated Cattle. In Department of Genetics (Dublin: University of Dublin), pp. 1-29.

Martin, J.-L., and Pitocchelli, J. (1991). Relation of within-population phenotypic variation with sex, season, and geography in the blue tit. Auk **108,** 833-841.

Mburu, D.N., Ochieng, J.W., Kuria, S.G., Jianlin, H., Kaufmann, B., Rege, J.E.O., and Hanotte, O. (2003). Genetic Diversity and relationships of Indigenous Kenyan camel (*Camelus dromedarius*) populations: Implications for their classification. Animal Genetics **34,** 26-32.

Miyamoto, M.M., and Goodman, M. (1986). Biomolecular Systematics of eutherian mammals: phylogenetic patterns and classification. Systematic Biology **35,** 230-240.

Nei, M. (1972). Genetic distance between populations. American Naturalist **106,** 283-292.

Ochieng, J.W. (2002). Genetic characterization of Kenyan and Chinese camels using microsatellites and mitochondrial DNA polymorphisms. Master of Science thesis, Southern Cross University, NSW, Australia

Ochieng, J.W., Henry, R.J., Baverstock, P.R., Steane, D.A., and Shepherd, M. (2007a). Nuclear ribosomal pseudogenes resolve a corroborated monophyly of the eucalypt genus Corymbia despite misleading hypotheses at functional ITS paralogs. Molecular Phylogenetics and Evolution **44,** 752-764.

Ochieng, J.W., Steane, D.A., Ladiges, P.Y., Baverstock, P.R., Henry, R.J., and Shepherd, M. (2007b). Microsatellites retain phylogenetic signals across genera in eucalypts (Myrtaceae). Genetics and Molecular Biology **30,** 1125-1134.

Parra-O, C., Bayly, M., Udovicic, F., and Ladiges, P.Y. (2006). ETS sequences support the monophyly of the eucalypt genus Corymbia (Myrtaceae). Taxon **55,** 653-663.

Penedo, M.C.T., Fowler, M.E., Bowling, A.T., Anderson, D.L., and Gordon, L. (1988). Genetic variation in the blood of llamas, *Llama glama* and alpacas, *Llama pacos*. Animal Genetics **19,** 267-276.

Peters, J., and Von-Driesch, A. (1997). The two-humped camel (*Camel bacrianus*): new light on its distribution, management and medical treatment in the past. Journal of Zoology **242,** 651-679.

Pigliucci, M., Murren, C.J., and Schlichting, C.D. (2006). Phenotypic plasticity and evolution by genetic assimilation. Journal of Experimental Biology **209,** 2362-2367.

Rubinsztein, D.C., Amos, W., Leggo, J., Goodburn, S., Jain, S., Li, S.-H., Margolis, R.L., Ross, C.A., and Ferguson-Smith, M.A. (1995). Microsatellite evolution - Evidence for directionality and variation in rate between species. Nature Genetics **10,** 337-343.

Russo, C.A.M., Takezaki, N., and Nei, M. (1996). Efficiencies of Different Genes and Different Tree-building Methods in Recovering a Known Vertebrate Phylogeny. Molecular Biology and Evolution **13,** 525-536.

Stanley, H.F., Kadwell, M., and Wheeler, J.C. (1994). Molecular Evolution of the family Camelidae: a mitochondrial DNA study. Proceedings of the Royal Society of London B **256,** 1-6.

Vidal-Rioja, L., Semorile, L., Bianchi, N.O., and Padron, J. (1987). DNA composition in South American camelids I characterization and *in situ* hybridization of satellite DNA fractions. Genetica **72,** 137-146.

Wiens, J.J., Kuczynski, C.A., and Stephens, P.R. (2010). Discordant mitochondrial and nuclear gene phylogenies in emydid turtles: implications for speciation and conservation. Biological Journal of the Linnean Society **99,** 445-461.

Wilson, R.T. (1984). The Camel. In Longman Group Limited (London), pp. 223.

Wu, M., and Tinoco, I. (1998). RNA folding causes secondary structure rearrangement. Proceedings of the National Academy of Sciences of the United States of America **95,** 11555-11560.

Young, C.C. (1932). On the artiodactyla from the *Sinanthropus* site at *Chouk'outien* (Beijing: Paleontologia Sinica), pp. 1-159.

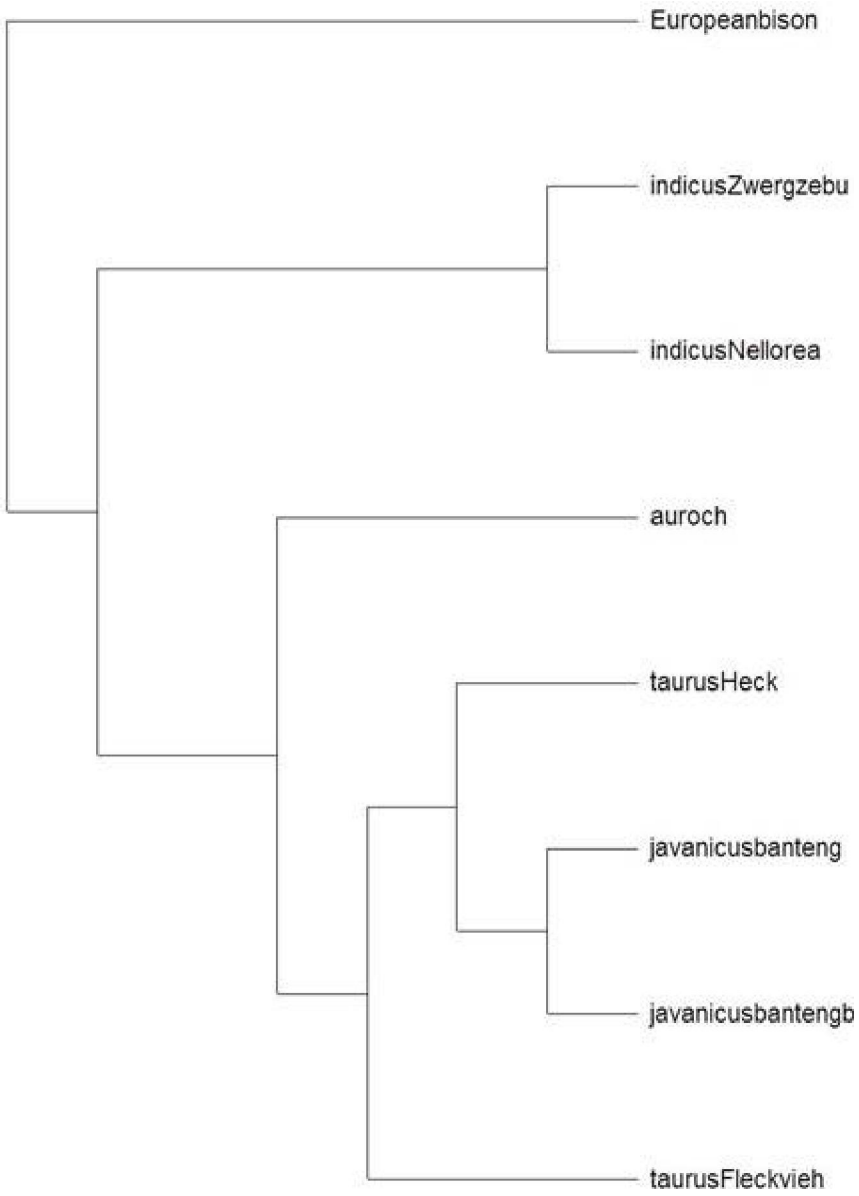
Zhu, Y., Queller, D.C., and Strassmann, J.E. (2000). A phylogenetic perspective on sequence evolution in microsatellite loci. Journal of Molecular Evolution **50,** 324-338.

Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Research **31,** 3406-3415.

Zuker, M., and Sankoff, D. (1984). RNA secondary structures and their prediction. Bulletin of Mathematical Biology **46,** 591-621.

APPENDICES

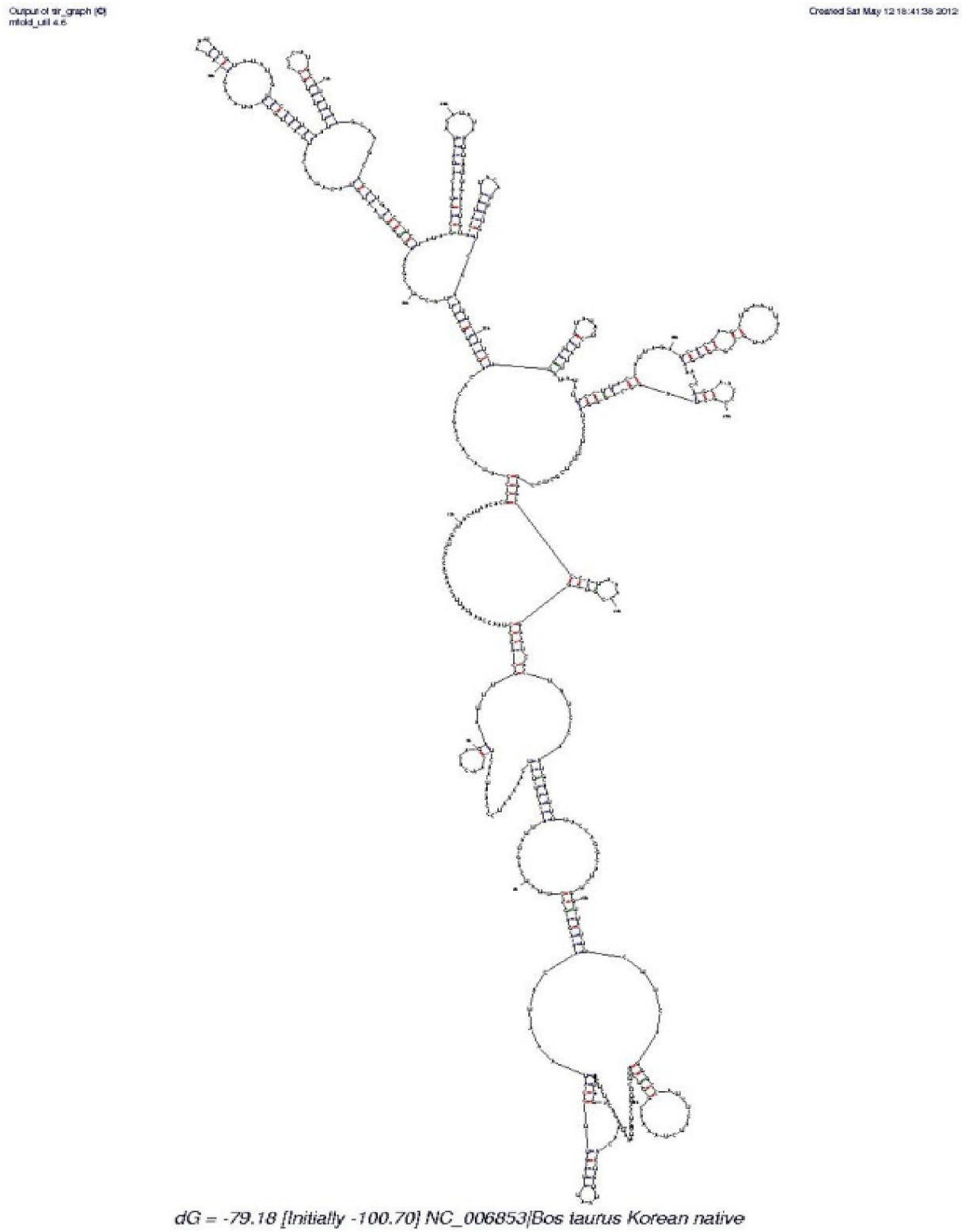
Appendix 1: Phylogeny of five Bovine species reconstructed from DNA sequences of mitochondrial Cytochrome *b* gene.



Appendix 2: Sequence alignment for the mitochondrial Cytochrome *b* gene used in reconstructing DNA-based phylogeny of five Bovine species. Only variable regions are shown (225 bases).

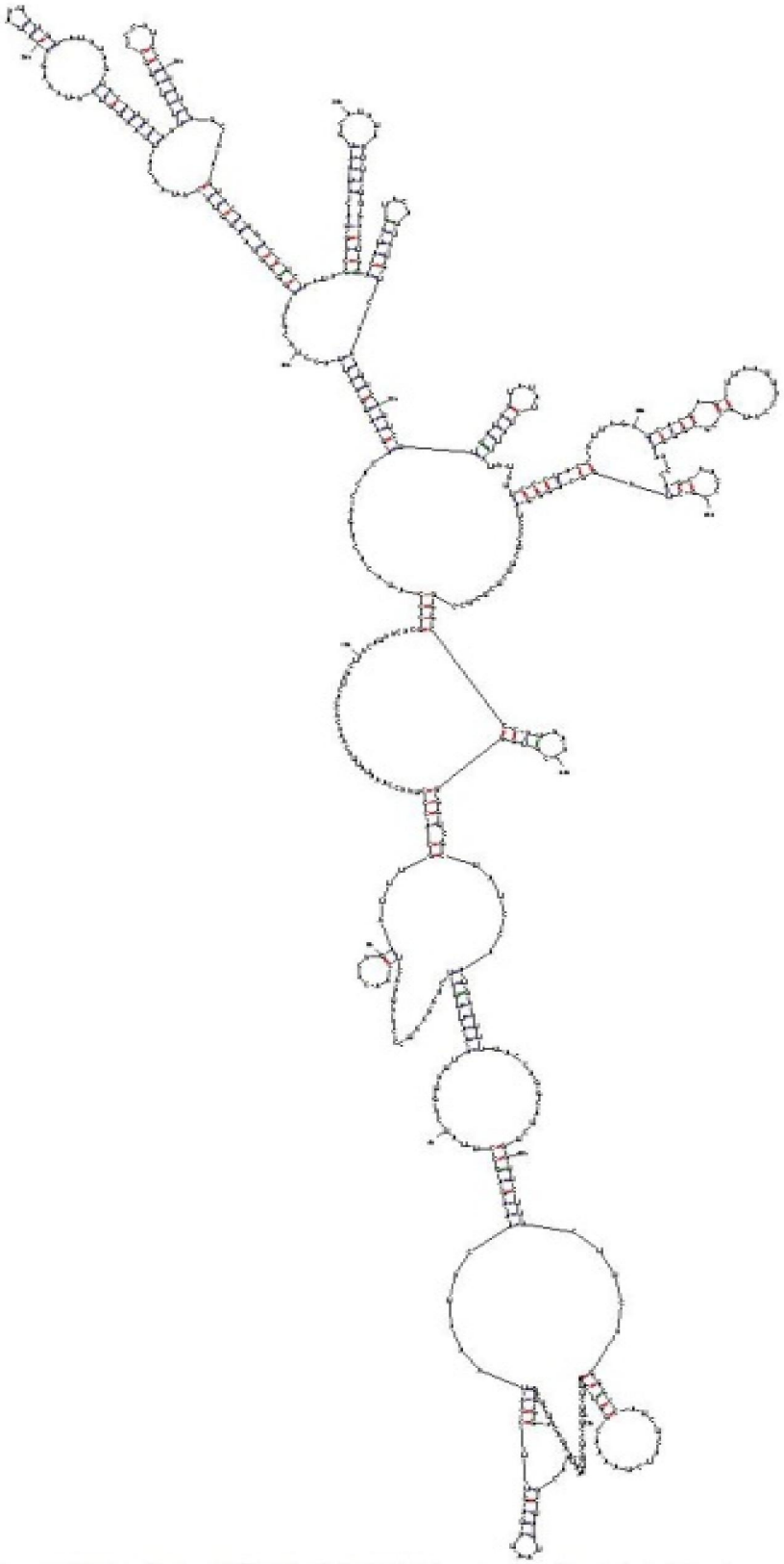
```
                    ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                             10         20         30         40         50         60         70         80
B.indicus Zwergz    GCCCGATTGT CGATCTCGAA TTTCGTCAGC TACGTCCCGA ATCGCTCCAT CGCAAGTAAT CGCATTACTT AGTTTATTCA
B.indicus Nellor    GTCCGATTGT CGATCTCGAG TCTCGTCAGC TACGCCCCGA ATCGCTCCAT CGCAAGTAAT AGCATTACTT AGTTTATTCA
B.taurus Heck       ACTTAACCGC TATCTTTGGA TTCTACTGAT CGTACTTTAG GCTATCTTGT CATGAACGGC CGCGCCGTCC AACATGCCTG
B.taurus Fleckvi    ACTTAACCGC TATCTTTGGA TTCTACTGAT CGCACTTTAG GCTATCTTGT CATGAACGGC CGCGCCGTCC AACATGCCTG
B.primigenius Au    ACTTAGCCAC TATCTTTAAA CTCTACTGAT CGCACTTCAG ACTATCTTGC TATGGACGGC CACGTCGCCT GACACGCCTA
B.javanicus 1       ACTTAACCGC TATCTCTGGA TTCTACTGAT CGCACTTTAG GCTATCTTGT CATGAACGGC CGTGCCGTCC AACATGCCTG
B.javanicus 2       ACTTAACCGC TATCTCTGGA TTCTACTGAT CGCACTTTAG GCTATCTTGT CATGAACGGC CGTGCCGTCC AACATGCCTG

                    ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                             90        100        110        120        130        140        150        160
B.indicus Zwergz    AGTTTATCGA TACCCGAACC CATGGCCCGC GGGCTCAGAG TACGCCTATC ATAGTGCGTG ATTGTCCCGT GCTCCCTTGG
B.indicus Nellor    AGTTTATCGA TACCTAAACC TATGGCCCGC AGGCTCAGGA TACGCCTACC ATAGTGCGTG ATTGTCCCGT GCTCCCTTGG
B.taurus Heck       GACCCACTAG CGTTCAAGTT CGTAACATAT GAATCTGAAG CGTAATCGTC GCAACATACC ACCACTCTAT ATCTTTCCGA
B.taurus Fleckvi    GACCCACTAG CGTTCAAGTT CGTAACATAT GAATCTGAAG CGTAATCGTC GCAACATACC ACCACTCTAT ATTTTTCCGA
B.primigenius Au    GACCCACTGG CGTTCAGGTT CGCAATCTAT GAATCTGAAG CGTAATCGTT GCAACATACG GCCACTTTAC ATCCCCTCAA
B.javanicus 1       GACCCGCTAG CGTTCAAGTT CGTAACATAT GAATCTGAAG CGTAATCGTC GCGACATACC ACCACTCTAT ATCTTTCCGA
B.javanicus 2       GACCCGCTAG CGTTCAAGTT CGTAACATAT GAATCTGAAG CGTAATCGTC GCGACATACC ACCACTCTAT ATCTTTCCGA

                    ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|
                             170        180        190        200        210        220
B.indicus Zwergz    ATCCACTTGA CCTTATTGCT ATCTGACACA CATCAACCGA ATGGTCGCTT CGTATACTAA GTTTG
B.indicus Nellor    ATCCACTTGA CATTACTGCT ATCCGACACA CATCAACCAA ATGGTCGCTT CGTATATTAA GTTTG
B.taurus Heck       GCTTGTACAT TCCCATCATT GCTTAGTATC TGCTGGTTAG GCGACTGTCC TACGCGCCGG ACTCA
B.taurus Fleckvi    GCTTGTACAT TCCCATCATT GCTTAGTATC TGCTGGTTAG GCGACTGTCC TACGCACCGG ACTCA
B.primigenius Au    GTTCGTACGT TCCCATCATT GCTTAGTGTC TACTTGTTAG GCAACTATCT TACGCGCCGG GCCCA
B.javanicus 1       GCTTGTACAT TCCCGTCATC GCTTAGTATC TGCTGGTTAG GCGACTGTCC TACGCGCCGG ACTCA
B.javanicus 2       GCTTGTACAT TCCCGTCATC GCTTAGTATC TGCTGGTTAG GCGACTGTCC TACGCGCCGG ACTCA
```

Appendix 3: Morphological layout of predicted RNA secondary sructures from eight bovine species.



Output of str_graph (0)
mfold_util 4.6

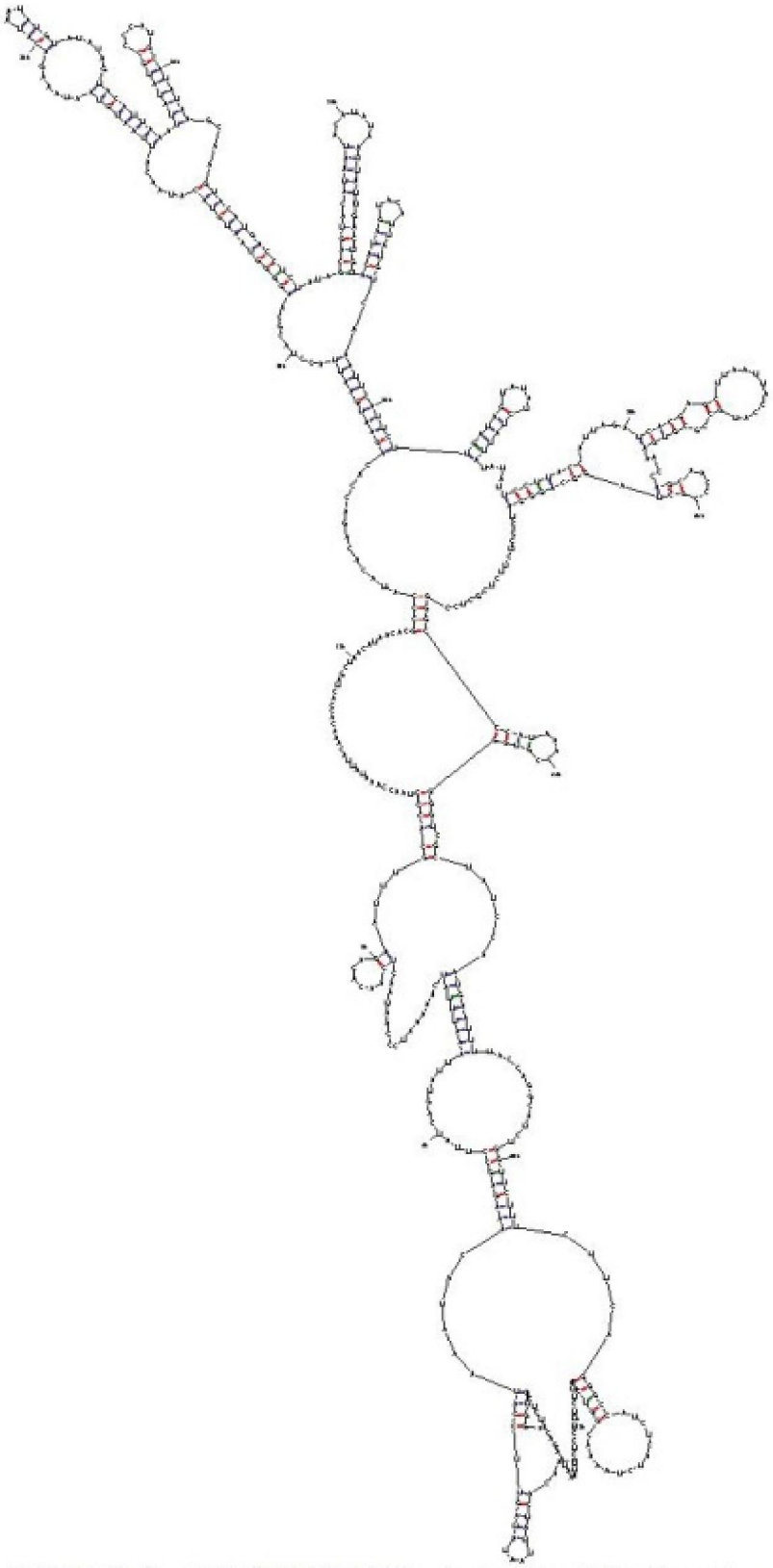
Created Sat May 12 16:51:43 2012



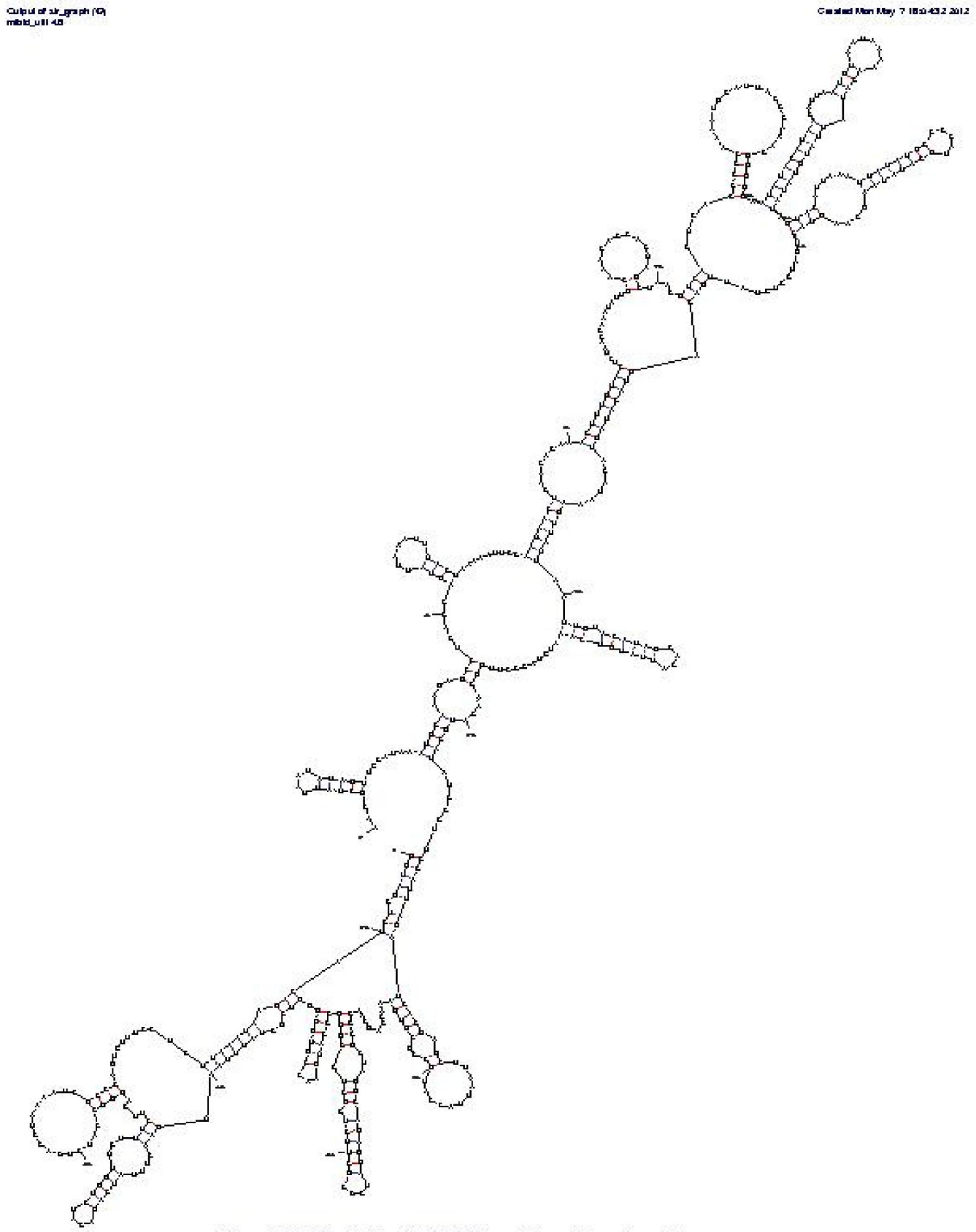
dG = -87.22 [Initially -108.90] AF492351|Bos taurus Fleckvieh breed

Output of sir_graph (0)
mfold_util 4.6

Created Sat May 12 16:53:50 2012



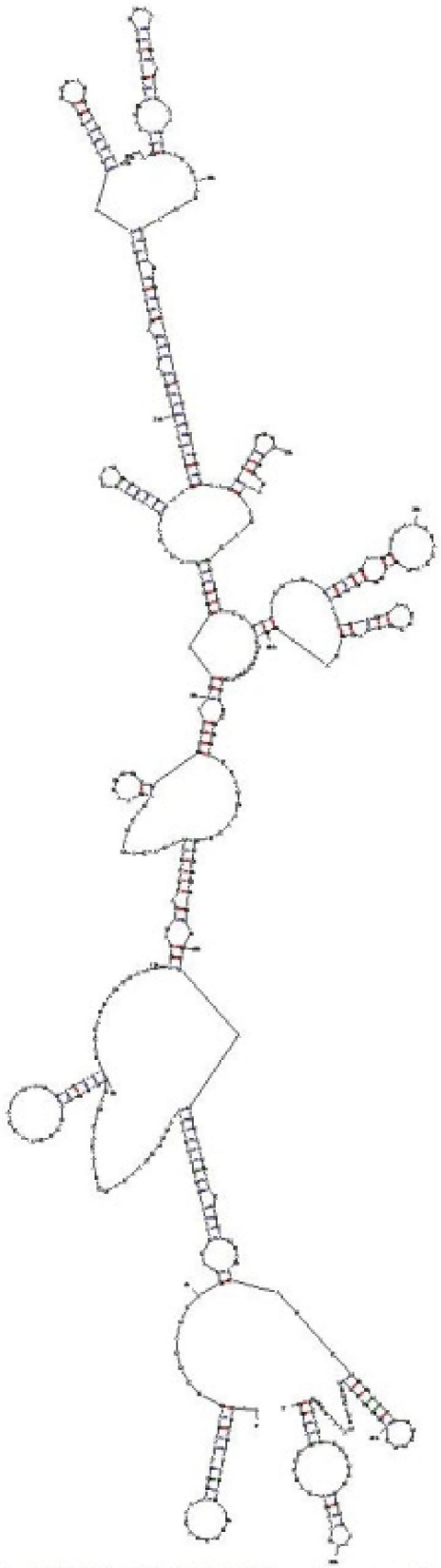
dG = -83.62 [Initially -105.30] GQ129208|Bos taurus breed Ukrainian grey



dG = -92.32 [Initially -106.40] Bison bison American bison

Output of ss_graph (C)
mfold_util 4.6

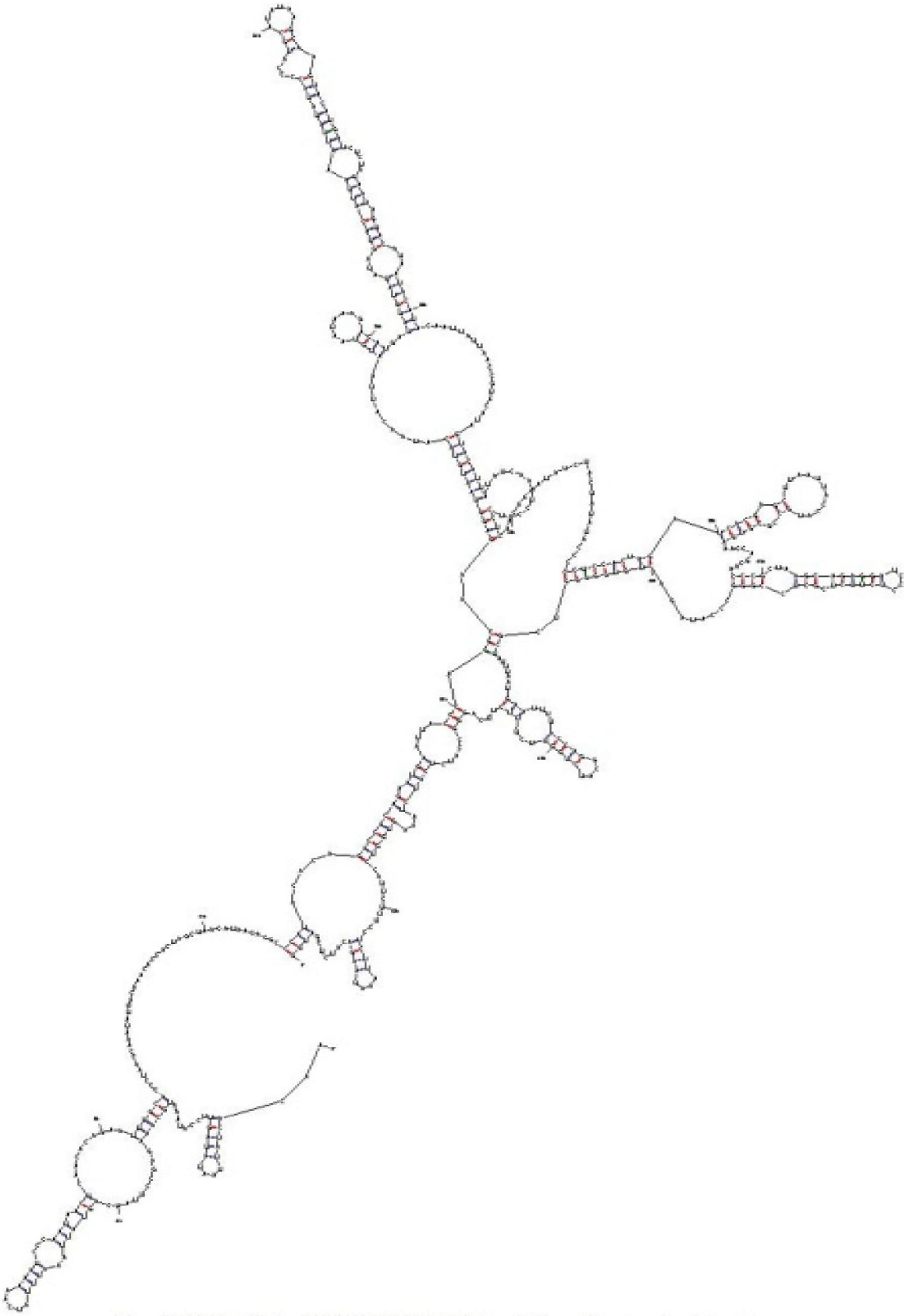
Created Mon May 7 19:26:33 2012



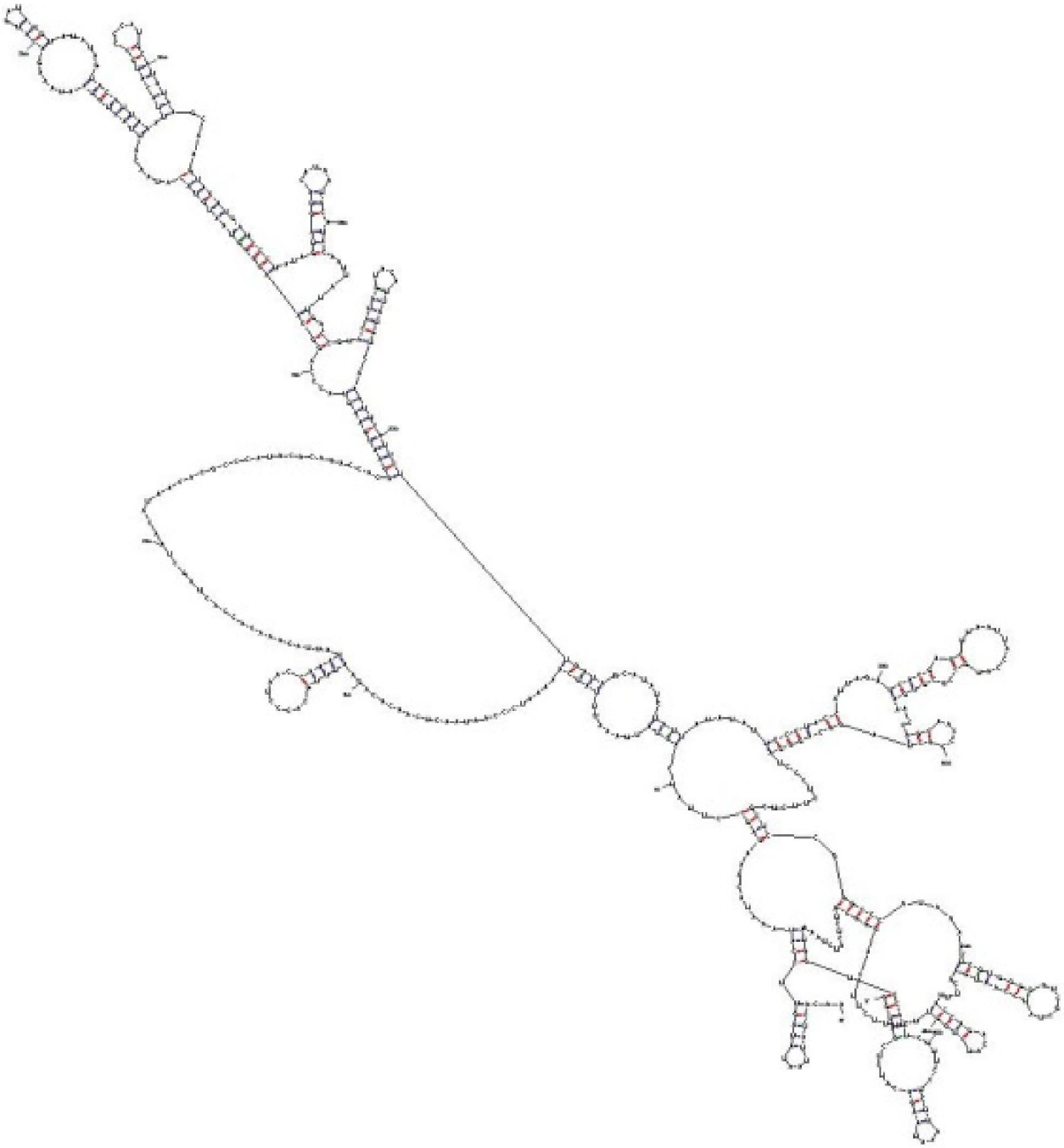
dG = -83.21 [Initially -102.80] NC_014044|Bison bonasus European bison

Output of ss_graph (ß)
mfold_util 4.6

Created Mon May 7 19:50:39 2012



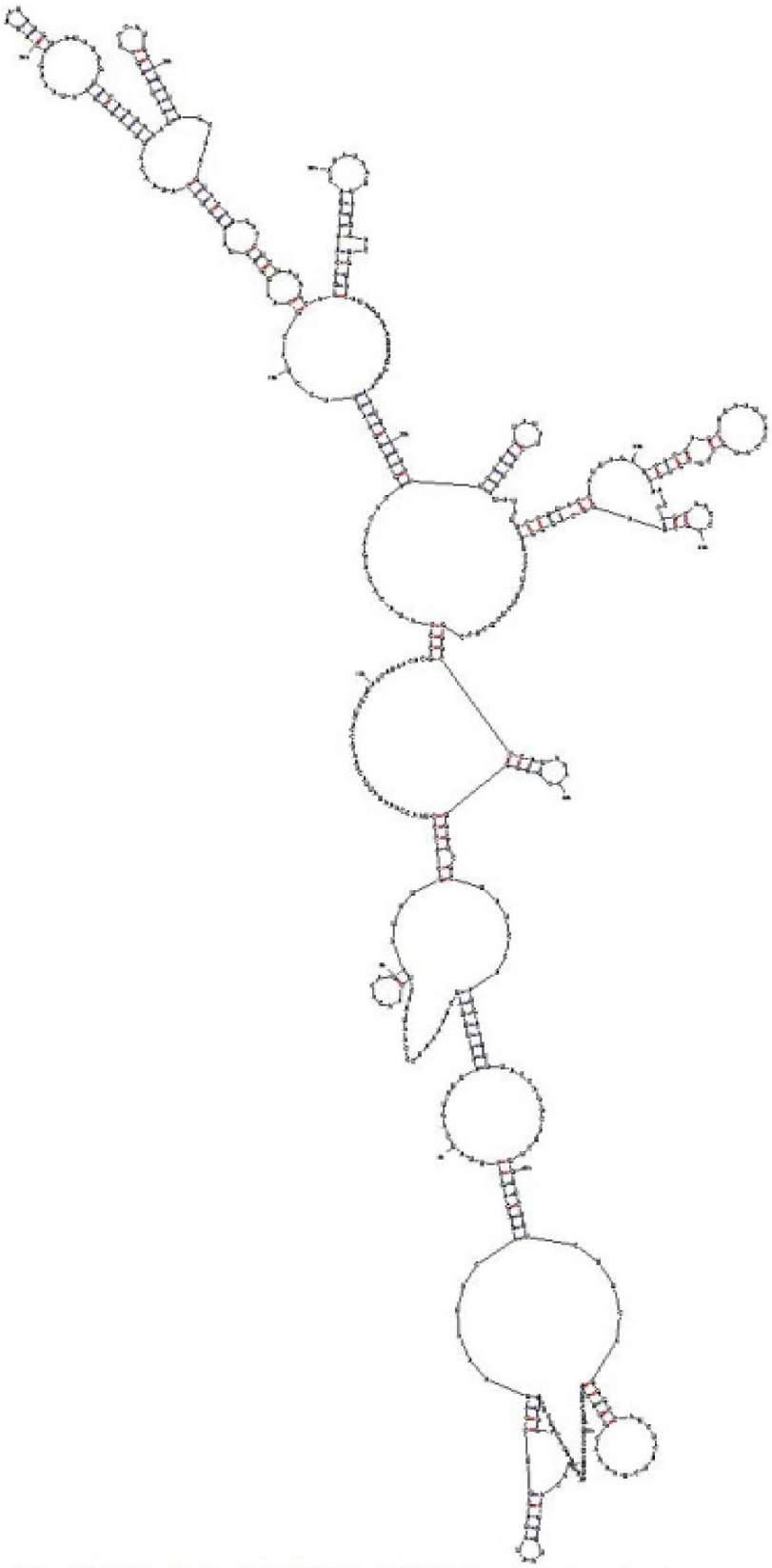
dG = -74.68 [Initially -94.90] AF492350|Bos indicus Zwergzebu breed



dG = -73.98 [Initially -103.00] NC_013996|Bos primigenius aurochs

Output of fir_graph (6)
mfold_util 4.6

Created Sat May 12 18:46:42 2012



dG = -76.97 [Initially -100.80] NC_012706|Bos javanicus banteng