NWO РФФИ

RESEARCH PROPOSAL
DUTCH-RUSSIAN RESEARCH COOPERATION 2003

1 **Please select the scientific field:**

2 **Please state the title of your project:** Development of algorithms and software for high-performance computing in genetic analysis of complex human traits

PERSONALIA OF APPLICANTS

Please complete these fields for the Dutch Project leader (applicant):

3 **Name and title(s):** Prof. C. M. van Duijn

4 **University or Institution:** Erasmus MC Rotterdam
Address: Dr. Molewaterplein 50, 3015GE Rotterdam, The Netherlands
Phone, fax and e-mail Phone:+31 10 4087394;Fax:+31 10 4089382; e-mail: c.vanduijn@erasmusmc.nl

Please complete these fields for the Russian co-leader (co-applicant):

5 **Name and title(s):** Prof. T. I. Axenovich

6 **University or Institution:** Institute of Cytology and Genetics (IC&G) SD RAS
Address: Lavrentjeva ave10,Novosibirsk,630090, Russia
Phone, fax and e-mail Phone:+7 383 2332840;Fax:+7 383 2331278; e-mail: aks@bionet.nsc.ru

………………………….. ……………………………….
Signature Dutch applicant Signature Russian co-applicant

PROJECT CONTENT

7 **Please insert the estimated duration and the start date of the project:**

36 Months January 04 Start Date

8 **Please give a summary of your project in terms of its content and goals:**
(max. 500 words, plus keywords):

The goal of the project is to develop a series of linked algorithms and software programs for high-performance computing in genetic analysis of complex human traits. This software should make semi-automatic discovery of genes involved in complex diseases possible. The algorithms must take into account evidence coming from different levels of genetic analysis (linkage, association studies, knowledge of the sequence of the human genome, literature data about disease). We will focus on exploiting highly parallelizable computation techniques in genetic analysis, building upon our previous join research. The software will specifically target the analysis of large pedigrees spanning 5 or more generations, as can be found in human isolated populations and life stock. A parallel computer system (cluster) will be constructed to support software testing and high-performance computing . The algorithms and software will be tested and validated using data simulated under various genetic models. Also, commercial software and software available in the public domain (if available) will be used as a golden standard for comparison. Finally, the data will be applied to the numerous data sets that have been obtained in ongoing research projects of Erasmus and IC&G.

Keywords:

computational genetics, genomics, complex disease and traits, linkage and association analysis, meta-analysis, linkage disequilibrium, pedigrees, model of inheritance

9 **Please describe briefly (by name) how young scientists will be involved in this project and what will be their role:**

From the Dutch side, Yurii Aulchenko will be primarily responsible for problem specification, every-day coordination and real-data testing of algorithms and software. He will be also responsible for the project documentation. He will be assisted by C. vanDuijn. From Russian side, A. Kirichenko will be responsible for software development and execution of this part of the project. A. Zykovich and I. Akberdin will take part in development of algorithms and software. D. Grigorovich will be responsible for development of a user interface.

10 **Please give a brief description of the previous research cooperation between your institutions.** Mention some recent joint publications:

Our institutions are involved in joint projects in computational genetics for more then two years. Particularly, we work on fundamental studies of human and animal populations and studies of genetics of Type 2 Diabetes Mellitus. Joint publications:

Aulchenko YS, Axenovich TI, Mackay I, van Duijn CM (2003) "'miLD and booLD Programs for Calculation and Analysis of Corrected Linkage Disequilibrium" Ann. Hum. Genet. (in press)

Aulchenko YS, Axenovich TI, Borodin PM (2003) "Inheritance of Litter Size at Birth in the House Musk Shrew and Brazilian Grass Mouse", XIXth International Genetics Congress, Melbourne, Australia (in press)

Aulchenko YS, Vaesen N, Heutink P, Mackay I et al. (2003) "Mapping Genes Influencing Type 2 Diabetes Mellitus in a recent genetically isolated Dutch population", Diabetes (in press)

Aulchenko YS, Heutink P, Mackay I, Bertoli-Avella A et al. (2003) "Linkage Disequilibrium in Young Genetically Isolated Dutch Population" Am. J. Hum. Genet. (submitted)

11 **Describe the state-of-the-art in the research field in the Russian federation and in the Netherlands:**

The Novosibirsk lab is the only lab in the Russian Federation that is working on development of methods for computer analysis of the inheritance of complex traits in humans and animals. It has developed a series of algorithms and software packages for segregation analysis of binary and complex traits (MAN1), pedigree drawing (Pedigree Query) and processing (MAIA) and modified other packages for specific purposes (Fisher-cvcrF2, Loki-QTLc). The lab is working on algorithm development using classical likelihood and modern Markov-Chain Monte-Carlo approaches. Although other Russian laboratories are working in the field of human genetics, their work is predominantly focused on studies of genetic polymorphisms in human populations. Those involved in studies of genetic epidemiology use standard software for analysis of association.

In the Dutch lab, several projects on the genetics of complex diseases were initiated, including the recently started Erasmus Rucphen Family (ERF) study, which concentrates on unraveling genes underlying quantitative traits in humans. The ERF study includes 2500 relatives from a recent genetically isolated population. Basically, the 2500 subjects consists of 100 2 or 3-generation families which go back to 30 founding couples living in the isolate in the period 1850-1900. Participants are screened for many quantitative traits related to cardiovascular disease, neuropsychiatric problems, eye pathology and endocrine disorders. The protocol includes a non-invasive assessment of the presence of atherosclerosis in the carotid artery and a low-radiation whole-body imaging yielding information of bone mass density, fat percentage and distribution and muscle content. The ERF study is the largest family study of this kind in Netherlands. The study is one of the key cohorts in the Center for Medical Systems Biology, one of the focus programs of the Dutch initiative for genomic research.

The analysis of multiple traits in a large kindred is imposing a major challenge on existing computational techniques. Basically the pedigrees with multiple loops, spanning over 5 generations can not be analyzed using software available in the public domain. The problem encountered is similar to that seen in life-stock research. To analyze the data obtained in the ERF project, new methods and software is to be developed, which can also be used in large pedigree studies in human and live stock populations elsewhere. The aim of the project proposed here is to develop a user friendly software application for such analysis.

12 **Discuss the content of your project:**

- What is the problem definition or central hypothesis of your project?

The whole program of development and functioning of living organisms is “written down” as a sequence of nucleotides in their DNA. Modern techniques make it possible to read these sequences. The main problem now is to make sense of this information, to find out which of the numerous polymorphic genes are involved in the pathogenesis of common and complex human diseases such as cardiovascular disease, cancer, neuropsychiatric disorders, bone disorders, asthma, and diabetes. To date several approaches and algorithms have been developed in order to link genes to diseases. A problem faced in practice is that only few have been translated into user friendly software packages.

Most of the existing programs are focused at completely automatic data processing and analysis. The efficiency of the strategies used to date is very low partly because it does not and cannot take into account specific features of specific diseases and features of the patients (pedigree and population structure, environmental influence, and so on). Further, none of the existing software programs is able to process extended multi-looped pedigrees spanning 5 generations or more.

The main goal of our project is to develop algorithms and software for genetic analysis of complex diseases in large human and livestock pedigrees, which takes into account the specific features of the disease (trait), population of subjects studied and existing knowledge about the disease (trait). We will make use of methods and algorithms that have been or are being proposed by us and others and incorporate those in our object-orientated software. If necessary, new algorithms will be developed.

We are going to develop an interconnected "language" which allows exploration of genetic data on various features, e.g., characteristics of the disease and the underlying genetic model. It will allow the researcher to consider patient and family specific characteristics and genetic models as objects having particular properties. This will allow sequential and flexible exploration of genetic data. Our final goal is to develop user-defined procedures with the help of object oriented programming. As we are planning to do an Open Source project, fully scalable and well-documented, we anticipate many third parties will use the results of the project.

- Discuss your method of approach:

Our approach is based on an iterative approach to discover genes for complex disease.

During preliminary analysis at the start of the study, specific features of the disease (diagnostic certainty, age of manifestation, risk factors and accompanying diseases) and the structure of the sample (pedigree, population structure and origin) are considered. For selected diseases, the literature data on the disease features and important covariates, results from previous genome scans and candidate-gene studies will be collected allowing for better model decision. The data on the study sample will be analyzed separately and in a meta-analysis together with the data available from the literature. The results of these analyses will make it possible to select a sub-set of genetic models which fit to these specific features of the disease.

Next, the gene discovery process enters iterative stage. As we will specifically focus on quantitative traits involved in complex disease, by definition there will be multiple genes and environmental cofactors involved in the trait which have different effects on the trait according to different genetic models. We will exploit in the analysis a standard but adjustable series of models of inheritance which take into account various genetic and environmental factors. A strategy for the analysis of these models has been developed at the Novosibirsk laboratory and several other laboratories. We will focus specifically on the Mixed Model of Major- and Polygenic Inheritance (MMMPI), assuming a hierarchy of gene effects. Within the framework of MMMPI, the first analysis will target the identification and localization of the (unknown) gene with the largest effect on the trait under the study. This is achieved by model-based linkage analysis. This approach is theoretically most powerful given that the right model is specified and has been shown to be reliable given correct model is used. After this step, the localized gene is considered as a covariate (fixed effect) in the next analysis. In the nex step, we can go for the next iteration, now focusing on the next (unknown) genes, which, after adjusting for the gene(s) localized in previous steps as covariates, has the strongest effect on the trait under study. For each new gene, we will select the optimal genetic model for data analysis. The iterative steps will be continued until there is no more evidence for unknown genes involved in the traits which can be detected within the power limitations of the study.

In our algorithms for gene localization, we will include an option which will allow weighting the results on previous findings as published in literature and genome databases. Again this weighting process will be based on meta analysis, as has been recently advocated (Lohmuller, Nat. Genet. 2003;33:177-82).

The same iterative approach will be implemented for fine-mapping, which will be done by the means of linkage disequilibrium analysis. For fine-mapping, again there will be an evaluation of the disease segregation and the linkage/association of the disease to the haplotype segregating in the family in order to select the individuals most informative to narrow down the region of interest. This information will be included in the subsequent analyses using a weighting procedure. Again, we will incorporate in weighing the effects of the various loci and other covariates simultaneously in the analysis as fixed effects.

Our approach will allow the researcher to make use of all information available on the disease and its distribution in population and pedigree. Analysis using variants of MMMPI is very complex. However, it allows correct estimation of power, unbiased assesment of significant levels and so on. We therefore intend to develop an automated, user friendly software application that will be made available for third parties. Another problem to overcome is that at any step, the genetic computations are very intensive, in particular in extended pedigrees. Though it is well-known that algorithms used in genetic analysis could be easily parallelized, only few programs exploit this feature. We will focus specfically on parallelization of our computations. One of the most obvious examples, where parallelization will allow for linear increase in speed with the number of nodes, is for example in statistical power and significance estimation. Other applications (parallelization by chromosome and pedigree) are less effective but are still straightforward and may in particular be effective when analyzing large pedigrees spanning 5 generation or more.

The programs will be developed using object orientated programming (Fortran 90/95 and C). The approach has been followed successfully earlier at Novosibirsk. Before using the data in gene discovery, we will evaluate its performance. As a test system, we will use extensive simulations and also a "proof-of-principle" project that is currently developed in Erasmus. The simulated data set will involve traits influenced by multiple genes, with different effects and underlying genetic models. Using such a set in where the models are known a priori, we can evaluate the statistical power and precision of the algorithms and software. The proof-of-principle project aims to re-confirm the role of the genes which are known to play a role in the pathogenesis of a disease under study using ERF. The main question will be to find a relation between for instance apolipoprotein E and serum cholesterol, or between apolipoprotein E and cognition or collagen 1A2 and bone mineral density. Successful re-discovery will validate the system for real-data analysis. Finally, we will also compare the performance of our software with that of existing (commercial) packages including Genehunter, Merlin, Allegro, Solar and Sage. As these programs are not be able to include the large pedigree(s) studied in ERF, the pedigree will be cut and then analysed as unrelated pedigrees.

- What are the objectives of the project?

The aim of our project is to deliver a user-friendly package of high-performance software and a cluster for step-by-step genetic analysis of complex diseases and other traits. This package will be able to take into account the specific features of the disease, population under the study and existing knowledge about the disease (trait). We will make use of methods and algorithms that have been or are being proposed by others, but are not yet implemented and integrated. If necessary, new ones will be developed. We will develop a system of commands, which will allow for flexible genetic analysis. We will specifically focus on the statistical analysis of extended pedigrees. We will apply the software and the cluster to discover new genes underlying human trait variations using the data from on-going Erasmus studies and other studies of extended pedigrees in the Netherlands and the research on human and life stock traits within the Russian Federations.

13 **Describe the division of tasks between the Dutch and the Russian researchers:**

Dutch Research tasks:

(1) setting up the task and requirements to algorithms and software (2) control and evaluation of the project (3) testing of software using real data

Russian Research tasks:

(1) development of the general strategy of the computing system (2) development of high-performance cluster, optimised for tasks for genetic epidemiology (3) development of algorithms and software (4) development of user-friendly interface and extensive documentation

14 **Describe the significance and innovative aspects of the project:**

Computer methods of identification and mapping of genes is very important for the development of prevention and treatment strategies for complex diseases. The progress in this direction is impeded by a tendency to make completely automated one-step methods. As a result the majority of these methods are not able to make use all information available about each particular disease in each particular population. We suggest a flexible step-by-step computer method in which every next step can use the results of the preceding analysis, including the analysis of already published results, and allows the user to choose the optimal method of gene identification and mapping taking into account the specific knowelege of the trait. Though it is well-known that algorithms used in genetic analysis could be easily and effectively parallelized, only few programs exploit this feature. Our project aims at parallelization of computations, an approach almost neglected up to date in software available within the public domain. The impact of this approach is particularly high when analyzing extended pedigrees in humans or livestock. The development of a formalized system of commands for operation with genetic data and models may allow better formalization and standardization of many tasks of modern genetics.

15 **Describe the expected results of the project (e.g. anticipated publications):**

The main results of the project will be a user-friendly package of software for step-by-step high-performance computing in genetic analysis of complex human diseases. Each step will give not only a knowledge on the genetic control of the disease, but also a background for further analysis (set of genetic models, estimate of information value of pedigrees, choice of the number and type of genetic markers, list of candidate genes and linked markers, literature data about this region and so on). In the course of the study we shall estimate the efficiency of various strategies, and create new algorithms. We will also analyze empirical data obtained in ongoing studies in Rotterdam, Novosibirsk and elsewhere. The results of these studies will be published in medical genetic and computer science journals such as American Journal of Human Genetics, Human Molecular Genetics, Bioinformatics, Genetic Epidemiology, American Journal of Medical Genetics

16 **Describe possible future application results:**

The scalable package of software for step-by-step high-performance computing in genetic analysis of complex human diseases (traits) resulted from this project will be of use in fundamental and applied human and medical genetics and genetic epidemiology (genetic analysis, gene-finding, genetic counseling). It would be also applicable in analysis of livestock data. The cluster will be extensively used for analysis of data obtained in on-going projects of Erasmus, and, if supported further, by the world community.

PERSONALIA CONTINUED

17 **Please provide names and contact information of the Russian participating researchers per institute**

Institute 1

Institute name nr.1: Institute of Cytology and Genetics SD RAS
Group Leader: Prof. T. I. Axenovich
Institute Address, Telephone/Fax Number and e-mail address:
Lavrentjeva ave. 10, Novosibirsk, 630090, Russia; Phone: +7 383 233 2840; Fax: +7 383 233 1278; e-mail: aks@bionet.nsc.ru

Key Researchers institute nr.1 (no more than 2):
Prof. P. M. Borodin

Dr. I. V. Zorkoltseva

Telephone and e-mail addresses:

Prof. Borodin: +7 383 233 28 13, borodin@bionet.nsc.ru; Dr. Zorkoltseva: +7 383 233 2813; zor@bionet.nsc.ru

Young scientists institute nr.1:

Dr. A. Kirichenko	Age: 27
A. Zykovich	Age: 22
I. Akberdin	Age: 22
D. Grigorovich	Age: 30

Telephone and e-mail addresses:

Dr. Kirichenko: +7 383 233 2840; kianvl@mail.ru; Zykovich:+7 383 233 2840; zykovich@bionet.nsc.ru; Akberdin: +7 383 233 2840; akberdin@bionet.nsc.ru; Grigorovich: +7 383 233 2840; grigorovich@bionet.nsc.ru;

18 Please give the names and contact information of the participating Dutch researchers per institute

Institute 1

Institute name nr.1: Erasmus MC Rotterdam

Group Leader: Prof. C. M. van Duijn

Institute Address, Telephone/Fax Number and e-mail address:

Dr. Molewaterplein 50, 3015GE Rotterdam, The Netherlands; Phone: +31 10 408 7394; Fax: +31 10 408 9382; e-mail: c.vanduijn@erasmusmc.nl

Key Researchers institute nr.1 (no more than 2):

Prof. P. van der Spek

Telephone and e-mail addresses:

+31 10 408 7299; p.vanderspek@erasmusmc.nl

Young scientists institute nr.1:

Dr. Y. S. Aulchenko	Age: 28
	Age:
	Age:
	Age:

Telephone and e-mail addresses:

+31 10 408 7486; i.aoultchenko@erasmusmc.nl

RESUMES OF RUSSIAN PARTNERS

19 **Please give short resumes of the Russian group leaders:**

Group leader 1:
Prof. Tatiana I. Axenovich (1949):
M.Sci. (Genetics) - Department of Cytology and Genetics, University of Novosibirsk (1971),
Ph.D. (Cytology) - Institute of Cytology and Genetics, Novosibirsk, (1977),
Dr.Sci. (Genetics) - Institute of Cytology and Genetics, Novosibirsk, (1996).
Career details:
Institute of Cytology and Genetics, Siberian Department of Russian Academy of Sciences
Postgraduate student (1971-1975)
Research Fellow (1975 - 1988)
Senior Research Fellow (1988 - 1994)
Head of the Laboratory of Methods of Genetic Analysis (1994 - to date)
Scientific interest: Statistical genetics, genetic analysis of complex traits, quantitative genetics.
Publications in: AmJHumGenet, GenetEpidem, GenetRes, AmJMedGenet, Genome

20 **Please give short resumes of the Russian key researchers:**

Key researcher 1:
Prof. Pavel. M. Borodin (1948):
M.Sci. (Genetics) - Department of Cytology and Genetics, University of Novosibirsk (1971),
Ph.D. (Genetics) - Institute of Cytology and Genetics, Novosibirsk, (1978),
D.Sci. (Genetics) - Institute of Cytology and Genetics, Novosibirsk, (1993).
Career details:
Institute of Cytology and Genetics, Siberian Department of Russian Academy of Sciences
Postgraduate student (1971-1973)
Research Fellow (1973 - 1983)
Senior Research Fellow (1983 - 1990)
Head of the Laboratory of Evolutionary Genetics (1990 – 1992)
Head of the Laboratoryof Recombination and Segregation Analysis (1994 - to date)
Scientific interest: Population genetics, genetic analysis of complex traits, cytogenetics.
Publications in: Science, GenetRes, AmJMedGenet, Genome

Key researcher 2:
Dr. Irina V. Zorkoltseva (1960):
M.Sci. (Applied Mathematics and Mechanics) - Department of Applied Mathematics, University of Novosibirsk (1983),
Ph.D. (Genetics) - Institute of Cytology and Genetics, Novosibirsk (2001),
Career details:

Postgraduate student of Institute of Applied Physics, Siberian Department of Russian Academy of Sciences (1983-1985)

Research Fellow of Institute of Theopretical and Applied Mechanics, Siberian Department of Russian Academy of Sciences(1983 - 1995)

Senior Research Fellow Institute of Cytology and Genetics, Siberian Department of Russian Academy of Sciences(1995 – to date)

Scientific interest: Computer science, statistical genetics, genetic analysis of complex traits.

Publications in: AmJMedGenet, RussJGenet

21 **Please give short resumes of the young scientists (include learning/working experience)**

Young scientist 1:

Dr. Anatolij V. Kirichenko (1976):

M.Sci. (Bioiformatics) - Department of Bioinformatics, University of Novosibirsk (2000),

Ph.D. (Bioiformatics) - Institute of Cytology and Genetics, Novosibirsk, (2003)

Career details:

Institute of Cytology and Genetics, Siberian Department of Russian Academy of Sciences

Postgraduate student (2000-2003)

Research Fellow (2003 – to date)

Publications in: RussJGenet

Young scientist 2:

Artem S. Zykovich (1981):

B.Sci. (Bioiformatics) - Department of Bioinformatics, University of Novosibirsk (2003)

M.Sci. student of Department of Bioinformatics, University of Novosibirsk (2003 –to date)

Young scientist 3:

Ilja R. Akberdin (1981):

B.Sci. (Bioiformatics) - Department of Bioinformatics, University of Novosibirsk (2003)

M.Sci student of Department of Bioinformatics, University of Novosibirsk (2003 –to date)

Young scientist 4:

Dmitrij A. Grigorovich (1972)

M. Sci. (Informatics and Applied Mathematics) - Department of Mathematics, University of Novosibirsk (1995)

Career details:

Institute of Cytology and Genetics, Siberian Department of Russian Academy of Sciences

Research Fellow (1995 – to date)

Publications in: InSilicoBiol., Bioinformatics, ComputApplBiosci., MolBiol