Introduction.

Autosomal Recessive Gripe Hyperalgia (ARGH) belongs to the group of autosomal recessive disorders, many examples of which were well described and mapped.

The gene responsible for ARGH is recessive and therefore the syndrome manifests itself only when individuals have two alleles of the gene. IV/2, IV/8, IV/10, V/1 and V/7 belong to this group.

The syndrome is extremely rare and therefore any marriages outside family are very unlikely to bring the gene in, and the out-of-family spouses are very unlikely to be the ARGH gene carriers.

The family tree allows to identify the heterozygous carriers of the gene, they are II/1, II/2, III/1 - III/4, IV/1 and IV/7. Also, the following individuals have 66% chance to be the carriers as well: IV/3-IV/6, IV/9, IV/11, IV/12, V/2 - V/6. Each of the 15 kids from the out of family marriages of IV/3 - IV/5, IV/9 and IV/12 has 33% chance of being a carrier for ARGH gene.

The family tree for the described scenario is attached.

This essay aims to outline the strategy for identification of gene responsible for ARGH disorder. The process can be divided into three logical steps described below.

TASK 1

Identification of the genetic locus that contains the genetic defect.

Before starting the experimental work, more data on the disorder should be collected. At present time, huge number of genetic disorders has already been studied, and it would be beneficial to see if any similar genetic defects have already been mapped or roughly located on the genetic maps. This information might be found using OMIM, the Online database of Mendelian Inheritance in Man (www.ncbi.nlm.nih.gov/**omim**/). The database can also be researched to see if any similar traits are known in experimental animals. It can't be ruled out that the gene responsible for the analogous disorder in animals was already indentified, in which case the job of identifying the human gene will become much easier.

To solve the task of identifying the genetic locus that contains the genetic defect responsible for ARGH, the genomic DNA of all affected individuals should be analysed. To obtain the DNA, it will be necessary to take blood samples from each individual (several mL) and extract the DNA from leukocytes (no cell separation

required). No tissue samples are required (and therefore no biopsy needed) since the genomic DNA will be the same in every cell of the body.

Since disorder is not heterogeneous, it means that the ARGH syndrome is caused by a single genetic defect and we can exclude the probability that the disease can be caused by mutations in different genes or different mutations in the same gene.

ARGH is an autosomal recessive disease which means that it can be fine-mapped using extended pedigrees and ancestral haplotypes, because mutated alleles can be transmitted for many generations. Homozygosity in the case of ARGH is autozygosity which means that the markers are identical by descent (Strachan & Read, 2004)). Autozygosity mapping is particularly useful if families with the same recessive condition are available with a number of individuals linked by inbreeding. As a result, a small inbred family can generate a large LOD score. The LOD score (the logarithm of the odds of departure from random segregation) is an important factor in the linkage analysis (Morton, 1955). Genetics is based on probability, and mapping of loci depends on calculating this probability. Good examples of the successful application of autozygosity mapping include the locating of the genes for autosomal recessive hearing loss (Guilford et al., 1994), benign recurrent intrahepatic cholestasis (using only four affected individuals from an isolated Dutch village) (Houwen et al., 1994), autosomal recessive profound congenital deafness (Chaib et al., 1996), and pulmonary hypertrophic osteoarthropathy (Uppal et al., 2008).

In practical terms, a suitable set of polymorphic markers has to be chosen for this work. The markers should be spaced at intervals no greater than 10-20 cM across the genome. Microsatellite markers (2-4 nucleotides repetitive sequences, of which more than 10,000 are mapped) are usually used for linkage analysis and tracing the DNA fragment through family. Good example of using the microsatellite markers is the mapping of achondroplasia-hypochondroplasia (Le Merrer et al., 1994). Rough linkage analysis (20 cM) can be done with ABI standard set of 192 pairs of primers. 811 pairs give resolution of 5 cM (ABI Human Linkage Mapping v2.5 Primer Pairs:https://products.appliedbiosystems.com/ab/en/US/adirect/ab?cmd=catProductDetail&productID=4329191&catID=600774&backButton=true).

The primers are unique for each heterogeneous marker because they flank the marker sequence and are not complimentary to the markers themselves. The LOD score analysis can be used to locate the disease gene. The LOD score of +3 is the criterion for linkage. The highest peak of LOD score marks the most likely location (Strachan & Read, 2004).

Genotyping can also be made using the SNP arrays, for example Affimetrix 10,000- or 50,000-SNP arrays. Successful application of this technique was described by Uppal and co-authors recently (Uppal et al., 2008). Comprehensive

genome scans involving many thousands of SNP assays requires significant amount of genomic DNA from each sample. The whole-genome amplification can be performed before the analysis to solve this problem (Barker et al., 2004).

Ideally, the candidate region should be as narrow as possible. This largely depends on the number of meioses available to study. With 100 informative meioses the region can be around 1 Mb. In our case the number of meioses will be much smaller, and therefore the candidate region is likely to be bigger. Good example of difficulties that are likely to be encountered is described by Kantarci and co-authors (Kantarci et al., 2007). The authors were trying to identify the gene defect leading to the development of Donnai-Barrow syndrome, which is another rare autosomal recessive disorder. The initial mapping with the use of Affimetrix 10K SNP arrays helped to locate a large (21 Mb) candidate region that was narrowed to 18 Mb with the use of different set of markers (microsatellite). This interval contained 51 known genes that needed to be analysed in order to identify the defective gene.

TASK 2

Identification of a candidate gene/s.

Positional information obtained on the previous step can reduce the number of candidates to approximately 30 genes located in the candidate region (Strachan & Read, 2004). Identification of a particular gene responsible for the disorder involves the following steps:

i) obtaining the clones of all the DNA of the candidate region;

ii) identification of all genes in the region;

iii) prioritizing them for mutation screening;

iv) testing the candidate genes for mutations in affected people.

The contig of clones (an arrangement of overlapping cloned DNA sequences) across the candidate region can be downloaded from the human genome sequence database. The most useful resources are: ENSEMBL - European Molecular Biology Organization gene browser (www.ensembl.org), gene browser of the University of California at Santa Cruz (www.genome.ucsc.edu), and National Centre of Biotechnology Information (www.ncbi.nlm.nih.gov). After establishing the contig, it is necessary to catalogue all the genes within it. A genome browser is used to display and analyse all definite and possible genes.

(Full list of internet resources for whole genome comparative analysis and associated tools is provided by Miller et al. (2004)). The database analysis can be supplemented by the transcript mapping. For instance, cDNA libraries can be screened using genomic clones from candidate region as probes. The techniques of exon trapping, zoo blotting and CpG island identification can help in identifying the genes as well (Sudbery, 2002; Strachan & Read, 2004).

The rationale for zoo blotting is that only the protein-coding sequences are likely to be conserved between humans and other mammals, while non-coding sequences are not. Therefore, only the DNA from exons will cross-hybridise with DNA of other mammals.

The cross-species DNA sequence comparison is also the primary method for identification of functional non-coding elements in the genome (Prabhakar et al., 2006). It is important to know the location of these elements in the candidate region since the mutations in the non-coding elements (such as promoters for mRNA synthesis) that are involved in the gene expression regulation are one of the potential reasons for the development of disease.

CpG island are often present in the beginning of the genes. The CpG sequences can be revealed with the help of restrictases such as *HpaII*that cuts the unmethylated CCGG sites.

Expression pattern of good candidate gene should be consistent with the disease phenotype. Candidate has to be expressed in the place where the pathology is seen. This will be difficult to check in our particular situation because the affected individuals are not willing to give the biopsy.

Finding out what is the function of the candidate gene might help a lot in deciding if this is a good candidate. For example, ion channels are involved in the transmission of nerve signals and therefore can be involved in the sensation of pain experienced by affected individuals. The presence of ion channel protein in the candidate region can suggest its involvement in the development of disease, but this will remain hypothetical until positively proven.

Unknown gene in the candidate region can be homologous to the known one. The mutations in that known gene may cause similar conditions.

It is very useful to see what is known about the homologous genes, similar to those found in the candidate region, in other species (mice, zebrafish, drosophila, etc.). Patterns of expression and phenotypes of mutations might be well studied on model organisms already. Homologies with mice provide particularly useful clues.

After identifying the candidate genes they have to be confirmed. Mutation screening is one of the most popular approaches (Strachan & Read, 2004). The presence of the same mutation in several affected individuals and its absence in

healthy unrelated people is a good confirmation that the correct gene was chosen. Additional evidences will be needed for formal proof, however.

Mutations are not necessarily pathological, they can be some rare neutral variants without major effects on gene expression. Also, mutations may affect not the coding parts of the gene, but some regulatory elements or promoters. In this case the expression of gene can be affected in the absence of mutations in the protein product of the gene.

After confirming the candidate gene, it is necessary to study its normal function.

TASK 3

Confirming that a particular mutation affects the gene function

The mutation can lead to a number of consequences - from slight changes in the conformation of the protein product (and therefore changing the rate of the reaction it catalyses) to the production of truncated or highly degradable protein that is not functional at all. Comparative studies of the mutated and normal proteins would help in establishing whether the mutation affects the normal function of the protein. To do so the sequence of abnormal protein should be obtained. By using the GeneSeeker browser (www.cmbi.ru.nl/geneseeker) and Disease Gene Prediction (DGP) (www.cgg.ebi.ac.uk/services/dgp) the cDNA and then the polypeptides corresponding to the gene sequences can be derived. The result depends on predicting the coding sequence which is not fully reliable.

Most of the RNA-encoded proteins have already been discovered by translating cDNA sequences and finding homologies with genomic sequences. The sequence data are available from international databases. The product of our gene might have been discovered and sequenced already. If this is the case, the cDNA of this gene have been cloned, and the actual clones are probably available commercially.

Calcium-specific ion channels are very well studied proteins (Yamakage & Namiki, 2002). The genes for many of them are mapped and sequenced. It is known that they contain several conservative motifs that are observed throughout the whole protein family. Mutations in these motifs are very likely to disrupt the protein function. A number of channelopathies has been recently discovered that are caused by mutations in ion channel proteins (Camerino et al., 2008).

Both mutated and normal gene can be cloned, expressed and translated *in vitro*, generating the protein products. Comparing the activity of these proteins may confirm that mutation affects normal function of the gene.

The product of defective gene can contain several mutations. It is not uncommon when the disorder can be caused by different mutations (heterogeneous disorders). For example, the Aicardi-Goutieres syndrome, an autosomal recessive neurological disorder, can be caused by mutations in the genes encoding any one of the three subunits of human ribonuclease H2 (Crow et al., 2006). (The same syndrome can be caused by mutations in the different chromosomal locus in the gene encoding exonuclease TREX1 (Ali et al., 2006; Rice et al., 2009)).

To find out which of the mutations is/are affecting the protein function, the experiments with partial restoration of normal sequence can be done. Site-directed mutagenesis can generate a set of mutants containing one mutation only. The structure and activity of their protein products can be studied to find which mutation/s affect/s the gene's normal function.

One of the ways to confirm that mutation affects the gene function is to construct a transgenic mouse model with the disease allele introduced into the animals with the knockout normal gene. The models are expected to show similarities to humans with the disease. Relevant mutation may be already known in mice, and therefore checking the databases can help.

Conclusion.

Due to the recent progress in genotyping technologies, the finding of linkage between a particular disorder and the genes responsible for the corresponding phenotype became more quick and straightforward. However, this remains a very difficult and time and resources-consuming task that can be fulfilled only via collective effort of large groups of scientists. There is no single protocol for solving the problem, since too many variables are involved in the process, and the best and shortest way towards the goal depends on what findings were made on the previous steps. In this essay I have tried to suggest a possible general strategy and particular methods that can be used for finding the gene for ARGH. This, however, doesn't mean that alternative strategies and other methods cannot be employed to achieve the same goal.

REFERENCES:

Ali, M. et al. (2006) A second locus for Aicardi-Goutieres syndrome at chromosome 13q14-21. *J. Med. Genet.* **43**, 444-450.

Barker, D.L. et al. (2004) Two methods of whole-genome amplification enable accurate genotyping across a 2320-SNP linkage panel. *Genome Res.* **14**, 901-907.

Camerino, D.C. et al. (2008) Therapeutic approaches to ion channel diseases. *Adv. Genet.* **64**, 81–145.

Chaib, H. et al. (1996) A gene responsible for sensorineural nonsyndromic recessive deadfess maps to chromosome 2p22-23. *Hum. Molec. Genet.* **5**, 155-158.

Crow, Y.J. et al. (2006) Mutations in genes encoding ribonuclease H2 subunits cause Aicardi-Goutieres syndrome and mimic congenital viral brain infection. *Nature Genet.* **38**, 910-916.

Guilford, P. et al. (1994) A non-syndrome form of neurosensory, recessive deafness maps to the pericentromeric region of chromosome 13q. *Nature Genet.* **6**, 24-28.

Houwen, R.H.J. et al. (1994) Genome screening by searching for shared segments: mapping a gene for benign recurrent intrahepatic cholestasis. *Nature Genet.* **8**, 380-386.

Kantarci, S. et al. (2007) Mutations in LRP2, which encodes the multiligand receptor megalin, cause Donnai-Barrow and facio-oculo-acoustico-renal syndromes. *Nature Genet.* **39**, 957-959.

Le Merrer, M. et al. (1994) A gene for achondroplasia-hypochondroplasia maps to chromosome 4p. *Nature Genet.* **6**, 318-321.

Miller, W. et al. (2004) Comparative genomics. *Annu. Rev. Genomics Hum. Genet.* **5**, 15-56.

Morton, N.E. (1955) Sequential test for the detection of linkage. *Am. J. Hum. Genet.* **7**, 277-318.

Prabhakar, S. et al. (2006) Close sequence comparisons are sufficient to identify human *cis*-regulatory elements. *Genome Res.* **16**, 855-863.

Rice, G.I. et al. (2009) Mutations involved in Aicardi-Goutieres syndrome implicate SAMHD1 as regulator of the innate immune response.*Nature Genet.* **41**, 829-833.

Strachan, T. & Read, A.P. (2004) *Human Molecular Genetics*, 3rd Edition, Garland Publishing, London, New York

Sudbery, P. (2002) *Human Molecular Genetics*. 2nd Edition, Pearson Education Ltd., Harlow, England

Uppal, S. et al. (2008) Mutations in 15-hydroxyprostaglandin dehydrogenase cause primary hypertrophic osteoarthropathy. *Nature Genet.***40**, 789-793.

Yamakage, M. and Namiki, A. (2002) Calcium channels - basic aspects of their structure, function and gene encoding; anesthetic action on the channels - a review. *Can. J. Anaesth.* **49**, 151-164.