BIOINFORMATICS APPLICATIONS NOTE Vol. 16 no. 12 2000
Pages 1153–1154

ASAP: analysis of peptide composition

Xavier Serra-Hartmann[1], Xavier Rebordosa[1], Jaume Piñol[1], Enrique Querol[1] and Marc A. Martí-Renom[1, 2,*]

[1]Institut de Biologia Fonamental, Universitat Autònoma de Barcelona, 08193 Bellaterra (Barcelona), Spain and [2]Laboratories of Molecular Biophysics, Pels Family Center for Biochemistry and Structural Biology, The Rockefeller University, 1230 York Ave, New York, NY 10021, USA

Received on June 3, 2000; revised on July 12, 2000; accepted on July 29, 2000

Abstract

Summary: *ASAP is a web tool designed to search for specific dipeptides, tripeptides and tetrapeptides in a protein sequence database. The server allows for: (a) identification of frequent and infrequent peptides and the creation of peptide probability tables for a given database of sequences (GenerNet program), (b) determination of the compatibility of an amino-acid sequence to the given peptide probability tables (ClonErrNet program); and (c) comparison of different protein databases based on peptide composition (CompNet program). ASAP server can be useful in protein engineering and/or protein classification studies.*

Availability: *http://guitar.rockefeller.edu/∼marcius/ASAP.html*

Contact: *martim@rockefeller.edu*

Different studies based on the determination of residue distribution in a particular organism try to reveal the so-called residue usage. Karlin and co-workers (Karlin *et al.*, 1990, 1994; Karlin and Bucher, 1992) have been studying the distribution of amino-acids and their relation to protein structure and codon usage for different organisms. The amino-acid neighborhood relationship in proteins has also been studied previously (Erhan *et al.*, 1980; Vonderviszt *et al.*, 1986). These two separate works demonstrate that the occurrence of di-peptides and tri-peptides are non-random and can be used as an objective key for classification of organisms and proteins. To date however, no-exhaustive analysis of the distribution of di-, tri-, and tetra-peptides for a given proteome has been reported. The ASAP (Analysis of Sequence and Amino acid Probabilities) application example presented here shows an organism-dependent peptide usage in protein synthesis in an approach similar to that of Burge and co-workers (Burge *et al.*, 1992), who analyzed the relative abundance of di-, tri-, and tetra-nucleotides within and between coding regions of complete genomes.

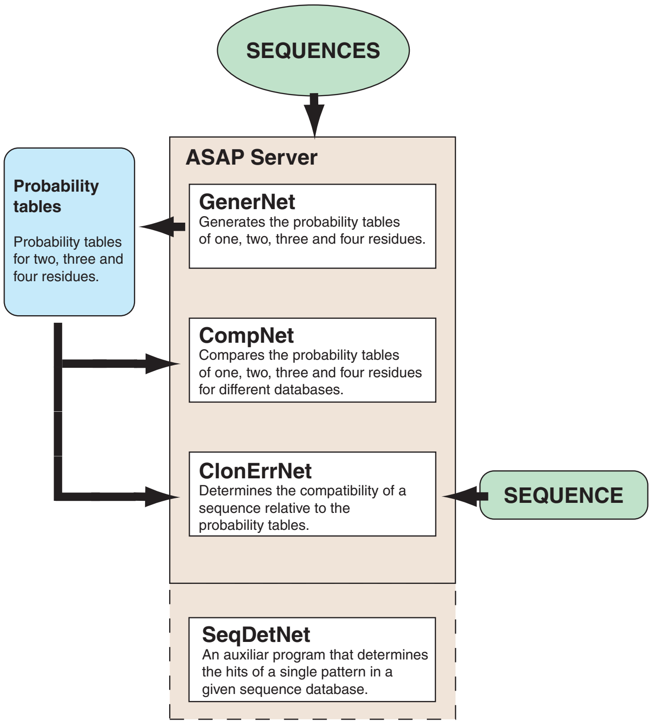
SEQUENCES
ASAP Server
Probability tables
Probability tables for two, three and four residues.
GenerNet
Generates the probability tables of one, two, three and four residues.
CompNet
Compares the probability tables of one, two, three and four residues for different databases.
ClonErrNet
Determines the compatibility of a sequence relative to the probability tables.
SEQUENCE
SeqDetNet
An auxiliar program that determines the hits of a single pattern in a given sequence database.

Fig. 1. General architecture of ASAP.

The ASAP web server analyzes the distribution of peptides (ranging from one to four residues) in complete genomes. The web server is based on three main programs: GenerNet, ClonErrNet and CompNet and an auxiliary program SeqDetNet (Figure 1). GenerNet generates three tables which contain all probabilities of occurrences of di-, tri-, and tetra-peptides. ClonErrNet determines the fitting of a single sequence to the given probability tables. CompNet compares different probability tables. SeqDetNet searches for a peptide sequence in a database of proteins. The example outlined below can be found on the ASAP URLs.

[*]To whom correspondence should be addressed.

GenerNet program

The input to this program is a sequence database (e.g. the proteome of *E.coli*) in the FASTA format. The program reads all entries in the database and generates probability tables for di-, tri-, and tetra-peptides. Six output files can be downloaded from the net. The first part of the output gives information about the database, including the total numbers of proteins and residues. The second part reports individual probabilities of the 22 residue types (including B and Z) in the database. The third part of the output report is composed of a series of plots showing the distribution of Z-scores for 2, 3 and 4 residue patterns. The average and standard deviation of the preference values are also reported. The Z-scores are computed for the 400, 8000 and 160 000 possible di-, tri- and tetra-peptide patterns, respectively. Finally, two files with the under- and over-represented sequences at significance of 1% are listed at the end of the report. These files give information about the patterns and their expected and actual occurrences.

ClonErrNet program

This program takes a single sequence as input and three files containing the probability tables of 2, 3 and 4 residue sequences (e.g. a pre-built probability tables for the *E.coli* proteome and an envelope polyprotein gE from BHV-1 (Rebordosa *et al.*, 1996)). The program reads the single sequence to be checked against the database and calculates its Z-score for all positions in the sequence depending on the probability tables chosen by the user. The first part of the HTML output gives information about the database, listing the average and standard deviations for the preference values. The second part reports individual Z-scores for the input sequence and the distribution of the Z-scores for two, three and four residue sequences. This section also reports the Z-score distribution for 100 randomized sequences with the same composition as the target sequence. This calculation is performed to assess the significance of the average Z-score of the input sequence. If the target sequence has an irregular distribution of Z-scores compared to that of the randomized sequences, the differences found for the target sequence are more significant. Finally, an average Z-score for the target sequence and its significance are reported at the end of each probability table.

CompNet program

This program takes as input a pair of probability tables for different databases and compares them (e.g. a pre-built database for *S.cerevisiae* is compared with a pre-built database for *E.coli*). The program reads the probability tables of two, three and four residue sequences and calculates the Z-scores for the preference values of actual matches in each database. The first part of the report includes the average and standard deviation of the differences between the databases for di-, tri-, and tetra-peptides and the distribution of Z-scores. Finally, the significant differences (at level of 1%) are shown.

A World Wide Web Server was constructed. The ASAP server is based on the three programs that analyze the distribution of peptides between one and four residues in an input sequence database. A user-friendly graphical interface is provided by the web browser. The probability tables generated by GenerNet program can be used to assess if a sequence is compatible to a specific database of sequences. Other applications of ASAP server can be the comparison of two peptide-probability distributions to asses the similarity or difference between two databases or the annotation of new proteins to a given family of sequences. The results derived from the ASAP server can also be used to re-design residue mutations that overcome low represented regions in a protein which may be difficult to express recombinantly.

Acknowledgements

The authors are grateful to Drs F.X.Avilés, E.Querol and A.Šali for encouragement about the project. The authors also wish to thank Dr X.Rebordosa for providing useful information of gE expression and Drs F.W.Studier, A.Fiser and A.Stuart for many helpful comments. M.A.M-R. is supported by a post-doctoral fellowship awarded by The Burroughs Wellcome Fund. The project has been aided by grant BIO2000-0647.

References

Burge,C., Campbell,A. and Karlin,S. (1992) Over- and under-representation of short oligonucleotides in DNA sequences. *Proc. Natl Acad. Sci. USA*, **89**, 1358–1362.

Erhan,S., Marzolf,T. and Cohen,L. (1980) Amino-acid neighborhood relationships in proteins. breakdown of amino-acid sequences into overlapping doublets, triplets and quadruplets. *Int. J. Biomed. Comput.*, **11**, 67–75.

Karlin,S., Blaisdell,B. and Brendel,V. (1990) Identification of significant sequence patterns in proteins. *Meth. Enzymol.*, **183**, 388–402.

Karlin,S. and Bucher,P. (1992) Correlation analysis of amino acid usage in protein classes. *Proc. Natl. Acad. Sci. USA*, **89**, 12 165–12 169.

Karlin,S., Zuker,M. and Brocchieri,L. (1994) Measuring residue associations in protein structures. Possible implications for protein folding. *J. Mol. Biol.*, **239**, 227–248.

Rebordosa,X., Piñol,J., Perez-Pons,J., Lloberas,J., Naval,J., Serra-Hartmann,X., Espuña,E. and Querol,E. (1996) Glycoprotein e of bovine herpesvirus type 1 is involved in virus transmission by direct cell-to-cell spread. *Virus. Res.*, **45**, 59–68.

Vonderviszt,F., Matrai,G. and Simon,I. (1986) Characteristic sequential residue environment of amino acids in proteins. *Int. J. Pept. Protein Res.*, **27**, 483–492.