Bioinformation Volume 5

open access

www.bioinformation.net Issue 3

Software

SeqCalc: A portable bioinformatics software for sequence analysis

Dhandapani Vignesh[1], Paul Parameswari[1], Kim Hae Jin[2], Lim Yong Pyo[1]*

[1]Molecular Genetics and Genomics Laboratory, College of Agriculture and Life sciences, Chungnam National University, Daejeon, Rep. of Korea – 305-764; [2] Ensoltek Co., LTD, 925 Tamnip-dong, Yuseong-gu, Daejeon, Rep. of Korea - 305-510; Yong Pyo Lim- Email: yplim@cnu.ac.kr, Phone: +82-42-8215739, Fax: +82-42-8218847 *Corresponding author

Received August 4, 2010; Accepted August 20, 2010; Published September 20, 2010

Abstract
Rapid genome sequencing enriched biological databases with enormous sequence data. Yet it remains a daunting task to unravel this information. However experimental and computational researchers lead their own way in analyzing sequence information. Here we introduce a standalone portable tool named "SeqCalc" that would assist the research personnel in computational sequence analysis and automated experimental calculations. Although several tools are available online for sequence analysis they serve only for one or two purposes. SeqCalc is a package of offline program, developed using Perl and TCL/Tk scripts that serve ten different applications. This tool would be an initiative to both experimental and computational researchers in their routine research. SeqCalc is executable in all windows operating systems.

Availability: SeqCalc can be freely downloaded at http://code.google.com/p/seqcalc

Keywords: Sequence analysis, SeqCalc, portable, software, bioinformatics.

Background
Genomic analysis gained momentum through genome projects and sequencing of model organisms. However, the rhythm of genome analysis is not matching with the rhythm of genome sequencing. This leads to the over accumulation of data in the repository. Initial interpretation of nucleotide and protein sequences leads to the identification of hidden genetic and genomic links. Experimental genome analysis is massive process and thus necessitates the demand to develop computational tools for predicting the sequences. We have numerous online software to support nucleotide and protein analysis **[1]**. In this work we introduce a new tool Sequence Calculator "SeqCalc" which is efficient in ten different ways. It can be used to perform various nucleotide, protein sequence analysis and also in automated experimental calculation. SeqCalc is developed using Perl **[2]** as scripting language and TCL/Tk widgets **[3]** was used for designing the graphical user interface.

This new tool provides the user with the basic requirements in his research all in one single window. It intends to satisfy the user with his maximum requirement. Its structure, application and memory makes it comfortable and easy for its users.

Methodology:
SeqCalc is independent software that provides a platform for various types of study as outlined in **Figure 1**. The software is subdivided into 3 major components that are tagged as their function called nucleotide, protein and common analyzers. The tool contains open reading frame identification, aliphatic, hydropathy and instability index calculator, codon calculator, protein secondary structural elements extractor, DNA melting and annealing temperature calculator, random protein, nucleotide generator, statistics calculator, complimentary and reverse complimentary generator, DNA transcript and translator and iso-electric point calculator. Scientifically accepted formulae and methods of calculations are attached in detail with the supplementary material.

First, we set few samples that will be re-used throughout this investigation. We selected nucleotide sequence of ubiquinol-cytochrome c reductase complex protein UQCR10 of *Bos Taurus* **[4]**, protein sequence of human insulin and from NCBI **[5]** and a new crystal form of LYS48-linke diubiquitin of *Bos taurus* for testing the efficiency of the tool. SeqCalc accepts fasta or raw format sequences as input. Besides, it requires PDB file as input for PDB fragmentor. The program also needs setting numeric input parameters for random sequence generator.

Result and Validation
SeqCalc is divided into three major regions as mentioned above. The nucleotide and protein calculators are sequence specific. Common calculator contains features that are shared by both protein and nucleotide sequences and they alter their calculations based on the user selection.

Nucleotide calc
The Nucleotide calculator consists of five categories in it. They are codon statistics calculator, sequence melting and annealing temperature calculator, complimentary and reverse complimentary convertor, DNA to protein convertor and ORF finder. The first option, codon statistics calculator is designed for the purpose of analyzing the frequency of codon used by determining each codons amino acid, number for hits and investigates the frequency per thousand base pairs. It converts the given sequence to its triplet code of amino acid and then calculates their occurrence. The melting and annealing temperature calculator analyses by taking the count of GC and AT content in the given sequence by using the formula given by Rychlik W, et al., **[6]**. The complementary option in this tool helps in obtaining the compliment and the reverse compliment of large genome sequence. Transcript and translate tool is used to retrieve transcribed mRNA and the translated protein sequences from the given input sequence. Open reading frame identification in a sequence is helpful in gene coding sequence identification. "*" represents the start codon and "/" indicates the stop codon in the output of ORF finder.

Protein calc
The Protein calculator is subdivided into 3 major subdivisions. They are Protein secondary structure extractor, Iso-electric point calculator and Protein properties calculator. The protein secondary structure extractor is used to extract the amino acids that encode the secondary structural regions of protein from a given PDB file. It also extracts each chain, α-helix and β-strand sequences in FASTA format. From the extracted secondary structural sequence it identifies the type of helix, the sense of strand with respect to the previous position and the percentage of occurrence from the chain sequence in the header line of FASTA **[7]**. The iso-electric point calculator's algorithm is developed based on the expasy pKa values **[8]**. Protein properties option is used to calculate properties like aliphatic index,

extinction coefficient, molecular weight and grand average hydropathy of the protein from the given protein sequence. In the case of sample report shown in **Figure 2**, algorithms of SeqCalc are highly accurate. The efficiency of SeqCalc is compared against the online expasy tool. The graph **(Figure 3)** correlates the results of SeqCalc against the expasy tools.

Common calc

The common calculator is used to perform calculations that are shared by both protein and nucleotide sequences. They have the two different tools in them random nucleotide, protein sequence generator and protein, nucleotide sequence statistics calculator. Tools have been modeled to produce random DNA and protein sequences. Each time this tool generates different random sequences according to the parameters given in the tool box. To generate random DNA, composition of each base pair, amino acid or the length of sequence should be given as input. The sequence statistics calculator accepts the nucleotide or protein sequence then calculates their molecular weight and the number and percentage of occurrences of each residue from the given sequence. The program will ignore numbers, spaces or characters which do not correspond with one of the 4 nucleotide bases. The parameters computed in the protein statistics include the molecular weight, amino acid composition, atomic composition and extinction coefficient. All these parameters in the scripts are calculated using universally accepted formulas and have been executed successfully.

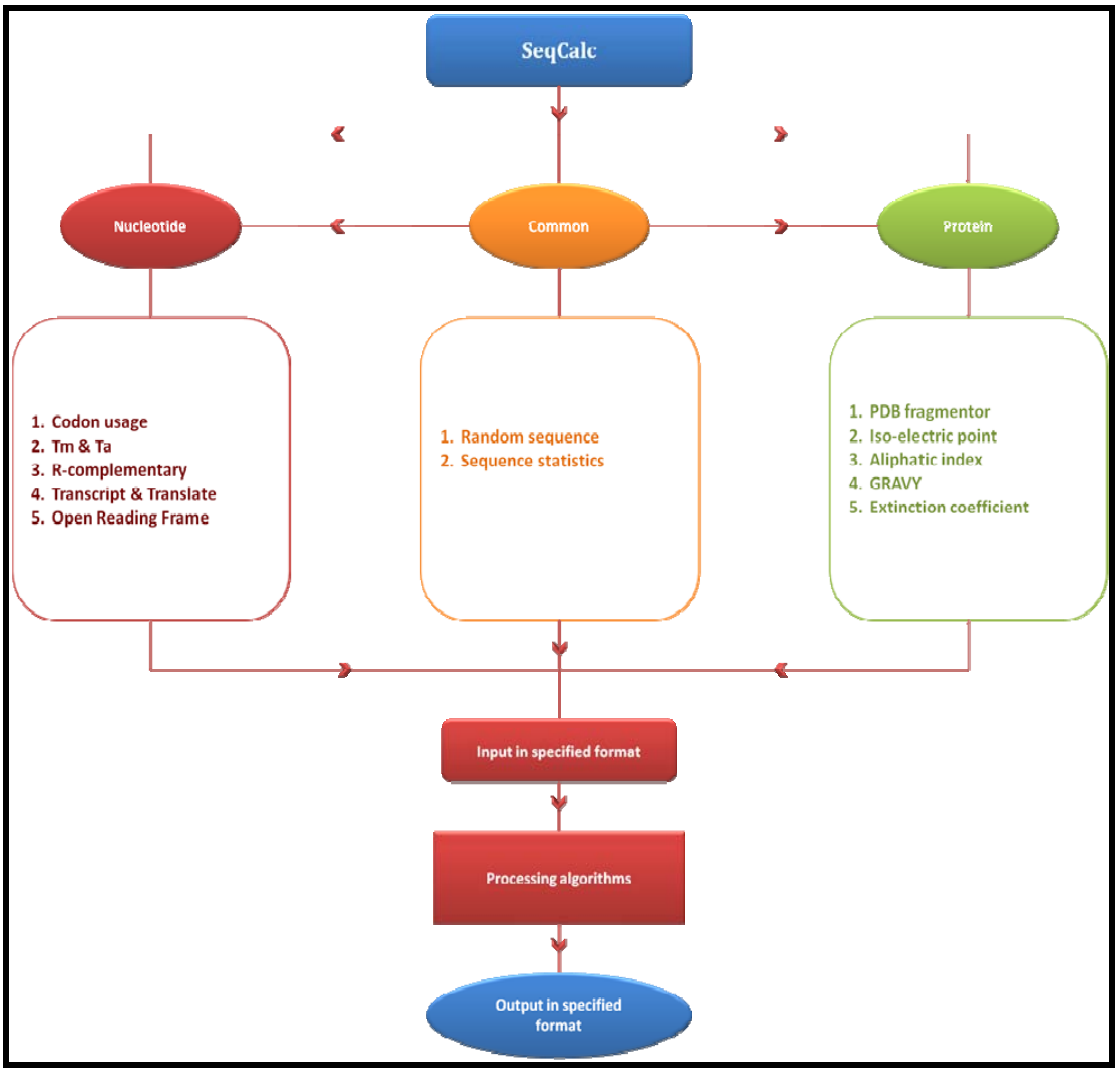

Figure 1: Schematic representation and application of SeqCalc.

Bioinformation Volume 5

open access

www.bioinformation.net Issue 3

Software

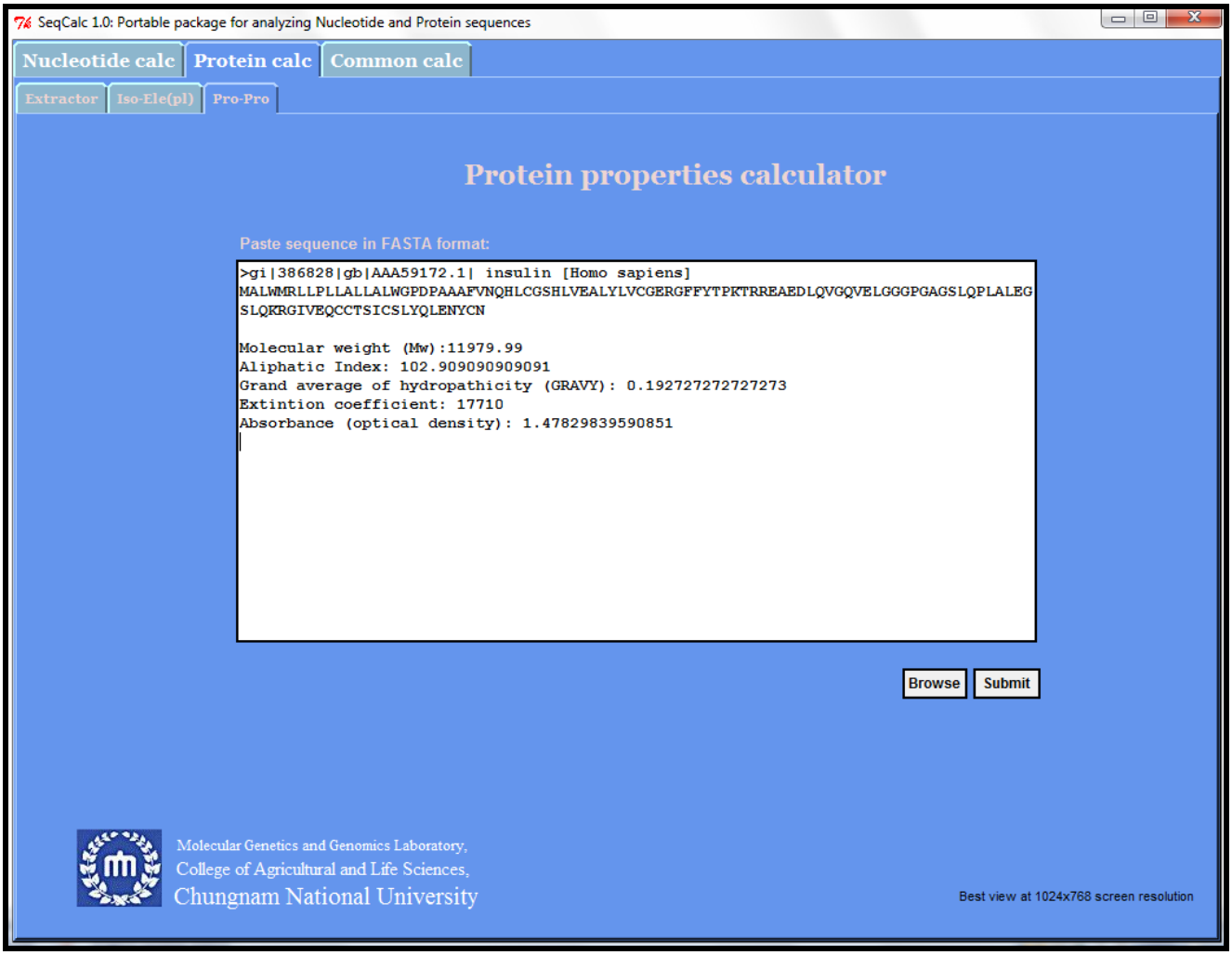

Figure 2: A snapshot of SeqCalc tool and protein properties calculator output are shown. Each calculator can be accessed by clicking on the named tabs.

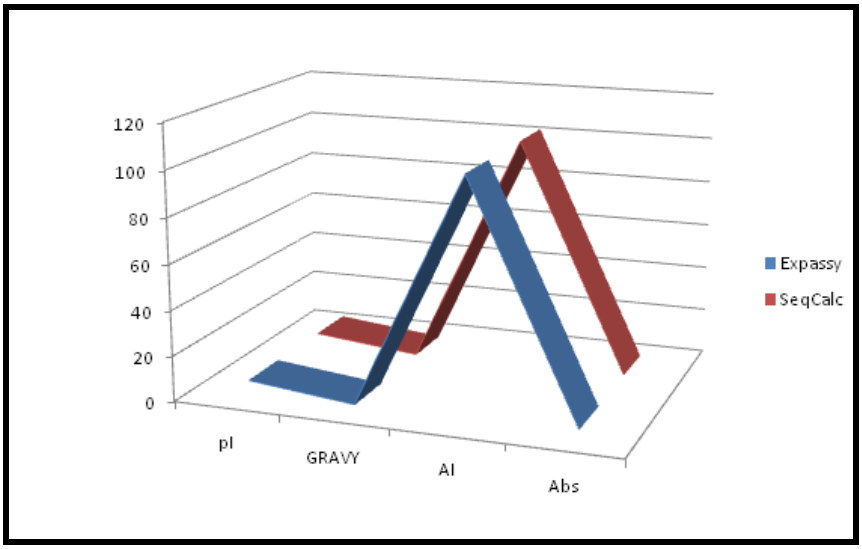

Figure 3: Efficiency of the protein property application in SeqCalc is plotted against the online expasy prot-param tool. Properties like Iso-electric point (pI), Grand average hydropathy (GRAVY), Aliphatic index (AI) and Absorbance (Abs) were considered and similar results were obtained respectively.

Bioinformation Volume 5

open access

www.bioinformation.net Issue 3

Software

Conclusion

SeqCalc is independent software that can be used in all types of windows operating system. It contains ten commonly used applications by research personnel. The software mainly aims to satisfy the users with their maximum requirement. Various sequence analysis and automated experimental calculators are included in the tool. Its structure, application and memory size it makes comfortable and easy for users. We believe this new tool will be a one stop tool and an initiative to both computational and experimental researchers in their routine research work. In future, we plan to further update the tool with more applications in a single window.

Acknowledgement

This tool is developed under "Human Resource Development Center for Economic Region Leading Industry" project, supported by the Ministry of Education, Science & Technology (MEST) and the National Research Foundation of Korea (NRF).

References

[1] http://www.bioinformatics.org/sms2/

[2] http://www.perl.org

[3] http://www.tcl.tk

[4] http://www.ncbi.nlm.nih.gov/nuccore/NM_001113723.1

[5] http://www.ncbi.nlm.nih.gov/protein/AAA59172.1

[6] W Rychlik *et al*. *Nucleic Acids Res*. **17**: 8543 (1989) [PMID: 2587212].

[7] D Vignesh *et al*. *Bioinformation* **4(9)**:390 (2010).

[8] K Joanna *et al*. *BMC Genomics* **8**:163 (2007) [PMID: 17565672].

Edited by P. Kangueane

Citation: Vignesh *et al.* Bioinformation 5(3): 85-88 (2010)