Coalescent vs. Time-forward Simulations in the Problem of the Detection of Past Population Expansion

Krzysztof A. Cyran, Dariusz Myszor

***Abstract*—** The objective of this article is to show advantages and disadvantages of time-forward simulations as compared to the simulations performed backwards in time. The posed general issue is then narrowed to the problem of detection of past population expansion. The detection of population expansion is relevant in population genetics studies and there are plenty of methods used for that purpose. One of them utilizes genetic information preserved in microsatellites present in great abundance in the genome. We address the problem by simulating microsatellites evolution in different population growth scenarios. Namely we use time-forward computer simulation methods and compare results with results obtained by other researchers whiu used coalescent methodology. We argue that feed-forward simulation which are based on real life scenarios when succeeding generations are picked from the preceding one are becoming more and more suitable tool for population genetics with the increasing computational capabilities of contemporary computerss.

***Keywords*—**time-forward computer simulations, coalescent methods, population growth detection tests, short tandem repeat motifs/

I. Introduction

COALESCENT theory enables creating huge amounts of samples in quite a short time [1], yet its methods were developed some years ago when computers were rather expensive and possessed relatively low computational power. Over the last years the situation has changed due to invention of multi-core processors and overall progress in technology, which makes contemporary hardware highly efficient in computations and available at reasonable price. What is more, some recent research shows that given circumstances, coalescent methods might return different results than time-forward simulation approach.

In both coalescent-based and time-forward simulation methods we often want to gain sample from population with experienced changes in amount of individuals between generations.

We are simulating changes of some genetic markers caused by mutation process and parenthood. Most popular genetic markers are microsatellites – short strains of DNA build from repeating motifs of length 2-6 nucleotides [2]. Length of microsatellite is denoted by amount of such repeated motifs, usually 60 or so [3].

Common mutation in microsatellites are changes in amount of repeated motifs [4] (change in length of microsatellite), usually we use one-step symmetric stepwise mutation model SSMM (microsatellite might change length by one, we assume that the probability of addition and deletion of one repeating motif is equal) [5].

Microsatellites became popular between researchers because of relative high mutation rate (about 10^{-4} – 10^{-5}), and the fact that they are spread all over genome [6] (in human genome there is over 10 000 known microsatellites [7]), most of them is in non coding DNA, so most of microsatellites probably don't have influence on reproductive capabilities of individuals. Microsatellites are easy in mathematical analysis.

During research work we created series of population that underwent different kind and magnitude of growth. To simulate develop of population we need model providing dynamic description of the evolution. We choose Wright – Fisher model, it is based on idealized population. Classical version of Wright – Fisher model assumes [8]:

- discrete and non overlapping generations,
- haploid individuals in populations,
- constancy of population size,
- equilibrium fitness of individuals in the population,
- lack of geographical or social structure in the population,
- no recombination in the population.

Because we simulated populations which size was changing in time, we modified model so it allowed for changes in

The scientific work financed by Ministry of Science and Higher Education in Poland from funds for supporting science in 2008-2010, as a research project number N N519 319035

K. A. Cyran is with the Institute of Informatics, Silesian University of Technology, Gliwice, 44-100 Poland, phone: +48-32-237-2500; fax: +48-32--237-2733; e-mail: krzysztof.cyran@ polsl.pl.

D. Myszor is with the Institute of Informatics, Silesian University of Technology, Gliwice, 44-100 Poland, phone: +48-32-237-2500; fax: +48-32--237-2733; e-mail: dariusz.myszor@ polsl.pl.

population size. We assume that all experiments were correct for Y [9] chromosome or mtDNA [10] in order to bypass recombination issues and provide haploid individuals.

When new generation was created the old one was deleted so there were no overlapping generations. During creation of new individual all parents could be chosen with equal probability, so we eliminate the problems of individuals' fitness and geographical or social structure.

II. Simulations Methodology

In general there are two ways of population history simulation. For that purpose we can use coalescent theory or time forward simulation

A. Time-forward Simulations

In time forward simulation we create succeeding generation based on previous one. Every individual in previous generation might have influence on current generation. Usually we don't have genetic samples composed of all members of a population, but just a few individuals (e.g. 40 as in [11]). Amount of analyzed individuals can make difference in outcomes (Fig. 1).

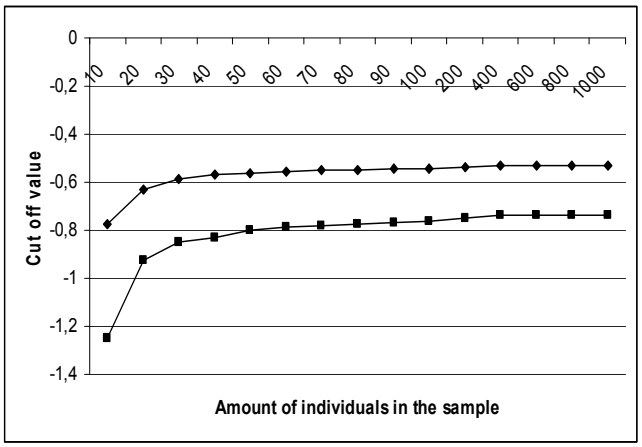


Fig. 1. Amount of individuals in the sample might have influence on coefficients values. Cut off values of $\ln\hat{\beta}_1$ ♦ and $\ln\hat{\beta}_2$ ■, were determined by the 0.05 quantile of the empirical distributions. We created 100 unlinked histories with constant amount of individuals $N = 20\ 000$. Each individual had 30 microsatellites, We took 100 samples from every generation that number were divisible by 100 000. For each history we simulated 1 000 000 generation. Mutation rate was $v = 5 \times 10^{-4}$.

For simulation we used designed by us software called GenSim. The screenshot presenting form used for simulation is presented in Fig. 2. The software was written in *C#* programming language in .NET framework. As a develompemnt tool the Microsoft Visual Studio 2005 was used. In our experiments we tried the same algorithm as a random number generator which was formerly used by the first author for simulating branching processes in the problem of Mitochondrial Eve dating [12]. However, fianally mainly due to implementation reasons we used so called Mersenne Twister generator. The training sets obtained from siulations were used by one layer and two layers perceptrons with the purpose of proposing new artificial neural network based test [13]. Perceptrons were utilized because these networks are universal, contrary to probabilistic neural networks which learn much faster but are dedicated primarily for classification [14]. Also, due to their fast learning probabilistic neural networks can be applied a criterion in optimization of feature space [15], but are not a good choice in approximation problems considered in our study.

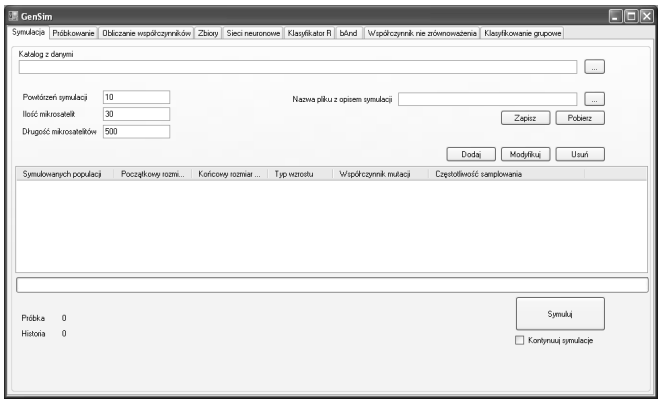


Fig. 2. Simulation form of the GenSim software.

To make our experiments more real, we took samples of n individuals from a populations and each sample contains fewer members than the whole population. Algorithm of time forward simulation consists of the following steps:

- Prepare initial population of N individuals. All individuals have the same amount of unlinked microsatellites, initialize each microsatellite with the same value. If we don't want to simulate vanishing microsatellites then initial size should be properly high.
- Run simulation for $2N$ to $4N$ iterations in order to reach mutation drift equilibrium [16] and gain real sample (see Fig 5).
- During each iteration create another generation of p individuals (p value is determined by changing of population size).
- For each member of new generation draw parent in the previous generation, take microsatellites from the parent, and for each, apply mutations according to SSMM model (one parent can have many children).
- Create as many generations as needed.

We can save individuals of every generation or just some

generations chosen, so there is a possibility to track changes of population's statistical properties such like those presented in the Fig 4 (for the definition of coefficients used in the Fig. 4, see Section 3).

B. Simulations of Coalescent

It is based on observation that not all individuals in a population will have children. For constant population size the probability of not having a child is about 37% [8]. Coalescent doesn't include those individuals without children into further consideration – genealogy of only those individuals which are in the sample is created (Fig. 3).

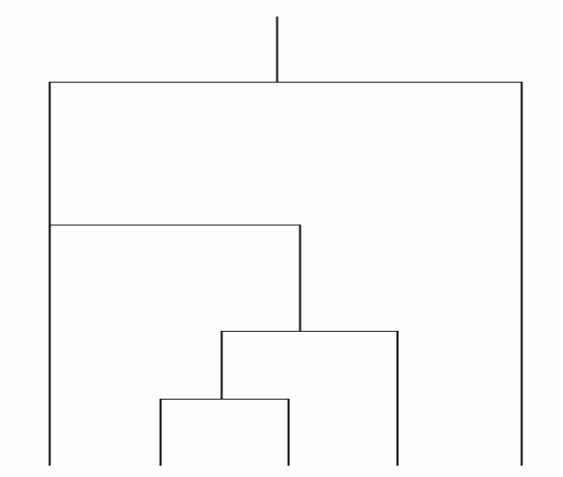
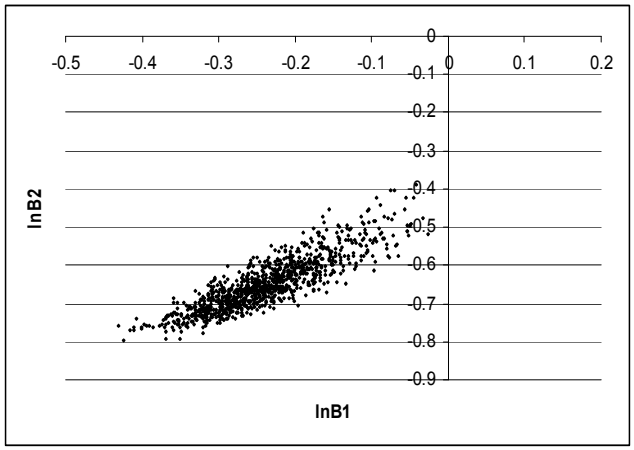


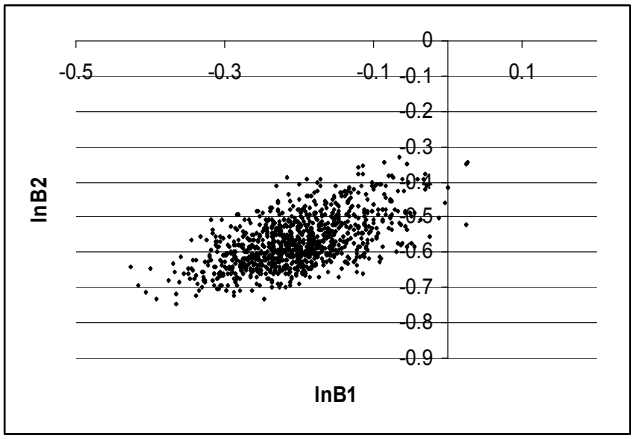
Fig. 3. The coalescent tree of the sample of five individuals.

Steps to create population using coalescent theory, assuming that for each microsatellite we create distinct genealogy, include: [8]

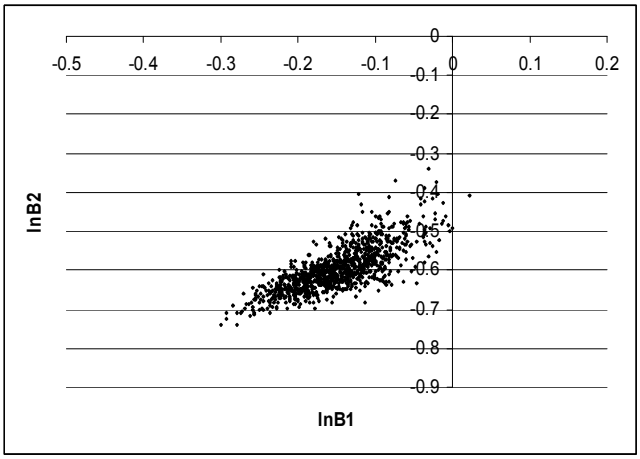
- Creation of a sample of *n* individuals, where *n* is not the size of the population, but just amount of individuals that we want to have in final sample. Each individual contains *m* unlinked microsatellites. At the beginning all microsatellites have the same length.
- repeating until there is only one individual in the sample of the following:
    - choosing whether the subsequent event is coalescent or mutation,
    - if its coalescent event choosing two individuals to coalescent and merge them, thus obtaining *n*-1 individuals in the sample,
    - if its mutation event choosing lineage to be mutated, determination of the type of mutation and application of it on some individuals.



(a)

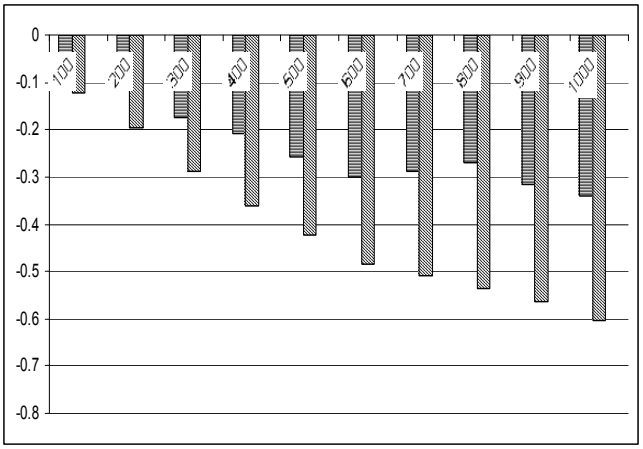


(b)

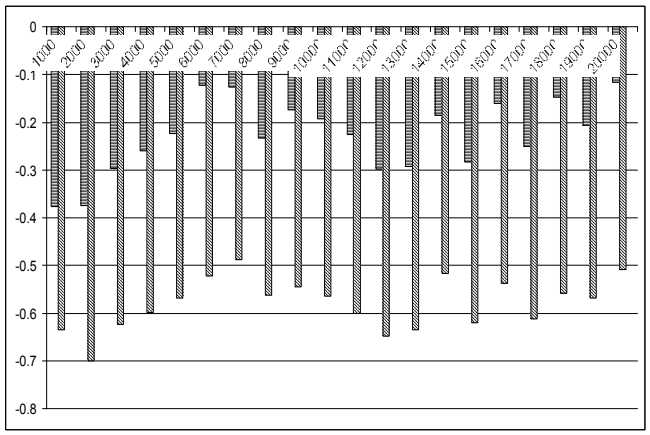


(c)

Fig. 4. $\ln\hat{\beta}_1$ as a function of $\ln\hat{\beta}_2$. Constant population with $N = 2500$ individuals, mutation rate $v = 5 \times 10^{-4}$, each individual has 30 microsatellites. We took 1000 samples, 40 individuals each from a) 10 000, b) 32 000, c) 64 000 generation, for each sample $\ln\hat{\beta}_1$ and $\ln\hat{\beta}_2$ were computed.



(a)



(b)

Fig. 5. Values of $\ln\hat{\beta}_1$ (horizontal lines) and $\ln\hat{\beta}_2$ (slant lines) for population of $N = 2500$ individuals directly after simulation start (a) and after 10 000 generations from simulation start. Each individual has 30 microsatellites, Mutation rate $v = 5 \times 10^{-4}$.

Alternatively, it is also possible to create genealogy tree for microsatellite and then to add mutations [17]. In any case, at the start of simulation we have a sample of n individuals coming from whole generation of N individuals. It is worth to notice that N is just parameter in coalescent method and doesn't influence time of simulation.

III. Experiments

The following sections describe experiments with simulations.

A. Imbalance Indices

As the statistical information about a population we used growth coefficient based on microsatellites , called the imbalance index. Characteristics of the imbalance index are described in article [11], authors of which created series of samples for populations undergoing growth of different types and magnitudes. Simulations performed there were based on coalescent methods. We repeated simulations described in this article using forward simulation method.

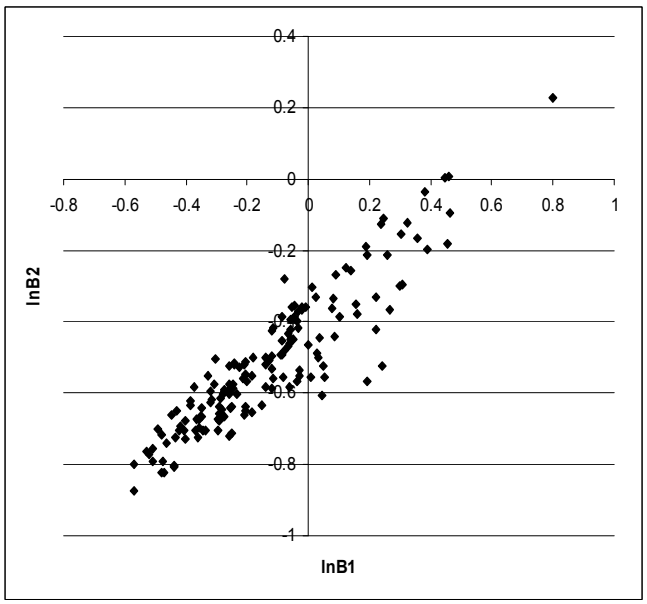


Fig. 6. $\ln\hat{\beta}_2$ as a function of $\ln\hat{\beta}_1$. 100 unlinked histories were created, each containing $N = 2500$ individuals, amount of individuals in generations were constant, mutation rate $v = 5 \times 10^{-4}$, each individual has 30 microsatellites. From 100 000 generation of each population 100 samples were taken. Each sample contained 40 individuals. For each populations mean of $\ln\hat{\beta}_1$ and $\ln\hat{\beta}_2$ were count. Those means were put on graph

There are two estimators of imbalance index:

$$\ln\hat{\beta}_1 = \ln\hat{\theta}_{\bar{V}} - \ln\hat{\theta}_{\bar{P}o} \tag{1}$$

and

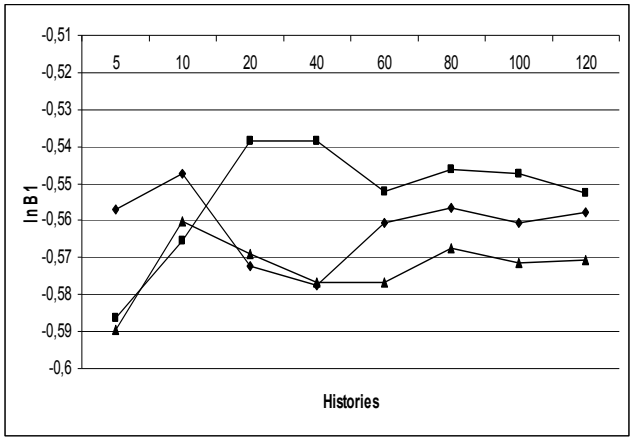
$$\ln\hat{\beta}_2 = \frac{1}{m}\sum_{i=1}^{m}\left((\ln\hat{\theta}_V)_i - (\ln\hat{\theta}_{Po})_i\right), \tag{2}$$

where m is the amount of microsatellites, $\hat{\theta}_V$ is the allele size variance estimator of θ, $\hat{\theta}_{Po}$ denotes the homozygosity estimator of θ, and finally the meaning of $\theta=4N\mu$ called composite parameter, is connected with the scale of the process. The reader interested in more in depth understanding of (1) and (2) should refer to [11] where these equations are explained in detail.

B. Initial Simulations

At the beginning of the simulation we have to initialize microsatellites' lengths in such a way that all microsatellites have the same length, and then we run simulation for $2N$ to $4N$ generations [16]. Observe, that the population in all generations has got the same amount of individuals. During

this time period values of coefficient are stabilising (Fig. 5) and populations reach mutation – drift equilibrium. After this pre simulation period, we are able to take significant samples from generation and simulate population growth.



(a)

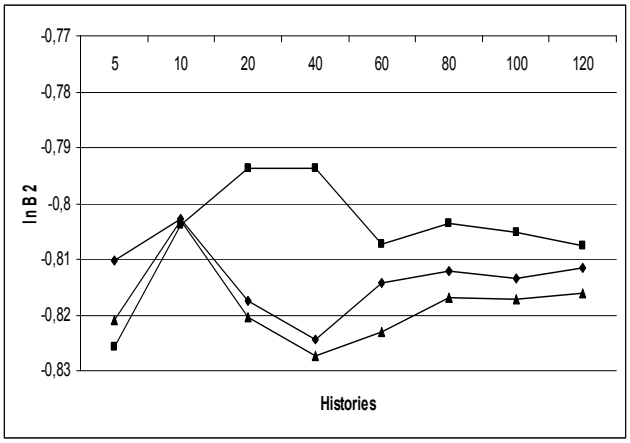
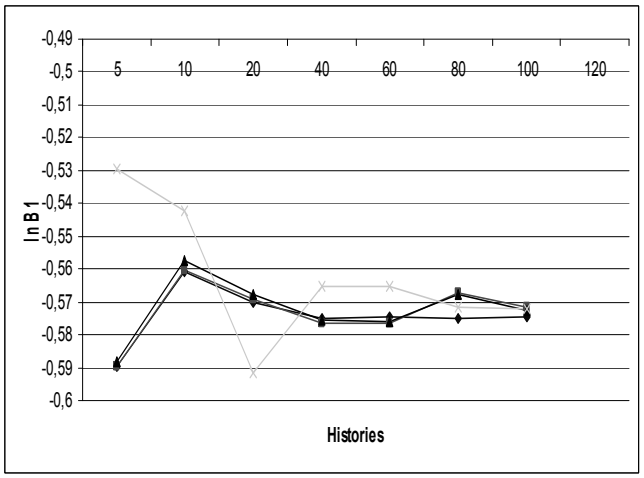


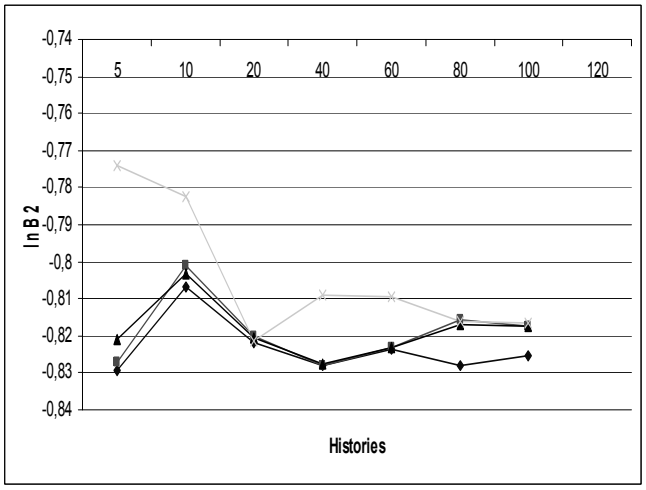
Fig. 7 Cut off values of $\ln \hat{\beta}_1$ and $\ln \hat{\beta}_2$ based upon population with different constant size 2500 ♦ 5000 ■ and 20 000 ▲. We used different amount of histories to count cut off values. We simulated 1 000 000 generations for each history, we remembered every population with number divisible by 10^5. For each population size, we created 120 unlinked histories. Each individual possessed 30 unlinked microsatellites. Mutation rate $v = 5 \times 10^{-4}$. From each history we took 100 samples each contained 4 individuals. For more than 100 generations cut off values are stabilizing.

C. Minimal Amount of Unlinked Histories

Important issue of forward computer simulation is minimal amount of unlinked histories that is needed in order to gain significant results. Every unlinked history has different coefficient values (Fig. 6) and our empirical tests showed us that for constant samples of different sizes 60 histories is enough to achieve stabilization of imbalance index estimators' cut off values (Fig. 7 and Fig. 8).



(a)



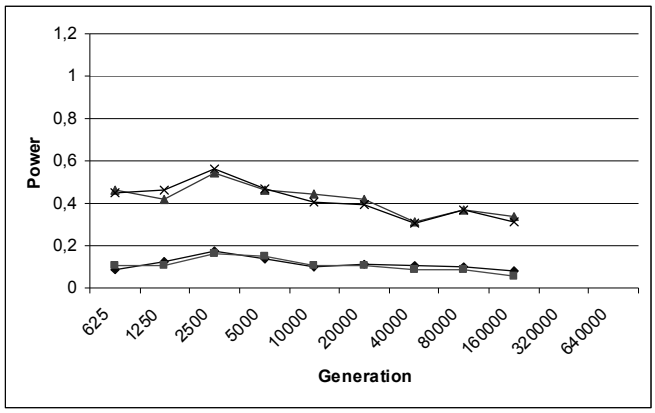
(b)

Fig. 8. Cut off values of (a) $\ln \hat{\beta}_1$ and (b) $\ln \hat{\beta}_2$ based upon constant size population N = 20 000 individuals and different amounts of unlinked histories from 5 to 100. Each sample contains 100 individuals. All graphs are based on the same histories but different samples taken from those histories.

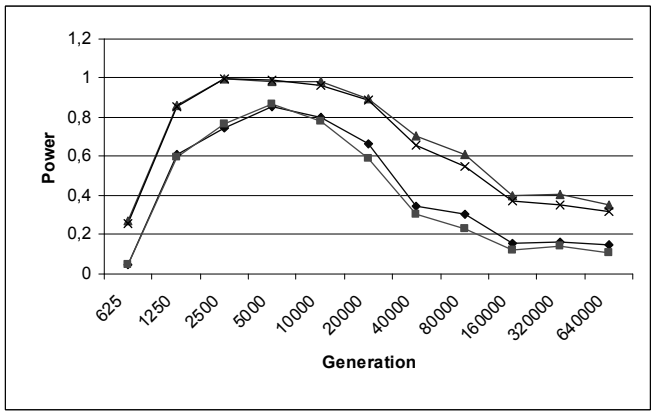
When we take second set of samples from population and count cut off values we should gain similar results to those obtained in the previous test. If we shuffle the samples the results will be more divergent but with increasing amount of individuals outcomes being similar.

In our experiment the test has been performed for constant sample of $N = 2 \times 10^4$ individuals, which we simulated for 10^6 generations and we used generations divisible by 10^5 to count cut off values. Each individual contained 30 microsatellites an we took from every generation 100 samples with 40 individuals each. In our experiments we always used 100 unlinked histories.

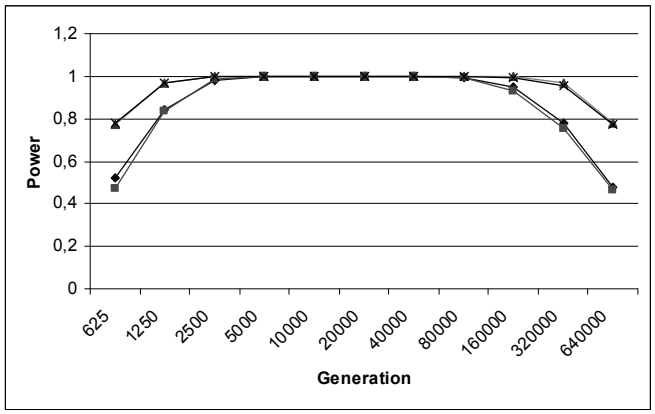
IV. Results



(a)

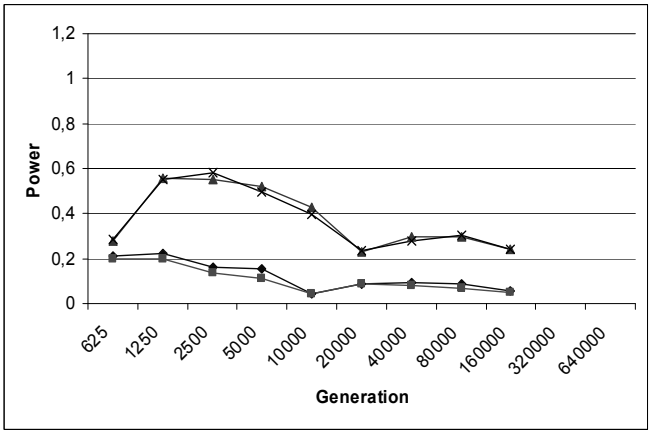


(b)

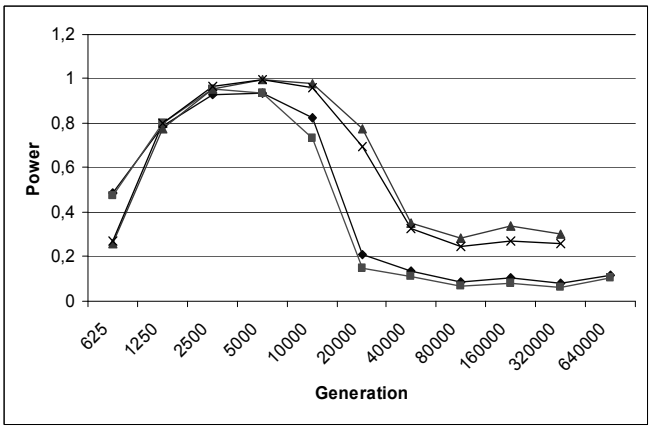


(c)

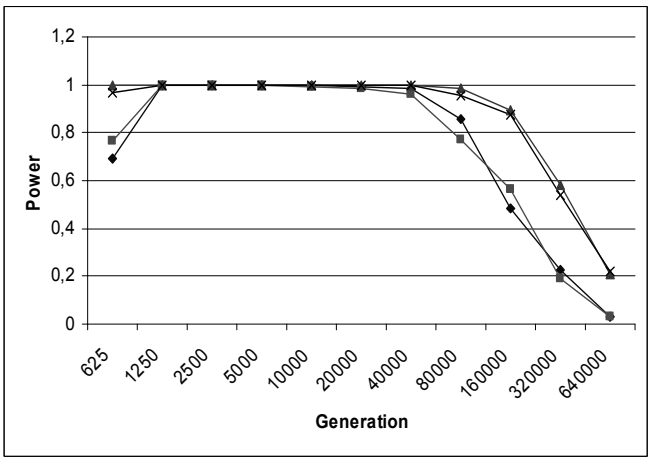
Fig. 9. Power of imbalance index's estimators for different cut off values $\ln \hat{\beta}_1$ ▲ and $\ln \hat{\beta}_2$ * based on coalescent methods, $\ln \hat{\beta}_1$ ♦ and $\ln \hat{\beta}_2$ ■ based on time forward computer simulation. Exponential growth from N = 2 500 individuals to (a) 5000, (b) 25 000 and (c) 250 000 individuals in different time period (amount of generations). We simulated 100 unlinked histories for each scenario. Mutation rate was $v = 5 \times 10^{-4}$, we took 100 samples from each unlinked histories each contains 40 individuals. Individual has 30 microsatellites.



(a)



(b)



(c)

Fig. 10. Power of imbalance index's estimators for different cut off values $\ln \hat{\beta}_1$ ▲ and $\ln \hat{\beta}_2$ * based on coalescent methods, $\ln \hat{\beta}_1$ ♦ and $\ln \hat{\beta}_2$ ■ based on time forward computer simulation. Stepwise growth from N = 2 500 individuals to (a) 5000, (b) 25 000 and (c) 250 000 individuals. We simulated 100 unlinked histories. Mutation rate was $v = 5 \times 10^{-4}$, we took 100 samples from each unlinked histories each contains 40 individuals. Individual has 30 microsatellites..

We created series of histories for different population growth scenarios. In our experiments we achieved different cut off values of imbalance index estimator than gained in [11]. Power of estimators for new cut off values $\ln\hat{\beta}_1$ = -0,55 and $\ln\hat{\beta}_2$ = -0,81 are lower than that estimated in article [11] for cut off $\ln\hat{\beta}_1$ = - 0,32 and $\ln\hat{\beta}_2$ = -0,65, as visible in Fig. 9 and Fig. 10.

All conditions of experiments were the same as in [11] and the only difference was different method of simulation used in our experiments, namely the time forward simulation, giving more reliable results.

V. Discussion

Our computer simulations performed showed that time forward simulation methods might give different results than those obtained from coalescent-based methods. Because time forward simulation methods are closer to real life scenario it might be appropriate to consider using this method of simulation especially when we have access to new fast computers or we want to count coefficients that we can directly apply to samples from real world.

Moreover, time forward computer simulations let us easier create complicated demographic histories and we can have access to every individual from every generation of the history and for example modify probabilities of having successors. It is also easier to model processes connected with geographic structure of the population and to test influence of such structured population on various, genetically important, factors.

VI. Acknowledgment

The authors would like to thank to Prof. Marek Kimmel from Department of Statistics at Rice University in Houston TX, USA, for long discussions and advice concerning population biology problems.

References

[1] P. Marjoram and **J**. D Wall, "Fast "coalescent" simulation", *BMC Genet.,* vol.7: no. 16., 2006.

[2] A. Renwick, L. Davison, H. Spratt, J. P. King, and M. Kimmel, "DNA Dinucleotide Evolution in Humans: Fitting Theory to Facts", *Genetics*, vol. 159, no. 2, pp. 737-47, 2001.

[3] D. B. Goldstein, D. D. Pollock, Launching Microsatellites: A Review of Mutation Processes and Methods of Phylogenetic Inference, J Hered 1997 Sep-Oct; 88(5):335-42.

[4] E. A. Sia, Ch. A. Butler, M. Dominska, P. Greenwell, Th. D. Fox, and Th. D. Petes, "Analysis of microsatellite mutations in the mitochondrial DNA of Saccharomyces cerevisiae", *Proc Natl Acad Sci U S A*,. vol. 97, no. 1, pp. 250–255, 2000.

[5] M. Kimura, T. Ohta, "Stepwise mutation model and distribution of allelic frequencies in a finite population", *Proc Natl Acad Sci U S A*, vol. 75, no. 6, pp. 2868–2872, 1978.

[6] L. A. Zhivotovsky, MW Feldman, SA Grishechkin. "Biased mutations and microsatellite variation ", *Mol Biol Evol,* vol. 14, no. 9. pp. 926-933, 1997

[7] I. Agrafioti1 and M. P. H. Stumpf**,** "SNPSTR: a database of compound microsatellite-SNP markers", *Nucleic Acids Res*., vol. 35, supplement 1, pp. D71-D75, 2007.

[8] J. Hein, M. H Schierup, and C. Wiuf, *Gene genealogies, variation and evolution: a primer in coalescent theory,* Oxford, New York, Oxford University Press, 2005.

[9] D. Bachtrog and B. Charlesworth, "Towards a complete sequence of the human Y chromosome", *Genome Biol*., vol. 2, no. 5, reviews1016.1–reviews1016.5., 2001.

[10] A. Eyre-Walker, Ph. Awadalla, "Does Human mtDNA Recombine?," *J Mol Evol*, vol. 53, pp. 430–435, 2001.

[11] J. P. King, M. Kimmel, and R. Chakraborty, "A Power Analysis of Microsatellite-Based Statistics for Inferring Past Population Growth," *Mol Biol Evol*. Vol. 17, no. 12, pp. 1859-1868, 2000.

[12] K. A. Cyran, "Simulating branching processes in the problem of Mitochondrial Eve dating based on coalescent distributions", *International Journal of Mathematics and Computers in Simulation*, vol. 1, no. 3, pp. 268-274, 2007.

[13] K. A. Cyran and D. Myszor, "New Artificial Neural Network based Test for the Detection of Past Population Expansion using Microsatellite Loci", send to International Journal of Applied Mathematics and Informatics.

[14] K. A. Cyran, "Image recognition with a diffractive optical variable device and probabilistic neural network: improvements of feature space", *WSEAS Trans. on Information Science and Applications*, vol. 2, no. 12, pp. 2212-2219, 2005.

[15] K. A. Cyran, "Comparison of neural network and rule-based classifiers used as selection determinants in evolution of feature space", *WSEAS Trans. on Systems*, vol. 6, no. 3, pp. 549-555, 2007.

[16] M. J. Donnelly, M. C. Licht, and T. Lehmann, "Evidence for recent population expansion in the evolutionary history of the malaria vectors Anopheles arabiensis and Anopheles gambiae", *Mol Biol Evol*., vol. 18, no. 7, pp. 1353-1364, 2001.

[17] J. Wakeley, *Coalescent theory an introduction*, Harvard University, 2008.

Krzysztof A. Cyran was born in Cracow, Poland, in 1968. He received MSc degree in computer science (1992) and PhD degree (with honours) in technical sciences with specialty in computer science (2000) from the Silesian University of Technology SUT, Gliwice, Poland. His PhD dissertation addresses the problem of image recognition with the use of computer generated holograms applied as ring-wedge detectors.

He has been an author and co-author of more than 70 scientific papers in journals (several of them indexed by Thomson Scientific) and conference proceedings. These include scientific articles like: K. A. Cyran and A. Mrózek, "Rough sets in hybrid methods for pattern recognition," *Int. J. Intel. Syst.*, vol. 16, 2001, pp. 149-168, and K. A. Cyran and M. Kimmel, "Interactions of Neanderthals and modern humans: what can be inferred from mitochondrial DNA?" *Math. Biosci. Eng.*, vol. 2, 2005, pp. 487-498, as well as a monograph: U. Stańczyk, K. Cyran, and B. Pochopień, *Theory of Logic Circuits*, vol 1 and 2, Gliwice: Publishers of the Silesian University of Technology, 2007. Dr. Cyran (in 2003-2004) was a Visiting Scholar in Department of Statistics at Rice University in Houston, US. He is currently the Assistant Professor and the Vice-Head of the Institute of Informatics at Silesian University of Technology, Gliwice, Poland. His current research interests are in image recognition and processing, artificial intelligence, digital circuits, decision support systems, rough sets, computational population genetics and bioinformatics.

Dr. Cyran has been involved in numerous statutory projects led at the Institute of Informatics and some scientific grants awarded by the State Committee for Scientific Research. He also has received several awards of the Rector of the Silesian University of Technology for his scientific achievements. In 2004-2005 he was a member of International Society for

Computational Biology. Currently he is a member of the Editorial Board of Journal of Biological Systems, member of the Scientific Program Committee of WSEAS international conferences in Malta (ECC'08), Rodos (AIC'08, ISCGAV'08, ISTASC'08) and multiconference in Crete (CSCC'08) as well as a reviewer for Studia Informatica and such journals indexed by Thompson Scientific as: Optoelectronic Review, Mathematical Biosciences and Engineering, and Journal of Biological Systems.

Dariusz Myszor was born in Pszczyna, Poland in 1983. He received MSc degree in computer science (2007) from the Silesian University of Technology SUT, Gliwice, Poland.

Currently he is a PhD student at Silesian University of Technology and is involved in scientific and didactiv activities of the Institute of Informatics at SUT.

His current research interests are focused in such areas as: neural networks, computational population genetics, computer simulations and bioinformatics.