DR ALEXEI FEDOROV

Reexamining introns

Dr Alexei Fedorov explains the background to his genomic investigations into non-coding DNA, and also computer modelling's research potential in solving questions that have eluded researchers to date

To begin, can you outline the core principles and objectives of your research?

My lifetime research goal is to progress our understanding of the organisation and operation of the human genome to new frontiers. I approach this problem from three directions: the first is the investigation of intron cellular roles; the second is the analysis of the mid-range inhomogeneity (MRI) of the human genome sequence and finding important, yet unknown, signals associated with this inhomogeneity; and finally, I am involved in the development of a new computer model (the MAGE program) for studying human genome evolution in order to answer questions that cannot be solved by mathematical equations.

Could you describe the characteristics of introns?

Genes are pieces of DNA that are used as templates for production of useful RNA molecules. Frequently, the primary RNA transcripts of genes are not functional until specific segments of RNA, called introns, are excised from them, thereby converting them into active molecules. The most known genes are those that code for proteins. In vertebrates nine tenths of these genes possess introns. On average, there are eight introns per gene, yet in extremes it could be up to 300 introns. Furthermore, these introns are usually very long, so in total they occupy up to 30 per cent of our genome, while the important protein-coding portion of genes represents only 1.5 per cent of human DNA.

What makes introns such a double-edged sword for multicellular organisms?

The enormous intron size creates several drawbacks. These include the considerable waste of energy during gene expression which is 'unwisely' spent on generation of long intronic segments in primary RNA transcripts that should be removed. Introns also cause a delay in production of proteins (it takes several hours to obtain an RNA transcript from a gene with extra-long introns). Finally, errors in the intron removal are a frequent cause of genetic diseases. Some benefits must be associated with introns to compensate for these disadvantages. We have described six constructive roles for introns in two reviews, including the existence of non-protein coding RNAs inside introns, and different patterns of intron removal, known as alternative splicing, which allows the production of protein isoforms from one gene.

Moreover, why are the origins of introns and their cellular roles subject to such controversy?

For two decades after the discovery of introns in 1977, introns were widely considered as 'junk' or 'selfish' genomic elements. A popular view at that time was the 'introns-late' theory, hypothesising that introns originated at the relatively late stages of evolution (several hundred million years ago) and propagated like viruses or transposable elements, infecting a host organism and causing mild harm. At the same time there existed a contrary 'introns-early' opinion that introns are among the most ancient of genetic elements existing in the RNA world. This theory claimed that due to intron early existence, all modern genes were built from short pieces of ancient protein-coding regions. This decades-long dispute has resulted in the common opinion that introns were present at the time of origin of eukaryotes and exist in every species from the kingdom of life.

Can you offer an insight into the computer programs that you are developing alongside the databases and the role they will play in extracting and processing data?

Since my area is pure Bioinformatics, what I mainly do is create computer programs for the analysis of a variety of public databases. Some of these programs may generate our own databases of processed and classified sequences. During the past 10 years, my lab has produced hundreds of programs for the analysis of genomic sequences. The vast majority of them are small scripts written in Perl language for internal usage that identify particular signals and their arrangements within genomic sequences written in four letter code representing A, T, C, and G nucleotides. The programs allow us to obtain novel information about genomes, generate a hypothesis based on our results, make predictions and finally verify these predictions with newly written programs. Often this entire cycle takes only a few days and we repeat it multiple times for each project until we reach something very interesting and worth publishing. Beside this, we have a couple of computationally intensive projects that require very fast supercomputers and more robust programming in Java or C++.

DR ALEXEI FEDOROV

© Joshua Klein, Department of Art, University of Toledo. The image represents an artistic depiction of the Bioinformatics research for the human genome performed in the Fedorov's lab

Open thinking

The multidimensional issues currently faced within genomic science have led the **University of Toledo** to employ a number of highly innovative methods, and their computational work is being made freely available for the benefit of the research community as a whole

THE MATRIX ALGORITHMS for Genome Evolution (MAGE) program has been under development for a number of years, and is now producing important results. The program operates with chromosomal sequences, both human and non-human, working with a given number of modelled individuals. The sequences are then changed in a similar manner to real mutations, whilst observing the frequencies and locations of transitions of one nucleotide into another. The generational form of the simulation means that the computational cycle can be repeated over thousands or millions of generations. The ultimate goal is to describe unknown types of behaviour existing in large-scale ensembles of mutations during evolution. Results have already been obtained which demonstrate intriguing links between genome length, number of chromosomes and the rate of meiotic crossovers on human fitness as a whole.

Such findings are of great relevance to the study of the human genome, since mutations are of such importance. Averaging the available data from a number of sources, it seems that each human individual has around 80 novel mutations not present in the parental DNA. The current understanding of population genetics is insufficient to predict the way in which such a large number of mutations will have on mankind. In fact, some geneticists predict the inevitable collapse in the near future, due to so called 'genetic entropy', and this is why MAGE's modelling of populations is so valuable. Furthermore, current theorists in the field are split between 'Neutralism' and 'Selectionism' as opposing camps. The former maintain that beneficial mutations are rare, and certainly far less frequently fixed than neutral or slightly deleterious mutations. Selectionists believe in the abundance of beneficial mutations. The issue is currently unresolvable because of the impossibility of evaluating the effect of single mutations in complex organisms through experiments, unless the mutation is deleterious.

INVESTIGATION OF INTRON CELLULAR ROLES

OBJECTIVES

To create and improve several databases representing gene structures, exploiting them in collaboration with three laboratories that experimentally study non-coding RNA genes, gene expression, and alternative splicing.

KEY COLLABORATORS

Shuhao Qiu

Arnab Saha Mandal

FUNDING

National Science Foundation – contract no. 0643542

CONTACT

Dr Alexei Fedorov
Associate Professor

Department of Medicine
Health Science Campus
The University of Toledo
3000 Arlington Avenue
Toledo, OH 43614-2598
USA

T +1 419 383 5270
E alexei.fedorov@utoledo.edu

www.utoledo.edu/med/depts/bioinfo

ALEXEI N FEDOROV is Associate Professor of Medicine at the University of Toledo, specialising in the origin and evolution of introns, computer mining of novel genes and prediction of constitutive and alternative splicing. He attended Moscow University, gaining an MSc in Physics before attaining his PhD from the Russian Academy of Sciences, Moscow in 1993. Since December 2010, he has held the position of Vice Director of the Bioinformatics and Genomics/Proteomics Program at Toledo University.

THE UNIVERSITY OF TOLEDO
1872

This, and other controversies in the field, have come to the point of being impossible to deal with using numerical mathematics, making modelling extremely attractive.

The team expect that MAGE will make an important intervention within this field. The program itself is a 120 page Java script, and has been designed specifically to mimic the nucleotide changes in thousands of genes. MAGE's central importance is in working generationally, with individuals forming mating pairs to produce the next generation. Each of these pairs will have been individually modelled by MAGE, generating meiotic recombinations between paternal and maternal chromosomes. At every stage of the program, the fittest children are chosen to produce the new generation, working as natural selection. Of course, what comprises the fitness value is hard to maintain, and it must at the very least be a multidimensional parameter, making it extremely hard to deal with mathematically. MAGE, however, is capable of dealing with this issue, and based on user defined environments, distributions are calculated, and for each individual a selection coefficient is calculated. Although still in relatively early stages, MAGE has already demonstrated that, even with large numbers of predominantly negative mutations, fitness may be maintained through thousands of generations with the help of frequent recombinations between maternal and paternal chromosomes. In addition, the longer genomes may improve the organism fitness. These findings raise further questions, which will continue to be investigated by the program.

INVESTING IN COMPUTATION

The Bioinformatics Lab at the University of Toledo has a number of other projects which utilise computational models in order to investigate genetic phenomena. They have recently published a paper on Nucleic Acids Research defining the new computational technique called Binary Abstraction Markov Models (BAMM) for sequence classification. Whilst these projects usually take upwards of a year to generate the final computer code and obtain reliable results, Toledo continues to invest in them as an excellent method for research. BAMM used the fastest supercomputers in Ohio in order to run, the Glenn cluster which has 8,000 powerful processors from the Ohio Supercomputer Station in Columbus. The data which was published derived from hundreds of hours on this supercomputer. Despite the massive investment to get to the workable stage, each programme developed is made publicly available, opening up the potential for global research. They can be obtained from publications, web pages or on request to the researchers.

SCIENCE OF SUPERCOMPUTING

The investment in computer modelling has become more prevalent over the past decade, and the University of Toledo has matched this trend. Given that a number of problems at the forefront of different types of research are currently impossible to solve mathematically, modelling is providing a way to approach these issues without needing to formulate or solve the mathematically complex calculations. Project Coordinator Dr Alexei Fedorov is excited by the possibilities that modelling can offer: "I consider a computer modelling of genome evolution as a special type of advanced Cellular Automata—'a new kind of science' that has been brilliantly described by Stephen Wolfram in his book of the same title. As the technology, as well as our understanding of computer modelling, improves, it is set to provide more answers to problems in different disciplines".

FURTHER PROJECTS

Fedorov's view of computer programming also feeds into his notions about DNA, whereby genomes are viewed as self-realisation programs that autonomously fulfil their tasks and are able to respond to environment signals and conditions. This overlap means that there are important links to be made with other sections of the work which is being completed by Fedorov's team. Genomic mid-range inhomogeneity (MRI) regions are sections of the DNA which do not conform to the normative pattern (B-form), which was revealed by Watson and Crick in 1953. These sequences can be dozens, or even hundreds of nucleotides long, and include (but are not limited to) Z-form, H-form and quadruplexes. Each of these unusual DNA conformations is associated with specific nucleotide compositions enriched by particular bases and in some cases with a particular nucleotide order. Just as with the computational programs, work on MRI regions has been made available through an internet resource, focussed on the identification and visualisation of such regions in longer genomic sequences. A recent breakthrough was the discovery that certain MRI distributions are not random, and these include both Z- and H-form DNA, with GC-, AT-, and GT-rich regions. These regularities have been published in part by Dr Ashwin Prakash and co-authors. This is opening up new possibilities for research into the relationship between MRI regions and the expression profiles of neighbouring protein coding regions, something which is going to be pursued in the next grant proposal forwarded. It is expected that our understanding of the human genome will continue to progress rapidly over the next decade, with potential being opened up for the therapeutic manipulation of genetic material. However the area does develop, Toledo's investment in research will keep it at the forefront of innovation, demonstrating the strength of openly available research and information.