John von Neumann Institute for Computing

A Parameterized Procedure for Consensus Sequences - Validation of the Method

Sławomir Walkowiak, Bogdan Lesyng

published in

NIC Workshop 2006,
From Computational Biophysics to Systems Biology,
Jan Meinke, Olav Zimmermann,
Sandipan Mohanty, Ulrich H.E. Hansmann (Editors)
John von Neumann Institute for Computing, Jülich,
NIC Series, Vol. **34**, ISBN-10: 3-9810843-0-6,
ISBN-13: 978-3-9810843-0-6, pp. 197-200 , 2006.

http://www.fz-juelich.de/nic-series/volume34

A Parameterized Procedure for Consensus Sequences - Validation of the Method

Sławomir Walkowiak and Bogdan Lesyng

Department of Biophysics, Warsaw University,
Zwirki i Wigury 93, 02-089 Warsaw, Poland
E-mail: {*swalk, lesyng*}*@icm.edu.pl*

A consensus sequence for a given protein family is an averaged, representative sequence which describes common features of the family. Once generated it can be used as a query to protein databases. The quality measure of a constructed consensus sequence can be defined as a number of newly found sequences, or as an average similarity to all family members. In this work we made an attempt to compare these two ways of estimation of the quality measure of consensus sequences, and/or find any correlation between them. Four kinase families were chosen, and for each of them seven consensus sequences were constructed. Each of newly generated sequences was designed based on different sets of threshold parameters. Each of the constructed sequences was used as a database query, and also for each of them an average similarity to its family was determined. The average level of similarity was computed by scoring all possible pairwise alignments with the aid of a semihomology algorithm[1], estimating statistical significance of the number of identities and the identity distribution. Although we haven't found any unique set of the threshold parameters which could provide best results in both methods, we observed some correlation between the chosen sets of the threshold parameters and the 'scores' obtained for the consensus sequences.

1 Introduction

A consensus sequence can be defined as a best, unique sequence representation for the given protein family. The construction process of the sequence depends on an arbitrarily chosen set of control parameters describing the residue type at a specific position. Various sets of the threshold parameters defining gaps, conservative or non specific residue types, usually provide different consensus sequences[2].

2 Estimation of the Quality of Consensus Sequence

In this study we estimate the quality measure of consensus sequences in two ways. Firstly a consensus sequence was used as a query for protein databases and an average similarity to its family was determined. Secondly we computed an average level of similarity to all given protein family members. This procedure was applied for each of seven different parameter sets describing different consensus sequences for four kinase families.

3 Consensus Sequence as a Database Query

Each consensus sequence was used in a BLAST[3] query, and from the BLAST output sets we selected only those hits for which e-values were higher then a preset threshold value. In order to cutoff statistically insignificant hits, the obtained sets were then compared with the results for the template sequences, for each of the protein families. Table 1 1 presents the results.

Family	parameter set	no. of new sequences	$-<log(e-value)>$
RKT	6	23	14,17
RKT	3	23	14
RKT	2	23	14
RKT	1	22	13,82
RKT	7	43	7,21
JAK	6	37	17,7
JAK	2	35	16,86
JAK	1	36	16,81
JAK	3	34	16,44
JAK	7	49	5,8
cAMP	6	15	29,7
cAMP	2	13	29,31
cAMP	1	13	29,31
cAMP	3	24	26,58
cAMP	7	34	12,56

Table 1. Results of using the consensus sequences for the database query. For the fourth chosen family, heksokinases, we didn't find any new hits.

4 Calculation of an Average Similarity Level of the Consensus Sequences

An average similarity level of a consensus sequence for a given parameter set was computed by carrying out all possible pairwise alignments within its protein family. Each particular alignment was then scored using a semihomology algorithm[1] along with two other methods, computing a statistical significance value of obtaining a given number of identities, and to analyze the identity distribution[4]. The semihomology algorithm is based on a three dimensional diagram of all possible transitions and transversions between residues for all existing codons, and claims a single mutation as the main cause of evolutionary changes. Statistical significance of the number of the identities allows to compare results of the aligned sequences of different lengths. Estimation of the identity distribution, provides information about importance of the quality measure (briefly speaking: more regularity - more importance).

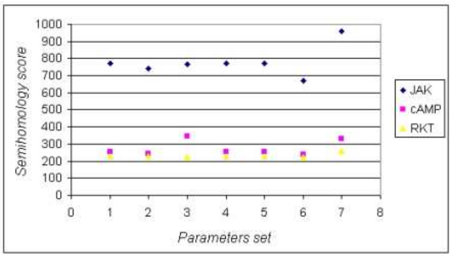

Figure 1. An average semihomology score for given parameters set.

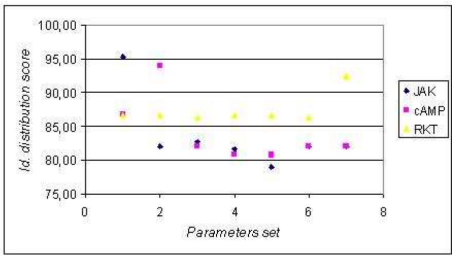

Figure 2. An average of the normalized identity distribution score
for given parameter sets.

5 Concluding Remarks

Some correlations between the parameter sets and the 'scores' were obtained for the consensus sequences. A most optimal set of the parameters for the studied estimation procedures couldn't be determined. Further studies using a larger group of protein families should provide more convincing results.

Acknowledgments

These studies were supported by the European CoE MAMBA and National CoE BioExploratorium.

References

1. Leluk, J.
A new algorithm for analysis of the homology in protein primary structure,
Computers & Chemistry **22**, 123–131 (1998).
2. Leluk, J., Fogtman, A., Lesyng, B.
Construction of Consensus Sequences of B-Spectrin Protein Family with variable Threshold Parameters and Validation of the Applied Approach, (2005).
3. Altschul, S.F. Gish, W., Miller, W., Myers, E.W., and Lipman, D.J.
Basic local alignment search tool,
J. Mol. Biol. **215**, 403-410 (1990).
4. Leluk, J., Mikoajczyk, A.
A new approach to sequence comparison and similarity estimation, XXIX Kongres FEBS, 26.06 -1.07.2004, Warszawa,
Eur. J. Biochem. **271**, Supplement, 29 (2004).