A non-statistical approach to protein mutational variability

Jacek Leluk *

Institute of Biochemistry and Molecular Biology, University of Wrocław, Tamka 2, 50-137 Wrocław, Poland

Received 16 August 1999; received in revised form 12 February 2000; accepted 17 February 2000

Abstract

The non-statistical, non-Markovian model of protein mutational variability is described. There are presented the essential features of the algorithm of genetic semihomology and some examples of its application. The results of genetic semihomology approach are compared with the results obtained by using statistical algorithms and matrices which are assumed in widely used programs such as ClustalW, FASTA, MultAlin and BLAST. The aim of the new algorithm elaboration is to improve the accuracy of the results of protein sequence comparison, avoid the wrong assumptions and misinterpretation of the results, and increase the amount of information available from such study. © 2000 Elsevier Science Ireland Ltd. All rights reserved.

Keywords: Protein homology; Mutations; Variability; Genetic semihomology

1. Introduction

There are over 400 amino acid indices and 42 mutation matrices described so far (Tomii and Kanehisa, 1996). Actually they are based on much fewer algorithms, most of which are subsequent modifications of the original ones. Some of them dominate in their usefulness and popularity, like the mutation data matrices derived from Dayhoff matrix (Dayhoff and Eck, 1968; Dayhoff et al., 1979) and PAM and BLOSUM indices. The algorithm of genetic semihomology (Leluk, 1998) differs from the others by its non-statistical approach and lack of scoring scale. Widely used statistical scoring matrices are very helpful for comparative studies of similarity and variability among different protein families. However, the values of such scoring may bring limited and confusing information. They are also dependent on the database used. For that reason the same algorithm can lead to different scoring matrix contingently upon the type and number of protein sequences used as database (Tomii and Kanehisa, 1996). The statistical approach is useful and necessary at certain steps of algorithm development and available knowledge. Such approach, provided with scoring matrices, enables elaborations on the regularities occurring in protein variability, amino acid replacements and relationship to their structural and physico-chemical character. However, they cannot support with explanation about the mechanism of variability or conservativity. It is hard to use them to predict new possible se-
Table 1
Multiple alignment of the gap containing fragments of seven chicken ovoinhibitor domains obtained with Markovian and non-Markovian methods

<table>
<thead>
<tr>
<th>ClustalW</th>
<th>MultAlin</th>
<th>Genetic semihomology</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Blosum62</strong></td>
<td><strong>Blosum62</strong></td>
<td><strong>Blosum62</strong></td>
</tr>
<tr>
<td>I VNCSLYASIGKDGSVWAC</td>
<td>I VNCSLYASIGKDGSVWAC</td>
<td>I VNCSLYASIGKDGSVWAC</td>
</tr>
<tr>
<td>II IDCSPYLVQVLDGNMVAAC</td>
<td>II IDCSPYLVQVLDGNMVAAC</td>
<td>II IDCSPYLVQVLDGNMVAAC</td>
</tr>
<tr>
<td>III VDCSKYPTVSDKGRTLVAAC</td>
<td>III VDCSKYPTVSDKGRTLVAAC</td>
<td>III VDCSKYPTVSDKGRTLVAAC</td>
</tr>
<tr>
<td>IV IDCQVPTKTTGKLKLLVC</td>
<td>IV IDCQVPTKTTGKLKLLVC</td>
<td>IV IDCQVPTKTTGKLKLLVC</td>
</tr>
<tr>
<td>V LDCTQYLSNT-QNEAITAC</td>
<td>V LDCTQYLSNT-QNEAITAC</td>
<td>V LDCTQYLSNT-QNEAITAC</td>
</tr>
<tr>
<td>VI LDCSYKTSTLKDGRQVAC</td>
<td>VI LDCSYKTSTLKDGRQVAC</td>
<td>VI LDCSYKTSTLKDGRQVAC</td>
</tr>
<tr>
<td>VII EHCREFPQKVS--------PIC</td>
<td>VII EHCREFPQKVS--------PIC</td>
<td>VII EHCREFPQKVS--------PIC</td>
</tr>
<tr>
<td>.:*</td>
<td>.:*</td>
<td>.:*</td>
</tr>
<tr>
<td><strong>Dayhoff (PAM)</strong></td>
<td><strong>Dayhoff (PAM)</strong></td>
<td><strong>Dayhoff (PAM)</strong></td>
</tr>
<tr>
<td>I VNCSLYASIGKDGSVWAC</td>
<td>I VNCSLYASIGKDGSVWAC</td>
<td>I VNCSLYASIGKDGSVWAC</td>
</tr>
<tr>
<td>II IDCSPYLVQVLDGNMVAAC</td>
<td>II IDCSPYLVQVLDGNMVAAC</td>
<td>II IDCSPYLVQVLDGNMVAAC</td>
</tr>
<tr>
<td>III VDCSKYPTVSDKGRTLVAAC</td>
<td>III VDCSKYPTVSDKGRTLVAAC</td>
<td>III VDCSKYPTVSDKGRTLVAAC</td>
</tr>
<tr>
<td>IV IDCQVPTKTTGKLKLLVC</td>
<td>IV IDCQVPTKTTGKLKLLVC</td>
<td>IV IDCQVPTKTTGKLKLLVC</td>
</tr>
<tr>
<td>V LDCTQYLSNT-QNEAITAC</td>
<td>V LDCTQYLSNT-QNEAITAC</td>
<td>V LDCTQYLSNT-QNEAITAC</td>
</tr>
<tr>
<td>VI LDCSYKTSTLKDGRQVAC</td>
<td>VI LDCSYKTSTLKDGRQVAC</td>
<td>VI LDCSYKTSTLKDGRQVAC</td>
</tr>
<tr>
<td>VII EHCREFPQKVS--------PIC</td>
<td>VII EHCREFPQKVS--------PIC</td>
<td>VII EHCREFPQKVS--------PIC</td>
</tr>
<tr>
<td>.:*</td>
<td>.:*</td>
<td>.:*</td>
</tr>
<tr>
<td><strong>Identity (unitary matrix)</strong></td>
<td><strong>Identity (unitary matrix)</strong></td>
<td><strong>Identity (unitary matrix)</strong></td>
</tr>
<tr>
<td>I VNCSLYASIGKDGSVWAC</td>
<td>I VNCSLYASIGKDGSVWAC</td>
<td>I VNCSLYASIGKDGSVWAC</td>
</tr>
<tr>
<td>II IDCSPYLVQVLDGNMVAAC</td>
<td>II IDCSPYLVQVLDGNMVAAC</td>
<td>II IDCSPYLVQVLDGNMVAAC</td>
</tr>
<tr>
<td>III VDCSKYPTVSDKGRTLVAAC</td>
<td>III VDCSKYPTVSDKGRTLVAAC</td>
<td>III VDCSKYPTVSDKGRTLVAAC</td>
</tr>
<tr>
<td>IV IDCQVPTKTTGKLKLLVC</td>
<td>IV IDCQVPTKTTGKLKLLVC</td>
<td>IV IDCQVPTKTTGKLKLLVC</td>
</tr>
<tr>
<td>V LDCTQYLSNT-QNEAITAC</td>
<td>V LDCTQYLSNT-QNEAITAC</td>
<td>V LDCTQYLSNT-QNEAITAC</td>
</tr>
<tr>
<td>VI LDCSYKTSTLKDGRQVAC</td>
<td>VI LDCSYKTSTLKDGRQVAC</td>
<td>VI LDCSYKTSTLKDGRQVAC</td>
</tr>
<tr>
<td>VII EhcrefPQKVS--------PIC</td>
<td>VII EhcrefPQKVS--------PIC</td>
<td>VII EhcrefPQKVS--------PIC</td>
</tr>
<tr>
<td>.:*</td>
<td>.:*</td>
<td>.:*</td>
</tr>
</tbody>
</table>

quences in a certain protein family, even if the probability of amino acid replacement is well described.

The aim of the new algorithm elaboration is to overcome the basic disadvantages of protein sequence analysis tools and to exclude some basic errors in the assumptions of the existing statistical methods. It is to be able to explain the mechanism and pathway of protein evolution and differentiation, not only limited to the description of the initial and final step of the observed changes. The algorithm of genetic semihomology departs from assuming that amino acid replacement is concordant with the Markovian model and explains why this model cannot be applied for these changes (Leluk, 1998). Another goal of this algorithm is to make it applicable to any group of proteins of any nature, function and location. It can be achieved for two reasons, (1) minimization of basic assumptions limited to the general amino acid/codon translation table and assuming that single point mutation is a principle, most common, mechanism of protein variability; (2) non-statistical approach (a statistical mutation matrix is replaced by three dimensional diagram representing all theoretically possible amino acid replacements). The approach presented in this article is to allow to study the multidomain structure of the proteins (at the primary structure level), prediction of the genetic code of the proteins with unknown gene sequence (useful for genetic probe construction), explaining the mutation mechanism
that took place at a certain fragment, precise location of insertion/deletion sites at non-homologous fragments, confirmation that the low homology of compared proteins is actual or casual, and prediction of the future or not reported mutations for the related proteins.

2. Materials and methods

The amino acid and nucleotide sequences selected for the purpose of this study concerned proteinase inhibitors from squash seeds (Joubert, 1984; Wieczorek et al., 1985; Otlewski et al., 1987; Heitz et al., 1989; Hakateyama et al., 1991; Chen et al., 1992a,b; Matsuo et al., 1992; Nishino et al., 1992; Ling et al., 1993; Hayashi et al., 1994; Lee and Lin, 1995; Hamato et al., 1995; Haldar et al., 1996; Stachowiak et al., 1996), the Bowman–Birk inhibitors (Odani and Ikenaka, 1978; Chen et al., 1992a,b; Baek et al., 1994; McGurl et al., 1995; sequences obtained from NCBI database), 31 sequences from the trypsin family (sequences obtained from SWISS-PROT (Bairoch and Apweiler, 1997; Bairoch and Apweiler, 1999; Apweiler et al., 1997; Bairoch, 1997), and PROSITE (Hofmann et al., 1999; Bucher and Bairoch, 1994) databases), chicken ovoinhibitor (Scott et al., 1987) and chicken ovomucoid (Catterall et al., 1980; Gerlinger et al., 1982; Lai et al., 1982). The Markovian matrices of amino acid replacement were applied by using programs ClustalW (Thompson et al., 1994), MultAlin (Corpet, 1988) and BLAST (Altschul et al., 1990; Gish and States, 1993). The algorithm of genetic semihomology was applied manually or with the program SEMIHOM (Leluk, 1998) (Table 1).

The dot matrix interpretation of the results was achieved with BLAST 2 SEQUENCES (Tatusova and Madden, 1999) and SEMIHOM (Leluk, 1998).

3. Results and discussion

3.1. General characteristics of non-statistical algorithm of genetic semihomology

The important feature of the algorithm of genetic semihomology concerns the minimal initial assumptions. It assumes the general codon-to-amino acid translation table, and the single point mutation as the basic and most common mechanism of protein differentiation (Fig. 1) (Leluk, 1998). This approach is not limited to single point mutation mechanism only. It may be used to localize those fragments, which undergo other kind of mutation.

The scoring scale can be assumed for genetic semihomology algorithm, but only for rough comparison studies. The identity of compared positions has score of 3, the semihomology of type I (residues encoded by triplets which differ by one nucleotide of the same type, e.g. both are purines) is scored as 2, and semihomology of type II (residues encoded by triplets which differ by one nucleotide of different type) as 1. In the other words, type I is represented by any single transition and type II by any single transversion in the nucleotide triplet. The non-semihomologous pairs are scored as 0.

The sense of this algorithm is not the scoring the comparison results but explaining the way of amino acid replacement and the mechanism of variability at both, amino acid and genetic code, levels. For that reasons the algorithm of genetic semihomology should be treated separately from other algorithms, especially statistical ones. This interpretation is often consistent with the amino acid replacement indices data (Fig. 2). Theoretical results obtained by algorithm of genetic semihomology should be treated separately from other algorithms, especially statistical ones. This coincidence is very important for correct interpretation of the mechanism of protein variability. The used mutation data matrix scoring is supported by PAM250 indices. The PAM250 values for each amino acid replacement were added. The result of such calculation is consistent with the result of theoretical genetic semihomology analysis. The higher score of the PAM indices’ sum is convergent with the genetic relationship of the studied set of amino acids at individual positions.

The goal of minimization of the initial assumptions in the algorithm of genetic semihomology is to (1) make the algorithm universal for any group
of proteins and (2) avoid errors and misinterpretation of the results. Basically, the algorithm assumes three general establishments. One of them is the codon-amino acid translation table, which is undoubtedly considered as true.

The second is the statement that single nucleotide replacement is the most often and easiest mechanism of mutational variability. The last general establishment comprises the three-dimensional diagram of amino-acid replacements by single point mutation (Fig. 1).

A single nucleotide replacement concerns two types of replacements—within the same group (purine to purine or pyrimidine to pyrimidine) and between two groups (purine to pyrimidine or conversely). Thus, the first case is described here as type I (transition) and the second as type II (transversion). It seems, that the replacement of type I is spatially more possible than the other. However, there are no distinct reports that such dependence exists. Another question is the position-dependence of replacement within the codon and its further consequence. The third nucleotide of the codon is known as the most tolerant and most flexible encoding unit. Most amino acids are encoded by first two nucleotides while the third position is optional. Therefore, in algorithm of genetic semihomology mutation at this position is considered separately and is defined as type III, which is independent of type I and II.

Amino acid residues as the products of genetic code translation are not equal units, when we consider both nucleotide and protein levels simultaneously. They are encoded by different number of codons from 1 to 6. That implies the difference in probability of their replacement by single point mutation. Such differentiation causes, that amino acids themselves cannot be treated uniformly in terms of mutational variability. It is important which exactly codon encodes a certain amino acid for its further transformation to another one. For example arginine encoded by AGG can be replaced by methionine (ATG) much easier than arginine encoded by CGY (Fig. 1). On the other hand, the probability of ArgCGY replacement by Cys (TGY) is higher than it is for ArgAGG. This leads to conclusion that the information of the amino acid residue evolutionary preceding the residue at a certain position is still present in its genetic code if the mechanism of replacement was based on point mutation. The probability of further replacement is also dependent on its codon. That reveals inconsistency of amino acid mutational substitution with the Markovian model which establishes all exchange probabilities as independent on the previous history (George et al., 1990). Markovian model however is assumed in most statistical matrices of protein mutational variability. This could be true only if each amino acid was encoded by one codon.

The algorithms assuming statistical mutation data matrix and Markovian model omit the process of cryptic mutations. These are the mutations, which considerably change the codon.

---

Fig. 1. Semihomologous relationships among proteinaceous amino acids. The codons of residues along each axis differ by only one nucleotide (Leluk, 1998). The diagram setting shows the codon changes at first (axis 1), second (axis 2) or third (axis 3) position in order A → G → C → U. This diagram forms the basis of the genetic semihomology algorithm instead of a matrix of statistical parameters or replacement indices, which support statistical algorithms.
sequence without changing the amino acid. Such phenomenon refers especially to residues encoded by up to six codons (Arg, Leu and Ser). Cryptic mutations enable code differentiation not restricted by conditions of environmental selection. They enlarge the spectrum of further possible replacements. Generally, most mutations of the third nucleotide of the codon are cryptic. For arginine, leucine and serine cryptic mutations at first or second positions are also allowed. Thus, they can be considered as intermediates between other two residues, exchange of which require more mutations. These three amino acids often occur at the positions, which are very variable and are occupied by six or more residues within homologous proteins (Figs. 3 and 4).

It is admitted that from evolutionary point of view the genetic code and amino acid ‘language’ have evolved simultaneously. They also act with strict coherence with each other. Therefore, in analysis of protein differentiation and variability, both of them should be considered as well. The approach of Siemion and Stefanowicz (1992) presents the periodical relation between chemical reactivity of amino acids and the primary structure of their genetic code. It also refers to one point mutation mechanism which is the assumption for amino acid reactivity properties and Chou–Fasman conformational parameters relationship (Siemion, 1994, 1995). Although, those elaborations concern the reactivity and secondary structure parameters, the theoretical strategy of approach is consistent.
3.2. Examples of application of genetic semihomology algorithm

3.2.1. Gap location

The commonly used alignment programs such as ClustalW, MultAlin or FASTA based on PAM and BLOSUM matrices generate gaps and locate them in the manner that aligns remaining fragments with each other. The accuracy of such strategy is good for the fragments of high homology. With the lower similarity degree of aligned sequences the accuracy also decreases. The strategy of these tools tends to combine the deleted positions into continuous fragments, which are set at most privileged site for entire alignment. The results are often inconsistent when provided with different comparison matrices and even when the same matrix is applied in different programs. An example of such discordant results are obtained when ClustalW and MultAlin are used for the alignment of chicken ovoinhibitor domains (Scott et al., 1987) (Fig. 5). The gap-containing fragment was thoroughly analyzed for genetic semihomology relationship at both levels — genetic and amino acid. The obtained results suggest discontinuous alignment of deletions with genetically reasonable location of non-homologous residues (Fig. 5). Their location seems to be proper with respect to the probability of one point mutation exchange between aligned residues.

3.2.2. Alignment of non-homologous fragments

For statistical matrices if aligned fragments of related proteins are not homologous, the observed probability of amino acid replacement is taken under consideration. This strategy works properly for protein families for which the mechanism of differentiation is similar. It concerns their general physico-chemical character, reactivity and biological function. But for very different proteins the matrices of replacement likeliness may have

Fig. 3. A simplified planar diagram of genetic relationship between the amino acids. The significance of the third position of the codon is omitted. The dashed curves show the cryptic mutational passages that do not cause change at the amino acid level.

Fig. 4. Frequency of six-codon amino acids as a function of position variability in homologous sequences. Results obtained from aligned trypsins (219 positions occupied by 1150 residues). Note that the serine occurrence at variable positions is more significant than arginine and especially leucine. See the text for the source and calculation method (Section 3.2.6).
not corresponding values. Genetic semihomology shows only the mutational probability of replacement, which is independent on the character of a protein and its biological function. It aligns the positions that are most similar with respect to their known or predicted genetic code. Taking the semihomologous positions into account completes the correct alignment initially obtained for identical positions.

3.2.3. Low, but significant homologies

The multiple alignment of evolutionary distant sequences is sometimes confusing and uncertain. It is hard to estimate if such sequences show actual relationship, or their similarity is casual. The problem is additionally increased by lack of unified methodology for actual homology estimation. The program SEMIHOM equipped with genetic semihomology analysis tool applicable also for proteins of unknown gene sequence is very useful for estimation whether the obtained low homology is actual or not. The double frame analysis (Leluk, 1998) allows to get rid of casual noise resulting from accidental single matches at different settings without erasing the essential ones (Fig. 5).

3.2.4. Internal homology of multidomain sequences

The similarity searching tools available in most protein databases and related servers are designed for fast database scanning for related sequences to the query sequence. The given results concern the optimal alignment and statistical parameters of comparison effect (FASTA, BLAST, ClustalW, MultAlin). The BLAST servers additionally provide with dot matrix interpretation of the results. Although, these tools are powerful for all types of proteins (and nucleic acids), they do not analyze the internal homology of multidomain sequences properly (Fig. 6). The domain alignment and analysis can be obtained by using program SEMIHOM with good dot matrix results, which can be additionally improved by semihomology analysis (Fig. 6). Even very distant domains (low similarity), which are still related to each other are visualized with this program. The results are achieved by comparison of a protein sequence with itself.

3.2.5. Location of fragments which are products of other mechanism than single point mutation

The single point mutation is considered as the most frequent and easiest mechanism of protein differentiation. However, there are other mechanisms such as deletion, insertion, duplication, inversion etc. The use of genetic semihomology algorithm for sequence multiple alignment and further analysis makes possible to identify the fragments which are the consequence of other type of mutation than single transition/transversion. The identification consists in finding the fragments where aligned residues are not semihomologous.

Fig. 5. Sequence comparison of the fragments of two Bowman-Birk inhibitors. Semihomology application and casual noise reduction for dot matrix results of low homology sequence comparison. The double frame setting shows homologous and semihomologous positions only for at least four identities along 12-unit string.
Fig. 6. Internal domain homology of chicken ovoinhibitor (7 domains) and ovomucoid (3 domains) with BLAST 2 SEQUENCES and SEMIHOM.

Fig. 7. Application of genetic semihomology algorithm for identification the fragments of possible different mechanism than single point mutation. The listing shows the amino acid variability at corresponding positions within 7 domains of chicken ovoinhibitor. The non-semihomologous residues (differing by more than one nucleotide in their codon) are separated by square brackets. The accumulation of these positions (in shadows) suggests different mechanism of differentiation from single transition/transversion.

3.2.6. Cryptic mutations

There are mutations that change composition of the gene without effect on amino acid sequence. Such cryptic mutations concern the
codons for amino acids that can be encoded by two or more triplets. These mutations seemingly do not have any effect at the protein level. However, they are important mechanism to the very variable sites in the manner that they increase the degree of variability available by single point mutation. For example a Ser\textsubscript{TCA} can be converted by single point mutation to Leu, Pro, Ala and Thr (Fig. 1). A cryptic mutation of Ser\textsubscript{TCA} to Ser\textsubscript{AGT} extends the spectrum of its conversion to Asn, Ile, Arg, Gly and Cys by single step mutation. The role of cryptic mutations is especially important for amino acids encoded by up to six codons (Ser, Arg and Leu). Among them serine is particularly interesting since its codons are the most differentiated (the genetic code matrix scores for them vary from 0 to +2). In fact serine participation in the most variable positions is outstanding as it is shown on the example of trypsins (Fig. 4). The sequences of 31 trypsins from different organisms were aligned. There were compared 219 positions (out of 247 corresponding to anionic precursor of bovine trypsinogen, SWISS-PROT accession number Q29463) occupied by 1150 residues. The study did not include the positions which were not represented in more than 70% of compared sequences (positions that were the result of insertion/deletion). The percentage of occurrence of the individual residues at a variable position was calculated in the following way:

$$p_j = \frac{n_j}{j} \times 100$$

where $p$ is the percentage of occurrence of an individual residue in positions occupied by $j$ different residues; $n_j$ the number of positions where an individual residue appears among $j$ occurring residues; and $j$ is the number of positions where $j$ different residues occur.

The planar diagram of genetic semihomology relationship between amino acids, reduced to the first two positions of the codon, shows the cryptic passages for six-codon amino acids (Fig. 3).

### 3.2.7. Genetic code prediction

Originally the algorithm of genetic semihomology has been designed for studies on proteins whose gene sequence is unknown. Similarly to PAM or BLOSUM methods the nucleotide sequence of the gene is not required (although, it is very helpful if we know it). One of the possible options of the genetic semihomology analysis is the prediction of possible codons for particular amino acids aligned. The method consists in comparison of all possible codons of the amino acids occurring at the corresponding position of the

![Table]

<table>
<thead>
<tr>
<th>1</th>
<th>4</th>
<th>5</th>
<th>7</th>
<th>8</th>
<th>12</th>
<th>13</th>
<th>15</th>
<th>22</th>
<th>33</th>
<th>37</th>
<th>49</th>
<th>50</th>
<th>59</th>
<th>61</th>
<th>62</th>
<th>63</th>
<th>66</th>
</tr>
</thead>
<tbody>
<tr>
<td>II</td>
<td>I</td>
<td>S</td>
<td>P</td>
<td>L</td>
<td>Q</td>
<td>R</td>
<td>D</td>
<td>N</td>
<td>R</td>
<td>F</td>
<td>N</td>
<td>H</td>
<td>T</td>
<td>E</td>
<td>K</td>
<td>L</td>
<td>E</td>
</tr>
<tr>
<td>III</td>
<td>V</td>
<td>S</td>
<td>K</td>
<td>P</td>
<td>S</td>
<td>K</td>
<td>D</td>
<td>R</td>
<td>R</td>
<td>F</td>
<td>N</td>
<td>R</td>
<td>T</td>
<td>K</td>
<td>R</td>
<td>Q</td>
<td>E</td>
</tr>
<tr>
<td>IV</td>
<td>I</td>
<td>D</td>
<td>Q</td>
<td>P</td>
<td>T</td>
<td>T</td>
<td>G</td>
<td>K</td>
<td>R</td>
<td>F</td>
<td>N</td>
<td>G</td>
<td>T</td>
<td>R</td>
<td>K</td>
<td>E</td>
<td>E</td>
</tr>
<tr>
<td>V</td>
<td>L</td>
<td>T</td>
<td>Q</td>
<td>L</td>
<td>S</td>
<td>Q</td>
<td>N</td>
<td>E</td>
<td>F</td>
<td>V</td>
<td>N</td>
<td>G</td>
<td>T</td>
<td>R</td>
<td>R</td>
<td>E</td>
<td>E</td>
</tr>
<tr>
<td>VI</td>
<td>L</td>
<td>S</td>
<td>K</td>
<td>K</td>
<td>T</td>
<td>K</td>
<td>D</td>
<td>R</td>
<td>M</td>
<td>V</td>
<td>S</td>
<td>R</td>
<td>T</td>
<td>R</td>
<td>E</td>
<td>E</td>
<td>D</td>
</tr>
<tr>
<td>VII</td>
<td>E</td>
<td>R</td>
<td>E</td>
<td>Q</td>
<td>K</td>
<td>-</td>
<td>-</td>
<td>M</td>
<td>V</td>
<td>N</td>
<td>N</td>
<td>R</td>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Prediction accuracy = 69.7%**

![Fig. 8](image-url) Genetic semihomology prediction of genetic code for chicken ovoinhibitor domains. The predicted gene sequence is compared with the known DNA sequence. Only those positions where prediction reduces possible codons are considered. The predicted codons that are consistent with those present in the ovoinhibitor gene are shadowed.
aligned sequences and choosing the codons which are most related to each other, as the most likely. Usually they can be converted to each other by single transition/transversion process. The prediction method gives promising results (Fig. 8). This option may be used to design sequence fragments for the unknown genes of the proteins. It may be helpful in searching the genes by using hybridization methods.

The simplified version of the program SEMI-HOM with dot matrix graphic interpretation for comparative and genetic semihomology studies of protein sequences is available to anyone upon request to the author (lulu@bf.uni.wroc.pl).

References


