Identification of genes in a genomic DNA sequence

Mammadova Sevinj Nadir¹,
Huseynova Lala Samaddin²

¹ Department of Medical Biology and Genetics, Azerbaijan Medical University; Republic of Azerbaijan
² Department of Medical Biology and Genetics, Azerbaijan Medical University; Republic of Azerbaijan

Abstract.
Today, newer pyrosequencing methods have drastically cut the cost of sequencing and may eventually allow every person the possibility of personalized genome information. Being able to read how our genes are expressed offers the promise of advanced medical treatments, but it will certainly require considerable work to generate, understand, organize, and apply this massive amount of data to human disease. This study describes the results of the CFTR, MEFV, GALT, PAH, BCKDHB and DBT genes analysis of Azerbaijanian patients. The molecular diagnostics methods using specific primers to identify CFTR, MEFV, GALT, PAH, BCKDHB and DBT genes many mutations. A total 15 different mutations (R261Q, V245V, P281L, R241C, L385L, V399V, E280K, R261X, A434D, R176X, R261Q, V399V, E280K, R261X, A434D, R176X, R241C, L385L, V399V, E280K, R261X, A434D, R176X, R261Q, V399V, E280K, R261X, A434D, R176X, R261Q, V399V, E280K, R261X, A434D, R176X, R261Q, V399V, E280K, R261X, A434D, R176X, R261Q, V399V) and 2 different polymorphisms (Q232Q,V245V) were detected of the PAH gene. 4 different mutations (P325L, H132Q, Q334K, N314D) were detected of the GALT gene, 6 different mutations (Phe508del, G138G, A165A, R202Q, R314X, G474G, G476G, G510D) were found in the CFTR gene, polymorphisms (D102D, G138G, A165A, R202Q, R314X, G474G, G476G, G510D) were detected of the MEFV gene, 2 different mutations were found of the BCKDHB gene and one mutation detected of the DBT gene in patients from Azerbaijani population. To prophylaxis the phenylketonuria, galactosemia, maple syrup urine disease, cystic fibrosis disease and Famillian Mediterranian Fever it is recommended to screen genetically newborns, to consult medical-genetically risky families, and to carry out prenatal diagnostics during pregnancies for those families.

Keywords:
DNA
gene
cases
amino acid
sequence

This work is distributed under the terms of the Creative Commons Attribution-ShareAlike 4.0 International License (https://creativecommons.org/licenses/by-sa/4.0/).
Sequencing DNA means determining the order of the four chemical building blocks-called "bases" - that make up the DNA molecule. The sequence tells scientists the kind of genetic information that is carried in a particular DNA segment. For example, scientists can use sequence information to determine which stretches of DNA contain genes and which stretches carry regulatory instructions, turning genes on or off. In addition, and importantly, sequence data can highlight changes in a gene that may cause disease [1].

In the DNA double helix, the four chemical bases always bond with the same partner to form "base pairs." Adenine (A) always pairs with thymine (T); cytosine (C) always pairs with guanine (G). This pairing is the basis for the mechanism by which DNA molecules are copied when cells divide, and the pairing also underlies the methods by which most DNA sequencing experiments are done. The human genome contains about 3 billion base pairs that spell out the instructions for making and maintaining a human being [2,3].

One new sequencing technology involves watching DNA polymerase molecules as they copy DNA - the same molecules that make new copies of DNA in our cells - with a very fast movie camera and microscope, and incorporating different colors of bright dyes, one each for the letters A, T, C and G. This method provides different and very valuable information than what's provided by the instrument systems that are in most common use [4].

Another new technology in development entails the use of nanopores to sequence DNA. Nanopore-based DNA sequencing involves threading single DNA strands through extremely tiny pores in a membrane. DNA bases are read one at a time as they squeeze through the nanopore. The bases are identified by measuring differences in their effect on ions and electrical current flowing through the pore. Using nanopores to sequence DNA offers many potential advantages over current methods. The goal is for sequencing to cost less and be done faster. Unlike sequencing methods currently in use, nanopore DNA sequencing means researchers can study the same molecule over and over again [5,7].

Empirical approaches, which historically came first, attempt to derive the characteristic frequencies of different
amino acid substitutions from actual alignments of homologous protein families. In other words, these approaches strive to determine the actual likelihood of each substitution occurring during evolution. Obviously, the outcome of such efforts critically depends on the quantity and quality of the available alignments, and even now, any alignment database is far from being complete or perfectly correct. Furthermore, simple counting of different types of substitutions will not suffice if alignments of distantly related proteins are included because, in many cases, multiple substitutions might have occurred in the same position. Ideally, one should construct the phylogenetic tree for each family, infer the ancestral sequence for each internal node, and then count the substitutions exactly. This is not practicable in most cases, and various shortcuts need to be taken.

Several solutions to these problems have been proposed, each resulting in a different set of substitution scores [8, 9].

The first substitution matrix, constructed by Dayhoff and Eck in 1968, was based on an alignment of closely related proteins, so that the ancestral sequence could be deduced and all the amino acid replacements could be considered occurring just once. This model was then extrapolated to account for more distant relationships. PAM (Accepted Point Mutation) is a unit of evolutionary divergence of protein sequences, corresponding to one amino acid change per 100 residues. Thus, for example, the PAM30 matrix is supposed to apply to proteins that differ, on average, by 0.3 change per aligned residue, whereas PAM250 should reflect evolution of sequences with an average of 2.5 substitutions per position. Accordingly, the former matrix should be employed for constructing alignments of closely related sequences, whereas the latter is useful in database searches aimed at detection of distant relationships. Using an approach similar to that of Dayhoff, combined with rapid algorithms for protein sequence clustering and alignment, Jones, Taylor, and Thornton produced the series of the so-called JTT matrices, which are essentially an update of the PAMs [10-12].

This extrapolation may not be fully valid because the underlying evolutionary model might not be adequate, and the
trends that determine sequence divergence of closely related sequences might not apply to the evolution at larger distances [13].

In 1992, Steven and Jorja Henikoff developed a different series of substitution matrices using conserved ungapped alignments of related proteins from the BLOCKS database. The use of these alignments offered three important advantages over the alignments used for constructing the matrices. First, the blocks collection obviously included a much larger number and, more importantly, a much greater diversity of protein families than the collection that was available to Dayhoff and coworkers in the 1970's. Second, coming from rather distantly related proteins, blocks alignments better reflected the amino acid changes that occur over large phylogenetic distances and thus produced substitution scores that represented sequence divergence in distant homologs directly, rather than through extrapolation. Third, in these distantly related proteins, blocks included only the most confidently aligned regions, which are likely to best represent the prevailing evolutionary trends.

It is remarkable that so far, throughout the 30-plus-year history of amino acid substitution matrices, empirical matrices have consistently outperformed those based on theory, either physico-chemical or evolutionary. This is not to say, of course, that theory is powerless in this field, but to point out that we currently do not have a truly adequate theory to describe protein evolution. Clearly, the last word has not been said on amino acid substitution matrices. A recently reported maximum-likelihood model for substitution frequency estimation has already been claimed to describe individual protein families better than the Dayhoff models. It remains to be seen how this and other new matrices perform in large-scale computational experiments on real databases. AAindex (http://www.genome.ad.jp/dbget/aaindex.html, see 3.6.3) lists 66 different substitution matrices, both ab initio and empirical, and there is no doubt that this list will continue to grow [14-17].

Protein sequence-structure threading is a family of computational approaches that, given a protein sequence, attempt to select, among all known 3D structures, the
structure that is best compatible with this sequence [18]. Metaphorically, the sequence is “threaded” through a variety of structures, and the method determines which one fits better than the others [19].

The underlying idea is already well familiar to us: protein structure is more conserved in evolution than sequence. Insertions and deletions that change the substrate specificity, thermal stability, and other properties of the protein mostly occur in the loop regions without changing the core set of α-helices and β-strands. Therefore, a comparison of the (predicted) secondary structure of the new protein against a library of known 3D structures could potentially identify distant homologs, even in the absence of statistically significant sequence similarity. Generally, threading methods involve calculating residue contact energy for the analyzed sequence superimposed over each structure in the database and ranking the structures by decreasing energy; the structure with minimal energy is the winner [20-22].

Several statistical models to estimate the probability of “native” fold detection have been developed. It has been consistently reported that combining the traditional contact-potential-based threading with the use of sequence profiles and secondary structure alignment leads to a substantially greater success rate of fold recognition than either threading or profile searches or secondary structure comparisons alone [23,24].

Further discussion of threading is beyond the scope of this book; a detailed review of the physical theory behind threading methods has been published recently. However, before ending this brief section with a list of threading software tools, which are available on the web, we must add a cautionary note based on our own research experience. Despite the well-documented success of threading approaches, using several different threading methods in the analysis of a variety of protein families. In contrast, we faced a considerable number of false leads that were associated with apparently statistically significant scores. We are relating this experience not to question the impressive performance of threading methods in fold recognition but to caution the reader that current threading approaches still might not be
robust enough for routine use in large-scale genome analysis [25,26].

Even in case-by-case manual analysis, before gleaning any far-reaching conclusions from the threading results, one has to be aware of the complexities of the approach and its potential pitfalls. It is important to carefully analyze the outputs to make sure that the reported secondary structure similarity is genuine and is not caused, for example, by “fatal attraction” of long (>15 aa) helices in two proteins that are otherwise entirely different. One should be extremely suspicious when the query protein produces high-scoring alignments with proteins known to have different folds [27].

Archaeal and bacterial genes typically comprise uninterrupted stretches of DNA between a start codon (usually ATG, but in a minority of genes, GTG, TTG, or CTG) and a stop codon (TAA, TGA, or TAG; alternative genetic codes of certain bacteria, such as mycoplasmas, have only two stop codons). Rare exceptions to this rule involve important but rare mechanisms, such as programmed frameshifts. There seem to be no strict limits on the length of the genes. Indeed, the gene rpmJ encoding the ribosomal protein L36 is only 111 bp long in most bacteria, whereas the gene for B. subtilis polyketide synthase PksK is 13,343 bp long. In practice, mRNAs shorter than 30 codons are poorly translated, so protein-coding genes in prokaryotes are usually at least 100 bases in length. In prokaryotic genome-sequencing projects, open reading frames (ORFs) shorter than 100 bases are rarely taken into consideration, which does not seem to result in substantial underprediction. In contrast, in multicellular eukaryotes, most genes are interrupted by introns. The mean length of an exon is ~50 codons, but some exons are much shorter; many of the introns are extremely long, resulting in genes occupying up to several megabases of genomic DNA. This makes prediction of eukaryotic genes a far more complex (and still unsolved) problem than prediction of prokaryotic genes [28-29].

We hope that after working through this study, interested readers will be encouraged to continue their education in methods of sequence analysis using more specialized texts.

Researchers now are able to compare large stretches of
DNA - 1 million bases or more - from different individuals quickly and cheaply. Such comparisons can yield an enormous amount of information about the role of inheritance in susceptibility to disease and in response to environmental influences. In addition, the ability to sequence the genome more rapidly and cost-effectively creates vast potential for diagnostics and therapies [30].

Although routine DNA sequencing in the doctor's office is still many years away, some large medical centers have begun to use sequencing to detect and treat some diseases. In cancer, for example, physicians are increasingly able to use sequence data to identify the particular type of cancer a patient has. This enables the physician to make better choices for treatments.

References:

[12] Huseynova L.S. Mutation Analysis of The Phenylalanine Hydroxylase


