LITTLE KEVIN: a program for the estimation of protein homology by analysing the amino acid compositions and sequences

J.P. Grillasca¹ ³, D. Aubert¹, R. Planells², N. Domingo¹ and H. Lafont¹

Abstract

We present here a computational method based on the analysis of amino acid composition for performing comparisons between proteins. This user-friendly and reliable test is aimed at rapidly identifying, from data-base subsets, sequences—if necessary, partial sequences—which share similar amino acid compositions to the input composition (deduced from experimental results). Apparent molecular weight (as determined by SDS-PAGE) and artefactual modifications due to the experimental determination of the amino acid composition are taken into account to perform the comparison. This program thus constitutes a useful tool in searching for the probable identification of either non-sequenced proteins or peptides from hydrolysed proteins.

Introduction

Comparison of homology, and similarity, between proteins has grown into a widespread and powerful tool in molecular biology. Many programs have been developed so far with this aim, but most of them need the availability of the primary structures of proteins that have to be compared. If this requirement is not fulfilled, comparison is not possible. Nevertheless, in some experimental cases, only partial information like amino acid molar composition and molecular mass of a given protein—or even a peptide fragment and not the entire protein—may be available.

Here we describe the program LITTLE KEVIN which is designed to perform comparisons between proteins by using these two parameters. Similar methods have already been reported (Sibbald et al., 1991; Shaw, 1993), but LITTLE KEVIN is more particularly devoted to performing comparisons between amino acid compositions, which may then represent either the composition of entire proteins or only the composition of peptide fragments. This program, using as input the experimental data, first calculates the probable number of residues in the polypeptide of interest. These putative patterns, in terms of the number of amino acids, are used to perform comparisons with the amino acid compositions deduced from total or partial sequences recovered from available data banks. This comparison allowed us to know whether the molar composition of the protein of interest could correspond, with respect to its molecular weight, to all or a particular part of a protein sequence and, if so, how high the degree of homology could be. This straightforward program allowed us to save much time in studying relationships between two polypeptides: the anionic polypeptide fraction (APF) and the liver fatty acid binding protein (L-FABP).

APF, although extensively studied since 1985 (Martigne et al., 1985) and purified to a high degree, still remained difficult to sequence. Indeed, this protein component of bile, which binds lipids and bilirubin, is blocked at its NH₂ terminus. The amino acid molar composition and the molecular weight of this protein are known. Biological arguments led us to suggest a link between APF and L-FABP (Chan et al., 1985): we demonstrated a high cross-reactivity between antibodies directed against these two proteins (J.P. Grillasca et al., unpublished data). Our problem was, therefore, to determine whether or not the hypothesis of a relationship between APF and L-FABP was valid, taking into account the different mol. wts (7 and 14 kDa for APF and L-FABP, respectively). If this hypothesis turns out to be true, then two hypotheses can be put forward: either L-FABP gives rise to APF by hydrolysis, or the L-FABP gene results in the production of both proteins (by alternative splicing or by using an alternative promoter). The development of LITTLE KEVIN allowed us to demonstrate that there was a strong probability that APF could be the C-terminal moiety of L-FABP, and this original approach promises to be very useful in the study of relationships between proteins.

System and methods

Materials

The equipment necessary for the use of this program includes a PC compatible computer (with mouse) belonging to the 386 DX class or higher. It must possess a clock speed of 40 MHz or more. The minimum RAM must be 4 Mbytes and the hard disk must possess a minimum of
5 Mbytes free to install LITTLE KEVIN, but the memory buffer must be able to evolve with the acquisition of new data, protein compositions and sequences. LITTLE KEVIN is a program compiled in the BASIC language in a WINDOWS environment.

Data input
Several parameters must be specified in LITTLE KEVIN before the calculations can begin.

(i) The experimental amino acid molar composition of the protein under analysis. The destruction of tryptophan and the permutation of asparagine and glutamine residues to aspartate and glutamate, respectively, will be discussed below.

(ii) The molecular weight (MW) of the protein under analysis, in Daltons (D).

These first two parameters are input through dialogue box number 1. Other parameters concerning optional data may be input through dialogue box number 2:

(iii) The approximate percentage of error of the protein molecular weight, if estimated from SDS-PAGE, this parameter can vary from 0 to 100%.

(iv) The percentage of window size variation during the alignment of the amino acid composition of the studied protein with the sequence of protein from the data bank. It is often desirable to vary the size of the window to optimize the search. When choosing this percentage one must take into account the quality of the composition and the precision of the determined molecular weight. This parameter represents the number of residues that LITTLE KEVIN uses to calculate the amino acid composition of the proteins from the data bank.

(v) The minimum percentage of homology (defined as below) for which the program retains the sequence as a putative solution.

A third dialogue box permits the input of the sequence against which the putative patterns deduced from the input composition are compared. These sequences are obtained from data banks and can be entered manually or by a 'cut and paste' function. Amino acids are specified by the international one-letter code. All proteins can be stored in memory and modified at will.

Algorithm
Using the amino acid molar composition, the molecular weight and the different parameters concerning the given protein, LITTLE KEVIN calculates the most probable amino acid patterns of the given protein in terms of actual numbers of residues. This is achieved by using information stored permanently in memory, i.e. the molecular weights of the 20 amino acids (module 1 in Figure 1).

The program then counts the number of each amino acid of the first matching sequence using a window, established as a function of the molecular weight of the protein under study, and the percentage of window size variation entered in options. This window moves from the beginning to the end of the protein sequence. The program compares all the amino acid patterns to the compositions deduced from amino acids nested by the window along the sequence(s).

Finally, the results are represented by shaded zones on the sequence(s) which could, with greater or lesser degrees of homology, correspond to the composition of the protein under study (Figure 2), LITTLE KEVIN then saves the solutions having a homology greater or equal to the value specified in the options (module 2 in Figure 1).

Method of calculation
Calculation of the quantity of a single amino acid in each pattern:

\[ \text{Solution}(X) = \left( \frac{\text{%Composition}(Y) \times \text{I} \times \sum_{Y} \text{%Composition}(Y) \times \text{AminoAcidWeight}(Y)}{100} \right) \]

Where %Composition(X) = molar percentage of the amino acid, X, entered by the operator; I = value representing the molecular weight of the protein (reiteration for: \( \text{MW-option} \times \text{size} < 1 < \text{MW + option} \times \text{size} \)); AminoAcidWeight(Y) = pre-programmed MW of the amino acid, Y, in Daltons; Solution Retained(X) = whole number value for the amino acid displayed in the table.

Remark: Int (exact value) is always rounded down; e.g. (Int 4.9 = 4), while Int (4.9 + 0.5) = 5.

Calculation of the deviation:

\[ \text{Deviation} = \sum_{X} \left( \frac{\text{Solution}(X) - \text{Int(solution}(X)) - 0.5}{\text{Solution}(X)} \right) \]

where Deviation = the total of all rounding errors calculated for each amino acid of a particular solution.

Per cent homology of the solution:

\[ \text{Homology} = \left[ 100 \times \frac{\sum_{X} \text{solutionretained}(X)}{\text{WindowCompositionAminoAcid}(Y)} \right] \]

Where WindowCompositionAminoAcid(Y) = number of amino acid, Y, present in the alignment window; WindowSizeAminoAcid = size of alignment window (in amino acid).
Output of the results

Results displayed on the screen (Figure 2) can also be saved to a file or directed to a printer. Results of module 1, i.e. amino acid per cent compositions, are presented with a deviation (see Method of calculation) and the putative number of amino acids in the sequence. Results of module 2 are presented in the form of shaded zones of the different matching sequence which correspond to the composition entered at the beginning of the calculations (Figure 2). With this result, several parameters are given which must be taken into account along with the analysis of the

Fig. 1. Algorithm of LITTLE KEVIN.
Fig. 2. Screen display of results obtained when a given composition is compared to a protein sequence. Even in the case of comparison to several proteins, solutions are listed according to decreasing homology (solution 1 is the most homologous). The shaded window represents the solution whose sequence appears in line 3 as amino acid composition and is to compare to line 2 where the calculated pattern deduced from a given composition (line 1) is displayed.

program's responses: (i) the experimental determination of the protein molecular composition that must particularly take into account the destruction of tryptophan, and the permutation of asparagine and glutamine to aspartate and glutamate, respectively. (ii) the homology between the sequence composition and the calculated pattern; (iii) the length of the resultant solution in comparison with the number of amino acids in the calculated pattern.

Results

The reliability of the program has been rigorously verified. Two verifications have been undertaken.

First, several (>10^6) imaginary sequences from 40 to 40,000 residues were created by an informatic random process. Using the Geneworks program and Genbank data base on CD-ROM, more than 200 recorded sequences matching, even partially, with only one imaginary random polypeptide were subloaded onto a sequence subset. Amino acid compositions of those imaginary sequences that were represented in the sequences subset were calculated and LITTLE KEVIN was always able to recover, for each calculated pattern, the corresponding protein from the pool of sequences.

Secondly, the ability of LITTLE KEVIN to recover a protein thanks to its amino acid composition, in spite of modifications brought during the experimental determination of this composition, was tested as follows. Amino acid molar compositions of 100 human proteins or parts of proteins (polypeptide hormones, transcription factors, other proteins of very particular composition, like proline-rich proteins) were calculated. These compositions were then deliberately modified in the following ways: elimination of tryptophans, permutation of asparagine and glutamine into aspartate and glutamate and slight modifications that were randomly brought to the final percentage of each amino acid within ±10% limits. We than checked for the efficiency of LITTLE KEVIN to recover the sequence (and even the partial domain) from which the composition was derived. Results obtained using LITTLE KEVIN were very encouraging. Each composition was matched against the 100 human protein sequences and in each case, using the default setting (molecular weight, ±10%; searching window size, ±5%; and threshold of amino acid homology for retaining the solution, 70%), the mother sequence was always recovered and no more than four or five windows around the entered domain were proposed.

It thus appears that LITTLE KEVIN can serve as an effective tool for orienting a plan of research: if the program does not find a given composition in any sequence or part of a sequence, the negative result may be considered as very reliable and the probability that a homologous protein may exist in the sequences submitted to analysis is very weak.

Discussion

LITTLE KEVIN is a new program which has proven to be easy to use, efficient (taking into account the whole amino
acid composition), and devoted to the sorting of proteins that share, at least partially, a given composition of amino acids. This kind of search is very important in studying proteins that resist sequencing attempts, which is the case for APF to date.

Two other methods had been described, but did not fulfil the same purposes. One of them (Sibbald et al., 1991) works with only a few amino acid entries to sort the matching sequences more rapidly, even partial sequences, but the risk of losing some of them or obtaining false-positive results is enhanced. The second method (Shaw, 1993) does not take into account the possible hydrolysis or post-translational modifications of primary structures of proteins.

LITTLE KEVIN is useful in the identification of fragments from a mixture of several hydrolysed proteins after separation of these fragments by SDS-PAGE: instead, the degradation of a fusion protein produced in a prokaryotic system, whether this is done to purify part of a protein with the goal of producing an antibody, to identify protease-sensitive regions or to discover ligand binding sites. With this program, it becomes possible to produce and use such a protein without having to sequence it.

A brief explanation, in a user-friendly environment, is available to the operator of LITTLE KEVIN for entering the various parameters that must be specified.

Nevertheless, we are aware of the limits of this program and intend to continue its evolution in several directions: use through a network; use with a Macintosh system, often used by biologists; capturing sequences directly by searching in the databases such as Swissprot or Genebank, and bypassing the 'cut and paste' procedures.

Acknowledgements

We wish to thank Jason Sanger for his skilful help in the preparation of the English version.

References


Received on January 2, 1996; accepted on April 23, 1996