Acta Biochimica Polonica

Vol. 48 No. 1/2001

21–33

QUARTERLY

Application of genetic semihomology algorithm to theoretical studies on various protein families★

Jacek Leluk✉, Beata Hanus-Lorenz and Aleksander F. Sikorski

Institute of Biochemistry and Molecular Biology, University of Wrocław, Wrocław, Poland

Received: 29 December, 2000; revised: 31 January, 2001; accepted: 5 March, 2001

Key words: genetic semihomology, multiple alignment, sequence consensus, cryptic mutations

Several protein families of different nature were studied for genetic relationship, correct alignment at non-homologous fragments, optimal sequence consensus construction, and confirmation of their actual relevance. A comparison of the genetic semihomology approach with statistical approaches indicates a high accuracy and cognition significance of the former. This is particularly pronounced in the study of related proteins that show a low degree of homology. The sequence multiple alignments were verified and corrected with respect to the questionable, non-homologous fragments. The verified alignments were the basis for consensus sequence formation. The frequency of six-codon amino acids occurrence *versus* position variability was studied and their possible role in amino acid mutational exchange at variable positions is discussed.

Theoretical comparative studies on proteins and nucleic acids have become powerful and advanced research tools commonly used in biochemistry, molecular biology, genetics, protein modeling and structure/function prediction. The informative and predictive value of such studies has been admitted in both protein and nucleic acid research. There are over 400 amino acid indices and at least 42 mutation matrices described so far (Tomii & Kanehisa, 1996). Actually they are based on much fewer algorithms, most of which are modifications of several original ones. The most current tools are based on principles derived from the Dayhoff matrix (Dayhoff & Eck, 1968; Dayhoff *et al.*, 1979). The indices used for comparative sequence analysis are mainly of BLOSUM or PAM type with different parameters according to the kind of protein sequences analyzed.

Most algorithms and programs use statistical matrices of amino acid replacement. They consider the probability of replacement from the statistical point of view, but do not refer to the biological mechanisms of replacement probability. The matrix indices (e.g., PAM250 used in Mutation Data Matrix) that reflect similarity and/or relationship as well as most methods used for se-

★A preliminary report of this material was presented at the International Conference on "Conformation of Peptides, Proteins and Nulceic Acids, Debrzyno, Poland, 2000.

✉To whom correspondence should be addressed: Jacek Leluk, Institute of Biochemistry and Molecular Biology, University of Wrocław, Tamka 2, 50-137 Wrocław, Poland; phone: (48 71) 375 2611; fax: (48 71) 375 2608; e-mail: lulu@bf.uni.wroc.pl

quence multiple alignment are focused entirely on the statistical calculations of the observed changes. The probability of amino acid replacement based on their genetic code is often not considered at all. Also there is no reference to the possible mutation mechanism or type. Many examples of such an approach are available within the tools accompanying the protein and/or genomic databases, like the Swiss-Prot Expert Protein Analysis System (ExPASy), the European Molecular Biology Laboratory (EMBL) database, the National Center of Biotechnology Information (NCBI), the European Bioinformatics Institute (EBI) database and GenBank.

The statistical scoring matrices are very useful for studies on similarity and variability within protein families. However, they are dependent on the database used. The same algorithm can lead to different scoring matrices depending on the type and number of protein sequences used as a database (Tomii & Kanehisa, 1996). Its application with identical parameters but in different programs may also give different results of the analysis (Leluk, 2000a). The statistical algorithms cannot give detailed information about the biological mechanism of protein variability. For example, their application to predict a new possible sequence within a protein family is quite limited, even if the probability of amino acid replacement among homologous sequences is well described.

The accuracy of the scoring matrix of amino acid replacement is dependent on the proper alignment of the sequences being compared. Therefore the most efficient programs combine both steps — a replacement scoring matrix and the alignment procedure. They are provided with several matrices, which give a possibility to choose the best one for specified purposes and proteins. However, all of them may be defined as statistical ones. The alignment strategy also varies in different programs (residue-to-residue, segment-to-segment, motif search etc.). Usually the alignment tools are well defined from the mathematical and statistical point of view, but they seldom refer to the biological principles of mutational mechanisms among proteins and/or the genetic code. A typical example of such a theoretical approach is the MAST sequence homology search algorithm (Bailey & Gribskov, 1998a) and the program MEME (Bailey & Gribskov, 1997; Bailey & Elkan, 1995; Bailey & Gribskov, 1998b; Grundy *et al.*, 1997). In this program the part concerning match scores, error parameters and cross validation estimators is very expanded, but the basic constituents of the biological molecule are assumed as just one-letter symbols of the string (the biological characteristics and genetic relationship are ignored).

The algorithm of genetic semihomology (Leluk, 1998; 2000a, b) assumes a close relations between amino acids and their addons for analysis of various relationships between proteins belonging to a given family and different protein families. It differs from the others by its non-statistical approach and a lack of scoring scale (the amino acid replacement within the related proteins is not represented by numerical index values, or replacement probability factors, as it is in PAM or BLOSUM matrices). Instead of a scoring matrix it is supported with a three-dimensional diagram including all theoretically possible amino acid replacements by one nucleotide exchange in their codons. More details concerning the algorithm of genetic semihomology, its construction and basic assumptions are described in the article devoted to the algorithm itself (Leluk, 1998). Besides the standard assignments it can be used for the study on cryptic mutations, the mechanisms of variability, prediction of the gene nucleotide sequence, and of new possible protein sequences within the same family. It was successfully applied to the study of long-distance mutations correlation and their effect on protein conformation (Leluk, 2000c). A special advantage concerns the confirmation of true relationship for proteins revealing low homology. The results of the genetic semihomology approach to various protein families are presented in this article.

MATERIALS AND METHODS

Amino-acid sequences of the analyzed protein families were taken from the original articles describing these proteins or directly from the protein databases. The amino-acid sequences selected for this study comprised proteinase inhibitors

from squash seeds (Chen *et al.*, 1992; Haldar *et al.*, 1996; Hamato *et al.*, 1995; Hatakeyama *et al.*, 1991; Hayashi *et al.*, 1994; Heitz *et al.*, 1989; Joubert, 1984; Lee & Lin, 1995; Ling *et al.*, 1993; Matsuo *et al.*, 1992; Nishino *et al.*, 1992; Otlewski *et al.*, 1987; Stachowiak *et al.*, 1996; Wieczorek *et al.*, 1985), Bowman-Birk inhibitors (Odani & Ikenaka, 1978; Chen *et al.*, 1992; Baek *et al.*, 1994; McGurl *et al.*, 1995; sequences obtained from NCBI database), 31 sequences from the trypsin family (sequences obtained from the SWISS-PROT (Bairoch & Apweiler, 1997; Bairoch & Apweiler, 1999; Apweiler *et al.*, 1997; Bairoch, 1997) and PROSITE (Hofmann *et al.*, 1999; Bucher & Bairoch, 1994) databases) and chicken ovoinhibitor (Scott *et al.*, 1987). The amino-acid sequences of human erythrocyte α- and β-spectrin, spectrin-like proteins and sequences similar to eglin C from *Hirudo medicinalis* were obtained from the SWISS-PROT, TREMBL and NCBI databases.

Preliminary sequence similarity searches were performed with the BLAST (Altschul *et al.*, 1990; Gish & States, 1993) and FASTA (Pearson & Lipman, 1988; Pearson, 1990) programs. Multiple sequence alignments and consensus sequences were obtained with several different methods. The statistical algorithms assuming the BLOSUM and PAM matrices were applied with the programs ClustalW (Thompson *et al.*, 1994), MultAlin (Corpet, 1988), BLAST series (Altschul *et al.*, 1990; Gish & States, 1993) and DIALIGN 2.1 (Morgenstern *et al.*, 1996; Morgenstern, 1999). The algorithm of genetic semihomology was applied either with the use of the program SEMIHOM (Leluk, 1998) or without computer support (manual step-by-step analysis of protein sequence relationship).

The dot matrix comparison of two sequences was achieved with SEMIHOM (Leluk, 1998) and verified with BLAST 2 SEQUENCES (Tatusova & Madden, 1999) and Dotlet (Junier & Pagni, 2000).

The frequency of occurrence of six-codon amino acids (Ser, Arg and Leu) as a function of the position variability was calculated and analyzed according to the principles of the genetic semihomology algorithm (Leluk 1998; Leluk, 2000a). The same algorithm supported the study on the cryptic mutation role, occurrence and mechanism within different protein families.

RESULTS AND DISCUSSION

The alignment strategy for proteins revealing high and low homology

The initial step of alignment with the algorithm of genetic semihomology and most other algorithms is the same. At first the sequences are checked for the best alignment that gives the maximum number of identities (without gaps). This alignment selection is the starting point to further adjustment of particular fragments with the use of gaps. For high homologies (60% and more) and low gap contribution the results are concordant regardless of the method used. Thus the multiple alignments for proteinase inhibitors from squash or Bowman-Birk inhibitors (Fig. 1) look almost identical when they are performed with ClustalW (Thompson *et al.*, 1994), MultAlin (Corpet, 1988) or the genetic semihomology algorithm (Leluk, 1998). The differences appear where gap contribution occurs (especially at non-conservative regions) and when proteins reveal low homology to each other (30% or less). For such proteins the application of the same statistical algorithm with the same analysis parameters, but in different programs, may bring different results (Leluk, 2000a). In order to set the actual alignment at those regions, the genetic semihomology algorithm considers the genetic relationship between particular residue pairs. This relationship assumes single point mutation of the codon as the most likely mechanism of change. This approach makes the algorithm different from the statistical approaches, where mainly the statistical frequency of amino acid replacement is considered. The genetic approach to amino acid exchange not only enables the proper alignment of non-identical (but related) fragments, but also gives some information about the evolutionary mechanism. It also allows predicting the other hypothetical residues that may occur at a certain position.

It is obvious that for cysteine-rich proteins cysteine distribution along the chain serves as the

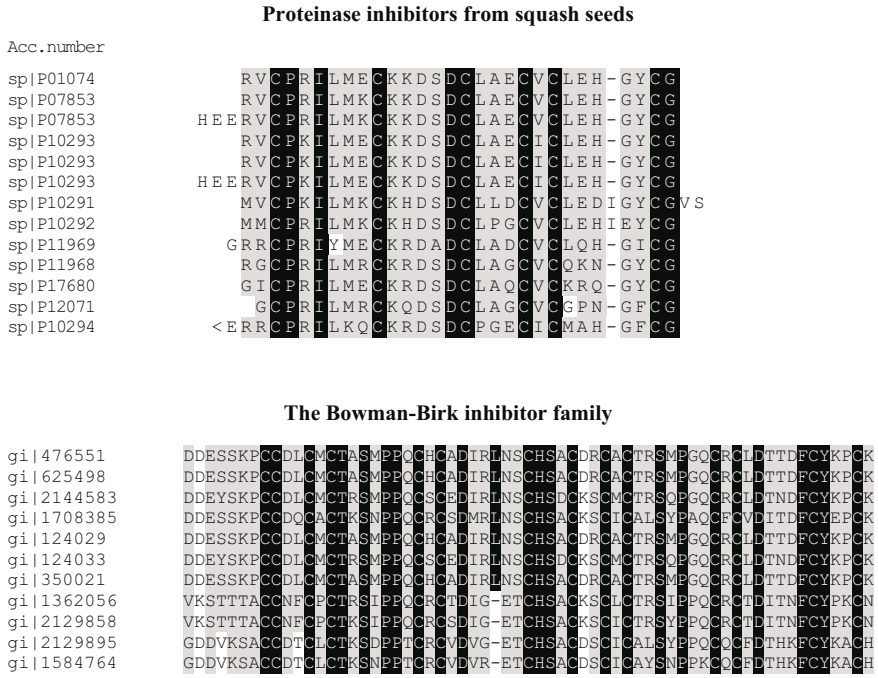

Figure 1. Two examples of multiple alignment of highly homologous, cysteine-rich proteins, verified by the algorithm of genetic semihomology.

The essential conservative residues are marked as white characters on black background. The shadowed characters indicate the semihomology relationship between the residues.

initial data for correct alignment (Fig. 1). Of course, it concerns the cysteines involved in the disulfide bridges formation. This is useful especially for studies on proteinase inhibitors which are divided into several families according to the cysteine (and disulfide bridges) topology. But the

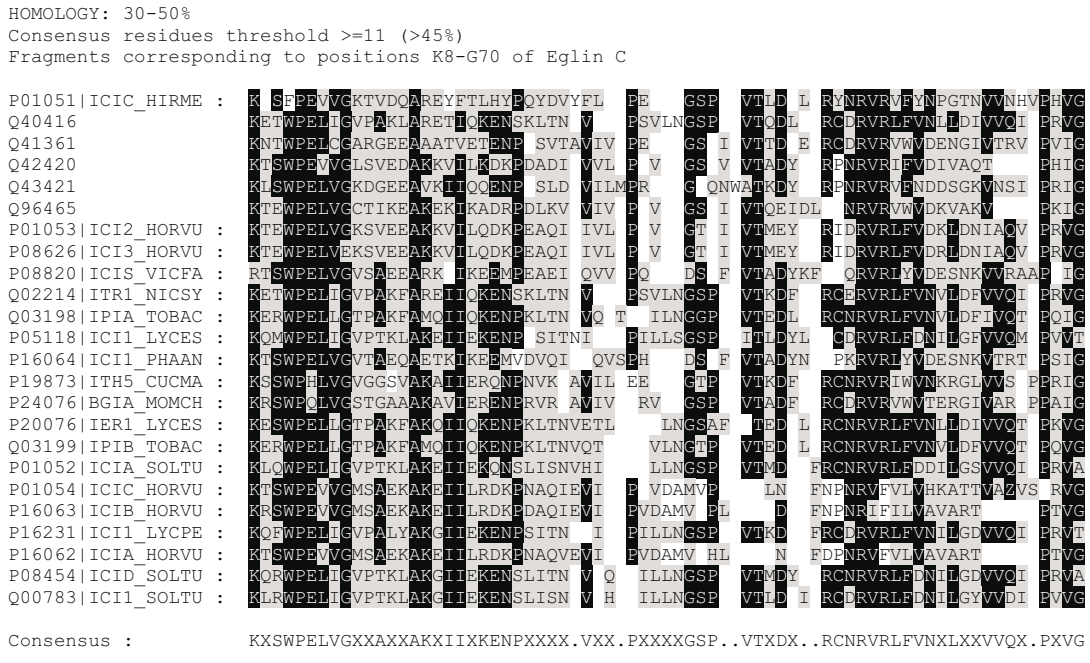

Figure 2. Multiple alignment of proteins homologous to eglin C from *Hirudo medicinalis*.

The results of genetic semihomology analysis. The labels' meaning is the same as in Fig. 1. See text for details.

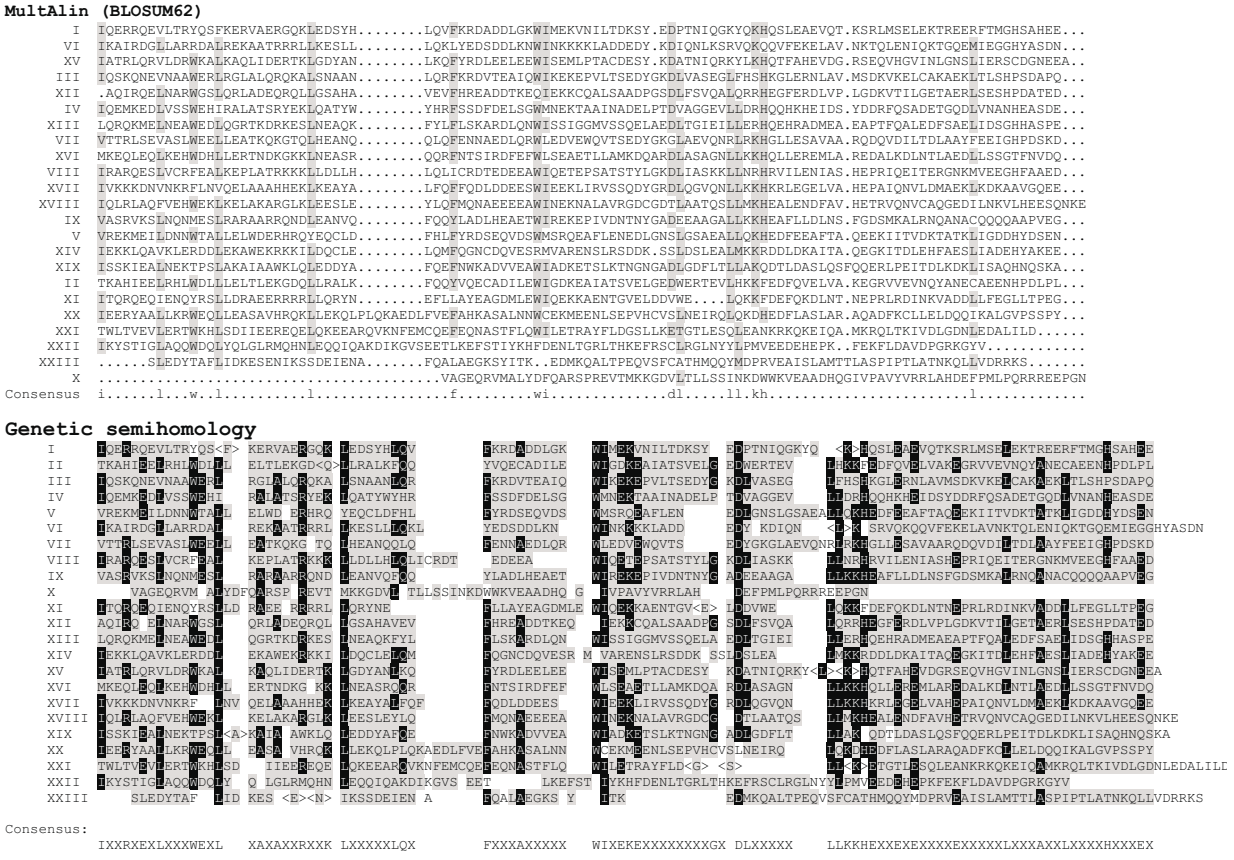

Figure 3. Multiple alignment of human erythrocyte α-spectrin 23 segments achieved with the MultAlin (BLOSUM62) method and the genetic semihomology approach.

The essential conservative residues are marked as white characters on black background. The shadowed characters stand for typical residues for aligned positions (significant conservativity) in the MultAlin alignment and for the semihomology relationship in the genetic semihomology approach. The segment consensus sequence calculated by each method is presented below each multiple alignment.

cysteine parameter cannot be used for such proteins like cysteine free eglin-like proteins (Fig. 2) or membrane bound spectrins (Fig. 3). In such cases it is necessary to recognize and localize the consensus positions which are sufficiently conservative. Depending on the protein group these may be the positions occupied by Trp, Pro, or long chain hydrophobic residues. The next step is to estimate the distance between the consensus residues for all proteins being aligned and to consider the occurrence of gaps. If gaps (deletions) are present, they must be located properly. For that purpose all pairs occurring between consensus residues (the consensus loop) are checked for the genetic relationship. Each pair is checked for the possibility of exchange of one residue to another by a single replacement of one nucleotide of its codon (actual or hypothetical). The residues that genetically do not correlate to any from the compared fragment of the other protein — are considered as inserted, and the gaps are located at the positions corresponding to them. If the fragments being compared are the same in length and do not show genetic semihomology, they can be interpreted as a result of a process different from single point mutation. The genetic semihomology analysis of the proteins revealing low similarity allows to establish whether the similarity is real or casual.

Figure 4 presents the alignments of sequences homologous to the 10th segment of human α-spectrin, obtained with ClustalW and MultAlin, and the alignment consistent with the genetic semihomology algorithm. According to the algorithm chosen the contribution of conservative consensus positions (white on black) and the re-

lated positions (grey) is different. The gap contribution is very limited and the identity is up to 60% – which proves the significant relationship between these sequences. However, the results are different for each analysis. ClustalW gives the least amount of information. The results obtained with MultAlin show higher similarity and the consensus is more complete (although this program uses the same scoring matrix as ClustalW). Significantly more information is obtained with the genetic semihomology approach. The contribution of conservative positions is the fullest – almost all residues at corresponding positions show possible genetic relationship. The concentration of non related residues (white background) suggests a variability mechanism different from single point mutation or more intensive mutational changes at these spots. The gap setting is also different in several cases. The occurence of specified consensus positions over "blank" positions is the highest. A meaningful consensus can also be obtained with the DIALIGN program that is a segment-to-segment approach to multiple sequence alignment (Morgenstern *et al.*,1996; Morgenstern, 1999). However, the informative value of this result is not very high. Not all positions are considered to be aligned, there is one additional gap close to the C-terminus and gap distribution conforms to the other alignments only in general.

Similar comparative studies were done for human erythrocyte α-spectrin repeats (23 domains) (Fig. 3). The homology among them does not exceed 30% (for most of them it is less than 25%). The high appearance of possible genetic relationship may be overestimated in this case, because of the very high position variability. The very variable positions (the ones that accept more than 8 residues) should not be analyzed for semihomologous relationship only because of too many possible ways of codon transformation. However, the genetic semihomology approach identifies many more consensus positions than the other methods do (see the chapter "Construction of a sequence consensus for related proteins").

The eglin-like proteins were aligned in the same manner as spectrin repeats (Fig. 2). The eglin itself is cysteine-free, therefore cysteine contribution in the structure alignment is almost none, but the homology between the sequences is higher than for spectrin repeats (up to 50%). A thorough analysis and verification with the genetic semihomology approach exposed a considerable amount of consensus residues.

Construction of a sequence consensus for related proteins

The aligned sequences of homologous proteins are used to construct a consensus sequence specifying the most conservative residues at the positions significant for a protein family. The important parameter in the consensus construction is the ratio (*r)* of a residue occurrence (*n*) in the aligned position per number of aligned sequences (*N*):

$$r = \frac{n}{N}$$

Usually the value of this ratio is proportional to the homology degree of the aligned proteins. For very coherent high homologies the consensus residues reveal the *r* value of 0.7 to 1.0. For lower homologies (30% or less) the accepted *r* value may be lower, but it should not be less than 0.4. To get a good consensus the ratio should not be lower than 0.5 if there are less than 20 sequences aligned. In questionable cases the accompanying residues at the aligned position may be considered as well. For example, if there is a leucine of the *r* value of about 0.4 and that position is also occupied by isoleucine (similar physicochemically, and semihomologous genetically to leucine) then leucine may be assumed as the consensus residue if the *r* value for both Leu and Ile is significantly high (e.g., 0.7). However, it is better to use a more restrictive design strategy for more correct recognition of the actual homologies.

The sequence similarity search results for the α-spectrin consensus constructed by different methods are shown in Fig. 5. The consensus refers to the general 106 amino-acid repeat of human erythrocyte α-spectrin (the repeats reveal homology less than 30% between each other; Fig. 3). The consensus designed by Sahr and co-workers (Sahr *et al.*, 1990) gives good results

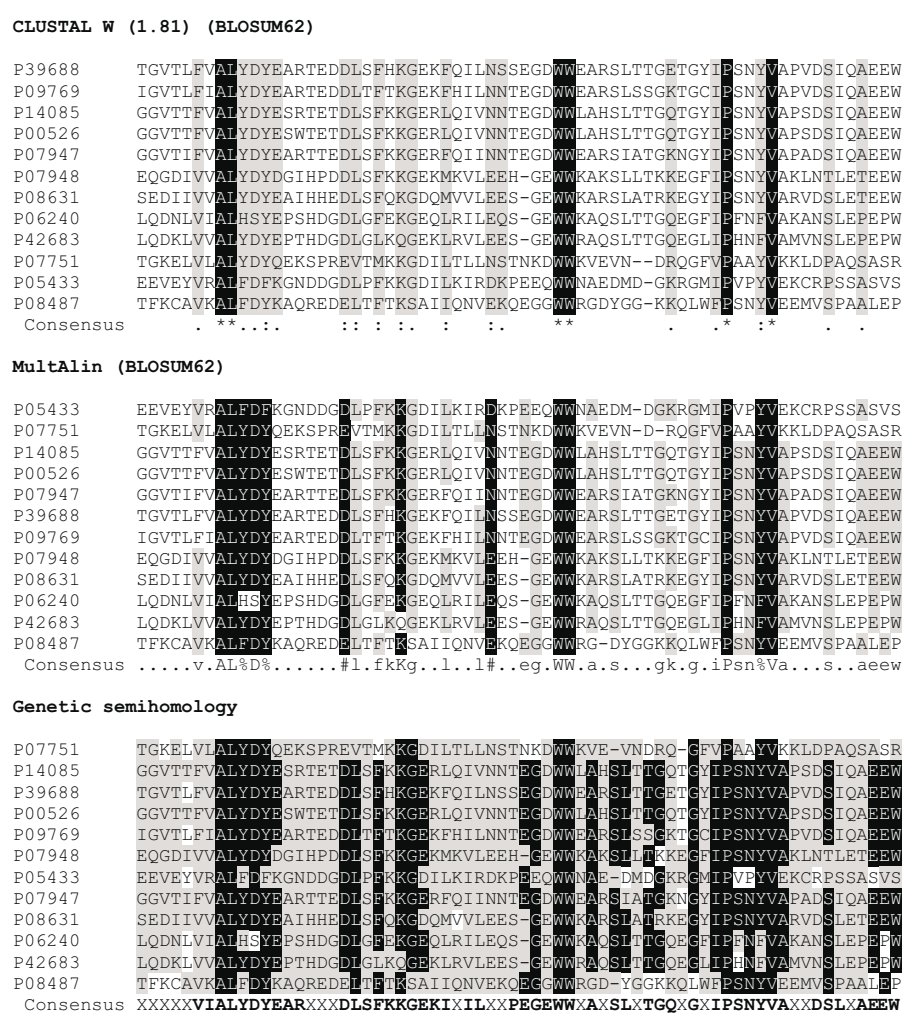

Figure 4. Multiple alignment of proteins homologous to the 10th segment of human erythrocyte α-spectrin.

The results of two BLOSUM approaches (ClustalW and MultAlin) and the genetic semihomology analysis. The essential conservative residues are marked as white characters on black background. The shadowed characters indicate the positions of significant relationship in the BLOSUM alignments and the semihomology relationship in the genetic semihomology approach. MultAlin consensus symbols: "%" is any of FY; "#" is any of NDQEBZ. ClustalW consensus symbols: "*" stands for identical or conserved residues in all sequences in the alignment; ":" indicates conserved subsitutions; "." indicates semi-conserved substitutions.

PROTEIN DESCRIPTIONS:

P07751	CHICKEN SPECTRIN ALPHA CHAIN, BRAIN (SPECTRIN, NON-ERYTHROID ALPHA CHAIN)
P14085	TYROSINE-PROTEIN KINASE TRANSFORMING PROTEIN SRC FROM AVIAN SARCOMA VIRUS (P60-SRC)
P39688	MOUSE PROTO-ONCOGENE TYROSINE-PROTEIN KINASE FYN (P59-FYN)
P00526	TYROSINE-PROTEIN KINASE TRANSFORMING PROTEIN SRC FROM ROUSE SARCOMA VIRUS(P60-SRC)
P09769	HUMAN PROTO-ONCOGENE TYROSINE-PROTEIN KINASE FGR (P55-FGR) (C-FGR)
P07948	HUMAN TYROSINE-PROTEIN KINASE LYN
P05433	P47(GAG-CRK) PROTEIN FROM AVIAN SARCOMA VIRUS
P07947	HUMAN PROTO-ONCOGENE TYROSINE-PROTEIN KINASE YES (P61-YES) (C-YES)
P08631	HUMAN TYROSINE-PROTEIN KINASE HCK (P59-HCK AND P60-HCK) (HEMOPOIETIC CELL KINASE)
P06240	MOUSE PROTO-ONCOGENE TYROSINE-PROTEIN KINASE LCK (P56-LCK) (LSK)
P42683	CHICKEN PROTO-ONCOGENE TYROSINE-PROTEIN KINASE LCK (PROTEIN-TYROSINE KINASE C-TKL)
P08487	BOVINE 1-PHOSPHATIDYLINOSITOL-4,5-BISPHOSPHATE PHOSPHODIESTERASE GAMMA 1 (PLC-GAMMA-1) (PHOSPHOLIPASE C-GAMMA-1)(PLC-II)(PLC-148)

and specifically considers features typical for α-spectrin. It works much better than the consensus obtained with the MultAlin (BLOSUM62) approach. The MultAlin consensus results could be obtained only when the expect threshold (the statistical significance threshold in BLAST similarity searches for reporting matches against database sequences) setting is extremely high. The genetic semihomology consensus does not separate α- and β-spectrins from each other as well as the consensus of Sahr and co-workers, but the expect threshold values are much lower. It means that the related proteins are easier to be found in the database as sequences of true homology.

Query sequence: Alpha spectrin segment consensus obtained with MultAlin program (BLOSUM62)

Sequences producing significant alignments:		Score (bits)	E Value
sp|P02549|SPCA_HUMAN	SPECTRIN ALPHA CHAIN, ERYTHROCYTE	17	21352
sp|Q01082|SPCO_HUMAN	SPECTRIN BETA CHAIN, BRAIN (SPECTRIN,...	16	27970
sp|P08032|SPCA_MOUSE	SPECTRIN ALPHA CHAIN, ERYTHROCYTE	16	27970
sp|P07751|SPCN_CHICK	SPECTRIN ALPHA CHAIN, BRAIN (SPECTRIN,...	16	27970
sp|Q13813|SPCN_HUMAN	SPECTRIN ALPHA CHAIN, BRAIN (SPECTRIN...	16	27970
sp|Q62261|SPCO_MOUSE	SPECTRIN BETA CHAIN, BRAIN (SPECTRIN, ...	16	27970
sp|P16086|SPCN_RAT	SPECTRIN ALPHA CHAIN, BRAIN (SPECTRIN, N...	16	36639
sp|Q00963|SPCB_DROME	SPECTRIN BETA CHAIN	16	47996
sp|P13395|SPCA_DROME	SPECTRIN ALPHA CHAIN	15	62874
sp|P16546|SPCN_MOUSE	SPECTRIN ALPHA CHAIN, BRAIN (SPECTRIN,...	14	141334
sp|P15508|SPCB_MOUSE	SPECTRIN BETA CHAIN, ERYTHROCYTE	14	185142
sp|P11277|SPCB_HUMAN	SPECTRIN BETA CHAIN, ERYTHROCYTE	14	185142
sp|P39254|Y04O_BPT4	HYPOTHETICAL 36.3 KD PROTEIN IN NRDC-M...	11	935541

Query sequence: Alpha spectrin segment consensus attained by Sahr et al. [1990]

Sequences producing significant alignments:		Score (bits)	E Value
sp|P13395|SPCA_DROME	SPECTRIN ALPHA CHAIN	43	2e-04
sp|Q13813|SPCN_HUMAN	SPECTRIN ALPHA CHAIN, BRAIN (SPECTRIN...	40	0.001
sp|P07751|SPCN_CHICK	SPECTRIN ALPHA CHAIN, BRAIN (SPECTRIN,...	39	0.003
sp|P16546|SPCN_MOUSE	SPECTRIN ALPHA CHAIN, BRAIN (SPECTRIN,...	39	0.004
sp|P16086|SPCN_RAT	SPECTRIN ALPHA CHAIN, BRAIN (SPECTRIN, N...	37	0.012
sp|P02549|SPCA_HUMAN	SPECTRIN ALPHA CHAIN, ERYTHROCYTE	37	0.016
sp|P08032|SPCA_MOUSE	SPECTRIN ALPHA CHAIN, ERYTHROCYTE	36	0.021
sp|P15508|SPCB_MOUSE	SPECTRIN BETA CHAIN, ERYTHROCYTE	35	0.048
sp|Q00963|SPCB_DROME	SPECTRIN BETA CHAIN	35	0.063
sp|Q01082|SPCO_HUMAN	SPECTRIN BETA CHAIN, BRAIN (SPECTRIN,...	34	0.082
sp|Q62261|SPCO_MOUSE	SPECTRIN BETA CHAIN, BRAIN (SPECTRIN, ...	34	0.11
sp|P11277|SPCB_HUMAN	SPECTRIN BETA CHAIN, ERYTHROCYTE	32	0.32
sp|P05095|AACT_DICDI	ALPHA-ACTININ 3, NON MUSCULAR (F-ACTIN...	21	797
sp|P34367|YLJ2_CAEEL	HYPOTHETICAL 256.3 KD PROTEIN C50C3.2 ...	20	1791
sp|Q03001|BPA1_HUMAN	BULLOUS PEMPHIGOID ANTIGEN 1 (BPA) (H...	19	3073
sp|P30427|PLEC_RAT	PLECTIN	19	4026
sp|P31670|GT27_FASHE	GLUTATHIONE S-TRANSFERASE 26 KD 47 (GS...	19	4026
sp|P46125|YEDI_ECOLI	HYPOTHETICAL 32.2 KD PROTEIN IN DSRB-V...	18	6908
sp|P42094|PHYT_BACSU	3-PHYTASE PRECURSOR (PHYTATE 3-PHOSPHA...	18	6908
sp|P56288|E2BG_SCHPO	PROBABLE TRANSLATION INITIATION FACTOR...	18	9049
sp|O00273|DFFA_HUMAN	DNA FRAGMENTATION FACTOR ALPHA SUBUNI...	18	9049
sp|P12311|ADH1_BACST	ALCOHOL DEHYDROGENASE (ADH-T)	18	9049

Query sequence: Alpha spectrin segment consensus attained by genetic semihomology algorithm

Sequences producing significant alignments:		Score (bits)	E Value
sp|P13395|SPCA_DROME	SPECTRIN ALPHA CHAIN	49	3e-06
sp|Q13813|SPCN_HUMAN	SPECTRIN ALPHA CHAIN, BRAIN (SPECTRIN...	47	1e-05
sp|P07751|SPCN_CHICK	SPECTRIN ALPHA CHAIN, BRAIN (SPECTRIN,...	42	3e-04
sp|P16086|SPCN_RAT	SPECTRIN ALPHA CHAIN, BRAIN (SPECTRIN, N...	41	6e-04
sp|P02549|SPCA_HUMAN	SPECTRIN ALPHA CHAIN, ERYTHROCYTE	38	0.005
sp|Q00963|SPCB_DROME	SPECTRIN BETA CHAIN	37	0.014
sp|P08032|SPCA_MOUSE	SPECTRIN ALPHA CHAIN, ERYTHROCYTE	36	0.019
sp|P15508|SPCB_MOUSE	SPECTRIN BETA CHAIN, ERYTHROCYTE	36	0.024
sp|P16546|SPCN_MOUSE	SPECTRIN ALPHA CHAIN, BRAIN (SPECTRIN,...	36	0.032
sp|Q01082|SPCO_HUMAN	SPECTRIN BETA CHAIN, BRAIN (SPECTRIN,...	34	0.094
sp|Q62261|SPCO_MOUSE	SPECTRIN BETA CHAIN, BRAIN (SPECTRIN, ...	34	0.094
sp|P11277|SPCB_HUMAN	SPECTRIN BETA CHAIN, ERYTHROCYTE	33	0.21
sp|P34367|YLJ2_CAEEL	HYPOTHETICAL 256.3 KD PROTEIN C50C3.2 ...	20	1195
sp|Q99001|AACB_CHICK	ALPHA-ACTININ, BRAIN ISOFORM (F-ACTIN ...	20	1566
sp|P35609|AAC2_HUMAN	ALPHA-ACTININ 2, SKELETAL MUSCLE ISOF...	20	1566
sp|P12814|AAC1_HUMAN	ALPHA-ACTININ 1, CYTOSKELETAL ISOFORM ...	20	1566
sp|P05094|AACT_CHICK	ALPHA-ACTININ, SMOOTH MUSCLE ISOFORM (...	20	1566
sp|P30427|PLEC_RAT	PLECTIN	19	2687
sp|P20111|AACS_CHICK	ALPHA-ACTININ, SKELETAL MUSCLE ISOFORM...	19	3520
sp|P47493|SYG_MYCGE	GLYCYL-TRNA SYNTHETASE (GLYCINE--TRNA L...	18	4611
sp|Q03001|BPA1_HUMAN	BULLOUS PEMPHIGOID ANTIGEN 1 (BPA) (H...	18	7913

Figure 5. Use of α-spectrin consensus achieved with different algorithms for sequence similarities search (BLAST).

See text for details.

Dot-matrix presentation of low but significant homologies

In this study 23 human erythrocyte α-spectrin repeats were taken as an example of low homology sequences of common origin. The homology between the segments is usually less than 25% (regarding identities). It is evident that all α-spectrin repeats have evolved from one ancestral gene encoding the initial 106-residue segment by several contiguous duplications (Speicher & Marchesi, 1984; Wasenius *et al.*, 1989). The sequence homology is hardly visible (Fig. 3), much more similarity concerns the secondary and tertiary structure of the spectrin segments, each possessing triple helical character (Speicher & Marchesi, 1984; Yan *et al.*, 1993). The dot plot reveals repeated internal homology along the α-spectrin chain when appropriate frame setting and identity threshold are used (Fig. 6). The

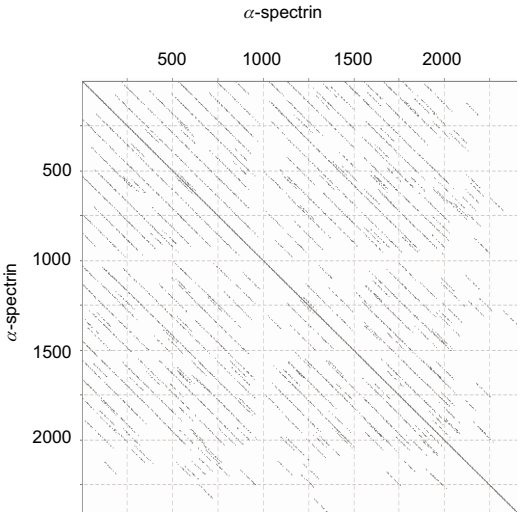

Figure 6. Dot matrix comparison of human erythrocyte α-spectrin with itself.

For the best visualization of the repeats, the identity threshold is set as 15, and frame size as 75. Details in the Results and Discussion.

common features of most repeats can be confirmed and more details can be concluded when the dot plot is run for α-spectrin *versus* the consensus segment sequence of Sahr and co-workers (Sahr *et al.*, 1990). This plot shows even more details when the genetic semihomology consensus is used and the plot is run in the semihomology mode (visualization of identical and semihomologous pairs) (Fig. 7). The repeating structure of the spectrin chain is then clearer, and the conservative spots can be localized more easily. The dot-plot analysis of α-spectrin chain with the consensus gives also some information about the evolutionary distance among the segments. The segments 10 and 21–23 are not visible on these plots at all, which indicates their high divergence in comparison with the other repeats. These conclusions are concordant with earlier reports (Speicher & Marchesi, 1984; Wasenius *et al.*, 1989). Additionally, this approach is a useful tool for detailed analysis of structurally and functionally essential fragments as well as of the differentiation mechanism of each segment.

Application of the genetic semihomology algorithm to the study of the six-codon amino acid distribution as a function of position variability A possible role of cryptic mutations in protein differentiation

Different numbers of codons (1 to 6) encode individual amino acids. Among them are three amino acids which have the maximum number of codons (6) – arginine, leucine and serine. The commonly used algorithms and programs do not consider the number of possible codons in the study of mutational replacement of a particular residue. The well-developed alignment procedures do not respect this parameter either. In this chapter the theoretical significance of multiple-codon residues and cryptic mutations in increased frequency of the amino acid replacement is discussed. Also the contribution and possible role of these residues in variable positions is presented.

The three six-codon amino acids differ in the diversity in codon composition. The least diversity is among leucine codons (CTX and TTR), the most difference is observed for the codons of serine (AGY and TCX). Arginine codons (AGR and CGX) are described as more diverse than leucine codons, since the latter always have a pyrimidine at the first position. There are mutations possible other than the replacement of the third nucleotide that do not change the encoded amino acid. For these amino acids even multiple mutations may not change the residue. Such mutations are de-

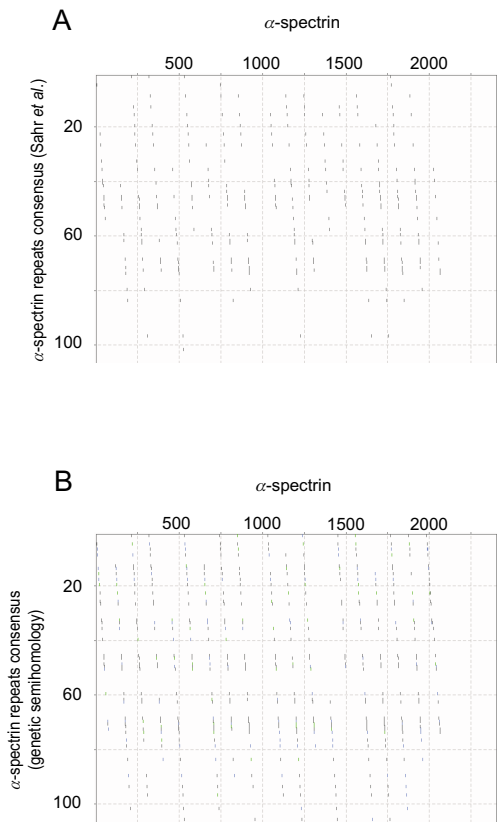

Figure 7. Dot matrix comparison of human erythrocyte α-spectrin with the consensus of 106-residue repeats.

The consensus is (A) as described by Sahr *et al.* (1990) or (B) achieved with the genetic semihomology approach (Leluk, 1998). The identity threshold and frame size are set as 8 and 40, respectively. In the genetic semihomology approach (B) the identity and genetic semihomology of the compared residues is visualized. Details in the text.

fined as cryptic. The genetic code matrix scores (GCM) (George *et al.*, 1990) for these specific cryptic mutations are within the range +1 to +2 for Arg and Leu and 0 to +2 for Ser. That means that for Arg and Leu a replacement of two codon positions does not change the amino acid and serine may remain even after replacement at all three positions of its codon (e.g., AGA → TCC). It is obvious that cryptic mutations should not have any evolutionary consequences at the protein level, since the protein remains identical. Therefore these mutations are not limited by structural or functional requirements of the protein. On the other hand, it is evident that among the mutations affecting the structure and function, the most common and most likely are single-base replacements. Thus the cryptic mutations may serve as a "passage" to increase the number of residues at a variable position. Theoretically a single point mutation can transform leucine to 10 amino acids. Arginine and serine may be changed to 12 amino acids each. For comparison the four-codon amino acids may be transformed to 7–8 other residues by a single point mutation. The genetic semihomology planar diagram (Fig. 8) shows the cryptic passages for Leu, Arg and Ser. If this mechanism is true then a higher frequency of these three amino acids should be observed at the very variable positions, where eight or more types of amino acids occur (Leluk, 1998; Leluk 2000a). Moreover, serine should dominate over arginine and arginine over leucine at those positions.

Several protein families were subjected to Ser-Arg-Leu occurrence analysis as a function of the degree of position variability (Fig. 9). Generally the theoretical prediction of their frequency was confirmed. However, for some protein families (e.g., eglin-like proteins) the results were dif-

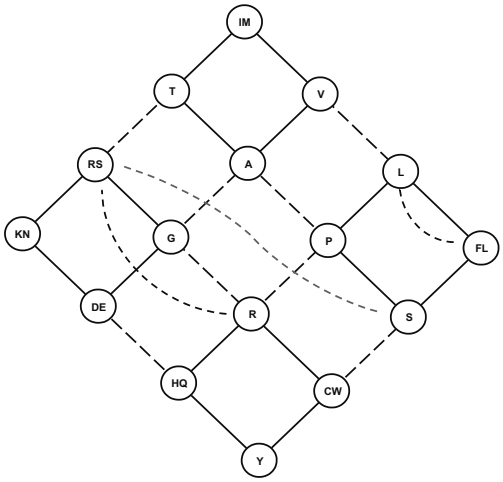

Figure 8. A planar diagram of genetic relationship between the amino acids showing the possibilities of reside-to-residue transformation by a single mutation at the first or second codon position (Leluk, 1998).

The related residues are placed on the same line. The significance of the third position of the codon is neglected in this diagram. The straight lines indicate the possibility of the amino acid exchange by one-nucleotide codon mutation (a solid line means transition type of nucleotide replacement, a dashed line means transversion). The dashed curves show the cryptic mutational passages (one or two single mutations are allowed) that do not cause a change at the amino acid level.

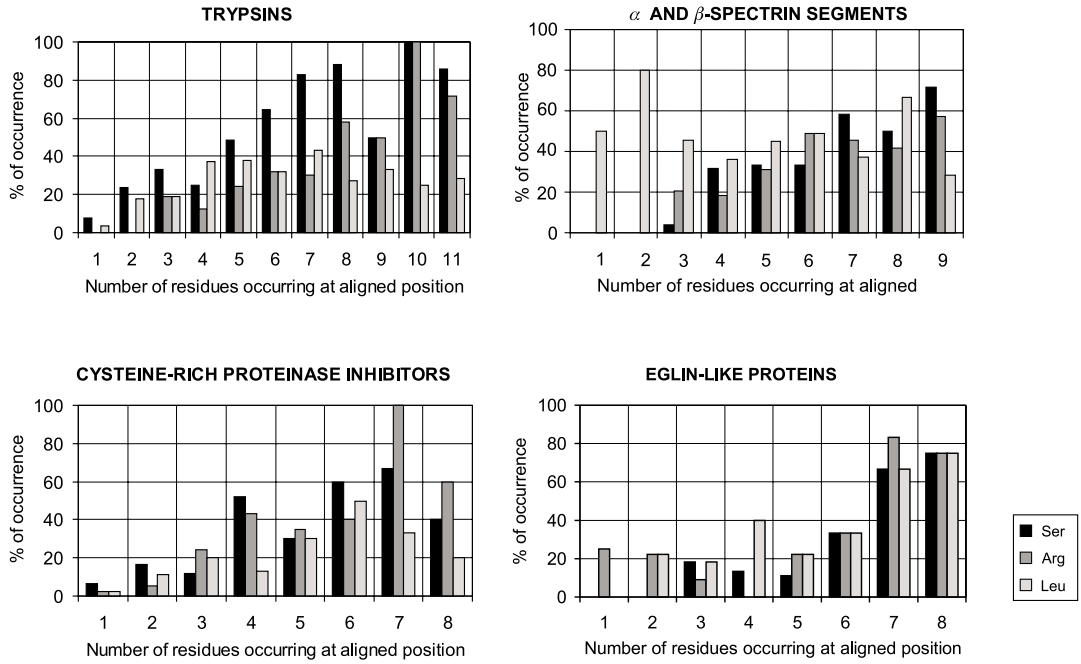

Figure 9. Frequency of six-codon amino acids as a function of position variability in different protein families.

Note that leucine frequency increase is not as regular as for arginine and serine. Except for the eglin-like proteins the serine and arginine occurrence is generally more significant at the most variable positions than leucine occurrence.

ferent than expected. The analysis of all sequences from all families combined showed a distinct domination of serine frequency at the positions occupied by seven and more residues (Fig. 10). Leucine contribution is the least, and the rise in frequency as the number of residues increases is not as clear for leucine as for serine and arginine. In conclusion, it may be assumed that the mechanism of cryptic passages of six-codon amino acids plays a special role in increasing the position variability.

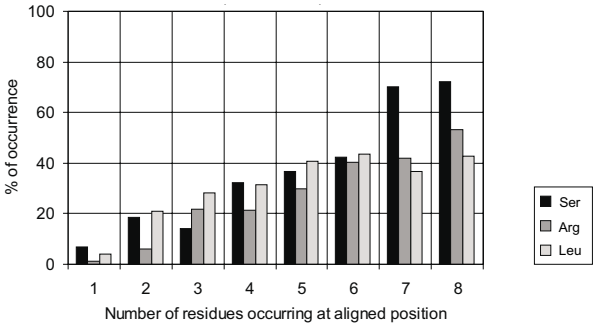

Figure 10. Frequency of six-codon amino acids as a function of position variability in randomly selected proteins of different origin and nature (see text for details).

The calculations concern 2686 residues at 606 corresponding positions.

The authors wish to thank Miss Monika Grabiec and Miss Monika Sobczyk for their contribution in the part of work concerning the multiple alignment verification and consensus formation.

REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.

Apweiler, R., Gateau, A., Contrino, S., Martin, M.J., Junker, V., O'Donovan, C., Lang, F., Mitaritonna, N., Kappus, S. & Bairoch, A. (1997) Protein sequence annotation in the genome era: The annotation concept of SWISS-PROT+TREMBL. ISMB-97; in *Proceedings 5th International Conference on Intelligent Systems for Molecular Biology*, pp. 33–43, AAAI Press, Menlo Park.

Baek, J.M., Song, J.C., Choi, Y.D. & Kim, S.I. (1994) Nucleotide sequence homology of cDNAs encoding soybean Bowman-Birk type proteinase inhibitor and its isoinhibitors. *Biosci. Biotechnol. Biochem.* **58**, 843–846.

Bailey, T.M. & Elkan, C. (1995) Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning J.* **21**, 51–83.

Bailey, T.M. & Gribskov, M (1997) Score distributions for simultaneous matching to multiple motifs. *J. Comput. Biol.* **4**, 45–59.

Bailey, T.M. & Gribskov, M. (1998a) Methods and statistics for combining motif match scores, *J. Comput. Biol.* **5**, 211–221.

Bailey, T.M. & Gribskov, M. (1998b) Combining evidence using p-values: Application to sequence homology searches. *Bioinformatics* **14**, 48–54.

Bairoch, A. & Apweiler, R. (1997) The SWISS-PROT protein sequence database: Its relevance to human molecular medical research. *J. Mol. Med.* **75**, 312–316.

Bairoch, A. & Apweiler, R. (1999) The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. *Nucleic Acids Res.* **27,** 49–54.

Bairoch, A. (1997) Proteome databases; in *Proteome Research: New Frontiers in Functional Genomics* (Wilkins, M.R., Williams, K.L, Appel, R.D., Hochstrasser, D.H., eds.) pp. 93–132, Springer Verlag, Heidelberg.

Bucher, P. & Bairoch, A. (1994) A generalized profile syntax for biomolecular sequence motifs and its function in automatic sequence interpretation. ISMB-94; in *Proceedings 2nd International Conference on Intelligent Systems for Molecular Biology* (Altman, R., Brutlag, D., Karp, P., Lathrop, R. & Searls, D., eds.) pp. 53–61, AAAIPress, Menlo Park.

Chen, X., Qian, Y., Chi, C., Gan, K., Zhang, M. & Chang-Qing, C. (1992) Chemical synthesis, molecular cloning, and expression of the gene coding for the *Trichosanthes* trypsin inhibitor — a squash family inhibitor. *J. Biochem.* **112**, 45–51.

Corpet, F. (1988) Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res.* **16**, 10881–10890.

Dayhoff, M.O. & Eck, R.V. (1968) *Atlas of Protein Sequence and Structure* (Dayhoff, M.O. & Eck, R.V., eds.) vol. 3, p. 33, Silver Spring, MD.

Dayhoff, M.O., Schwartz, R.M. & Orcutt, B.C. (1979*) Atlas of Protein Sequence and Structure* (Dayhoff, M.O., ed.) vol. 5, Suppl. 3, p. 345, Washington, DC.

George, D.G., Barker, W.C. & Hunt, L.T. (1990) Mutation data matrix and its uses. *Methods Enzymol.* **183**, 333–351.

Gish, W. & States, D.J. (1993) Identification of protein coding regions by database similarity search. *Nat. Genet.* **3**, 266–272.

Grundy, W.N., Bailey, T.M. & Elkan, C.P. (1997) ParaMeme: A parallel implementation and a Web interface for a DNA and protein motif discovery tool. *CABIOS* **12**, 303–310.

Hakateyama, T., Hiraoka, M. & Funatsu, G. (1991) Amino acid sequence of the two smallest trypsin inhibitors from sponge gourd seeds. *Agric. Biol. Chem.* **10**, 2641–2642.

Haldar, U.C., Saha, S.K., Beavis, R.C. & Sinha, N.K. (1996) Trypsin inhibitors from ridged gourd (*Luffa acutangula* Linn.) seeds: Purification, properties, and amino acid sequences. *J. Protein Chem.* **15**, 177–184.

Hamato, N., Koshiba, T., Pham, T., Tatsumi, Y., Nakamura, D., Takano, R., Hayashi, K., Hong, Y. & Hara, S. (1995) Trypsin and elastase inhibitors from bitter gourd (*Momordica charantia* Linn.) seeds: Purification, amino acid sequences, and inhibitory activities of four new inhibitors. *J. Biochem.* **117**, 432–437.

Hayashi, K., Takehisa, T., Hamato, N., Takano, R., Hara, S., Miyata, T. & Kato, H. (1994) Inhibition of serine proteinases of the blood coagulation system by squash family protease inhibitors. *J. Biochem.* **116**, 1013–1018.

Heitz, A., Chiche, L., Le-Nguyen, D. & Castro, B. (1989) 1H 2D NMR and distance geometry study of the folding of *Ecbalium elaterium* trypsin inhibitor, a member of the squash inhibitor family. *Biochemistry* **28**, 2392–2398.

Hofmann, K., Bucher, P., Falquet, L. & Bairoch, A. (1999) The PROSITE database, its status in 1999. *Nucleic Acids Res.* **27,** 215–219.

Joubert, F.J. (1984) Trypsin inhibitors from *Momordica repens* seeds. *Phytochemistry* **23**, 1401–1406.

Junier, T. & Pagni, M. (2000) Dotlet: Diagonal plots in a Web browser. *Bioinformatics* **16**, 178–179.

Lee, C. & Lin, J. (1995) Amino acid sequences of trypsin inhibitors from the melon *Cucumis melo*. *J. Biochem.* **118**, 18–22.

Leluk, J. (1998) A new algorithm for analysis of the homology in protein primary structure. *Computers Chem.* **22**, 123–131.

Leluk, J. (2000a) A non-statistical approach to protein mutational variability. *Biosystems* **56**, 83–93.

Leluk, J. (2000b) Regularities in mutational variability in selected protein families and the Markovian model of amino acid replacement. *Computers Chem.* **24**, 659–672.

Leluk, J. (2000c) Serine proteinase inhibitor family in squash seeds: Mutational variability mechanism and correlation. *Cell. Mol. Biol. Lett.* **5**, 91–106.

Ling, M.-H., Qi, H. & Chi, C. (1993) Protein, cDNA and genomic DNA sequences of the towel gourd trypsin inhibitor. *J. Biol. Chem.* **268**, 810–814.

Matsuo, M., Hamato, N., Takano, R., Kamei-Hayashi, K., Yasuda-Kamatani, Y., Nomoto, K. & Hara, S. (1992) Trypsin inhibitors from bottle gourd (*Lagenaria leucantha Rusby* var. Depressa Makino) seeds. Purification and amino acid sequences. *Biochim. Biophys. Acta* **1120**, 187–192.

McGurl, B., Mukherjee, S., Kahn, M. & Ryan, C.A. (1995) Characterization of two proteinase inhibitor (ATI) cDNAs from alfalfa leaves (*Medicago sativa* var. Vernema): The expression of ATI genes in response to wounding and soil microorganisms. *Plant Mol. Biol.* **27**, 995–1001.

Morgenstern, B., Dress, A. & Werner, T. (1996) Multiple DNA and protein sequence alignment based on segment-to-segment comparison. *Proc. Natl. Acad. Sci. U.S.A.* **93**, 12098–12103.

Morgenstern, B. (1999) DIALIGN 2: Improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics* **15**, 211–218.

Nishino, J., Takano, R., Kamei-Hayashi, K., Minakata, H., Nomoto, K. & Hara, S. (1992) Amino-acid sequencers of trypsin inhibitors from oriental pickling melon (*Cucumis melo* L. var. Conomon Makino) seeds. *Biosci. Biotech. Biochem.* **56**, 1241–1246.

Odani, S. & Ikenaka, T. (1978) Studies on soybean trypsin inhibitors, XII. Linear sequences of two soybean double-headed trypsin inhibitors, D-II and E-I. *J. Biochem.* **83**, 737–745.

Otlewski, J., Whatley, H., Polanowski, A. & Wilusz, T. (1987) Amino-acid sequences of trypsin inhibitors from watermelon (*Citrullus vulgaris*) and red bryony (*Bryonia dioica*). *Biol. Chem. Hoppe-Seyler* **368**, 1505–1507.

Pearson, W.R. & Lipman, D.J. (1988) Improved tools for biological sequence analysis, *Proc. Natl. Acad. Sci. U.S.A.* **85,** 2444–2448.

Pearson, W.R. (1990) Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol.* **183**, 63–98.

Sahr, K.E., Laurila, P., Kotula, L., Scarpa, A.L., Coupal, L., Leto, T.L., Linnenbach, A.J., Winkelmann, J.C., Speicher, D.W., Marchesi, V.T., Curtis, P.J. & Forget, B.G. (1990) The complete cDNA and polypeptide sequences of human erythroid alpha-spectrin. *J. Biol. Chem.* **265**, 4434–4443.

Scott, M.J., Huckaby, C.S., Kato, I., Kohr, W., Laskowski, M., Jr., Tsai, M.-J. & O'Malley, B.W. (1987) Ovoinhibitor introns specify functional domains as in the related and linked ovomucoid gene. *J. Biol. Chem.* **262**, 5899–5907.

Speicher, D.W. & Marchesi, V.T. (1984) Erythrocyte spectrin is comprised of many homologous triple helical segments. *Nature* **311**, 177–180.

Stachowiak, D., Polanowski, A., Bieniarz, G. & Wilusz, T. (1996) Isolation and amino-acid sequence of two inhibitors of serine proteinases, members of the squash inhibitor family, from *Echinocystis lobata* seeds. *Acta Biochim. Polon.* **43**, 507–514.

Tatusova, T.A. & Madden, T.A (1999) BLAST 2 SEQUENCES – a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol. Lett.* **174**, 247–250.

Thompson, J.D., Higgins, D.G. & Gibson, T.J. (1994) CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680.

Tomii, K. & Kanehisa, M. (1996) Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. *Protein Engng* **9**, 27–36.

Wasenius, V.-M., Saraste, M., Salven, P., Eramaa, M., Holm, L. & Lehto, V.-P. (1989) Primary structure of the brain α-spectrin, *J. Cell Biol.* **108**, 79–93.

Wieczorek, M., Otlewski, J., Cook, J., Parks, K., Leluk, J., Wilimowska-Pelc, A., Polanowski, A., Wilusz, T. & Laskowski, M., Jr. (1985) The squash inhibitor family of serine proteinase inhibitors. Amino acid sequences and association equilibrium constants of inhibitors from squash, summer squash, zucchini, and cucumber seeds. *Biochem. Biophys. Res. Commun.* **126**, 646–652.

Yan, Y., Winograd, E., Viel, A., Cronin, T., Harrison, S.C. & Branton, D. (1993) Crystal structure of the repetitive segments of spectrin. *Science*, **262**, 2027–2030.