BIOINFORMATICS APPLICATIONS NOTE Vol. 25 no. 15 2009, pages 1976–1977
doi:10.1093/bioinformatics/btp300

Structural bioinformatics

Ulla: a program for calculating environment-specific amino acid substitution tables

Semin Lee* and Tom L. Blundell

Department of Biochemistry, University of Cambridge, 80 Tennis Court Road, Old Addenbrooke's Site, Cambridge CB2 1GA, UK

Received on March 17, 2009; revised on April 18, 2009; accepted on April 28, 2009

Advance Access publication May 5, 2009

Associate Editor: Anna Tramontano

ABSTRACT

Summary: Amino acid residues are under various kinds of local environmental restraints, which influence substitution patterns. Ulla,[1] a program for calculating environment-specific substitution tables, reads protein sequence alignments and local environment annotations. The program produces a substitution table for every possible combination of environment features. Sparse data is handled using an entropy-based smoothing procedure to estimate robust substitution probabilities.

Availability: The Ruby source code is available under a Creative Commons Attribution-Noncommercial License along with additional documentation from http://www-cryst.bioc.cam.ac.uk/ulla.

Contact: semin@cryst.bioc.cam.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

In the evolution of proteins, individual amino acid residues are under various kinds of local environmental restraints such as secondary structure type, solvent accessibility and hydrogen bonding patterns. Previous study of amino acid substitutions as a function of local environment has showed that there are clear differences among substitution patterns under various environmental restraints (Overington *et al.*, 1992). The unique patterns of amino acid substitutions have been successfully exploited to predict the stability of protein mutants (Topham *et al.*, 1993), to identify potential interaction sites (Chelliah *et al.*, 2004; Gong and Blundell, 2008) and to detect remote sequence-structure homology (Chelliah *et al.*, 2005).

However, estimating amino acid substitution probabilities is not a trivial problem, especially when there are a very small number of observations in specific combinations of environments. To cope with the sparse data problem, an algorithm was developed by Sali (1991) as an extension of the method used by Sippl (1990) to derive robust potentials of mean force. Several variants of the generalized procedure such as Makesub (Topham *et al.*, 1993) and SUBST (Mizuguchi, unpublished results) have been subsequently implemented for smoothing substitution probabilities. Nevertheless, each lacks crucial features implemented in the other, and they use slightly different procedures for smoothing substitution probabilities, which may lead to very different amino acid substitution matrices.

To overcome these problems, we developed Ulla, a program for calculating environment-specific substitution tables (ESSTs), to unify all the major features of the previously developed programs and to provide additional functionalities. The program also generates heat maps from substitution tables to visualize the degree of conservation of amino acids under the environmental restraints.

2 DESCRIPTION

Ulla reads multiple sequence alignments and annotations for local environments in JOY template format (Mizuguchi *et al.*, 1998a). Users can provide their own definition of environment features, and an environment feature can be constrained to count substitutions only when the environment of residues is conserved. Ulla also supports confining percent identity (PID) range of sequence pairs to be considered and uses BLOSUM-like weighting scheme (Henikoff and Henikoff, 1992) to minimize sampling bias from highly similar sequences.

Ulla uses entropy-based smoothing procedures to reduce problems caused by sparse data. It is an iterative procedure that estimates probability distribution by perturbing the previous probability distribution with the successive measurement (Sali, 1991; Sippl, 1990). Hence, starting from a uniform frequency distribution, the estimated probability distribution at each step serves as an approximation for the next probability distribution (see Supplementary Material for details).

3 EXAMPLE USAGE

As an illustration, we generate ESSTs from HOMSTRAD alignments (Mizuguchi *et al.*, 1998b) with environment feature definitions of secondary structure type and solvent accessibility (Fig. 1a):

```
# name of feature (string);\\
# values adopted in .tem (alignment) file (string);\\
# class labels assigned for each value (string);\\
# constrained or not (T or F);\\
# silent (used as masks)? (T or F)
secondary structure and phi angle;HEPC;HEPC;F;F
solvent accessibility;TF;Aa;F;F
```

*To whom correspondence should be addressed.

[1]Ulla is a traditional Korean percussion instrument.

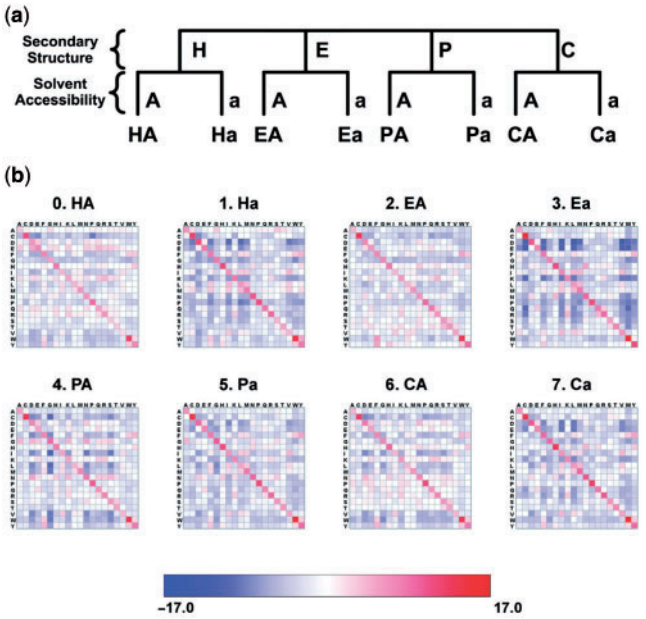

Fig. 1. Environment feature combinations and ESST generation. (**a**) The environment features are secondary structure type (H: helix, E: beta sheet, P: positive phi, C: coil) and solvent accessibility (A: solvent accessible, a: solvent inaccessible). Eight sets of combinations of environment features are generated. (**b**) Heat maps from each of resultant ESSTs. Blue to red indicates log-odds ratio of substitution probabilities.

Actual annotations for the environment features need to be provided in PIR format:

```
>P1;1mnma
sequence
QKERRKIEIKFIENKTRRHVLLVVSETGLVYTFSTPKFEPIVTQQEGR…*
>P1;1egwa
sequence
--GRKKIQITRIMDERNRQVTFTKRKFGLMKKAYELSVLCDCEIALII…*
>P1;1mnma
secondary structure and phi angle
CPCCCCCCCCCCCCCHHHHHHHHHHHHHHHHHHHHHHHHHHHPCCCEEEE…*
>P1;1egwa
secondary structure and phi angle
--CCCCCCCCCCCCCHHHHHHHHHHHHHHHHHHHHHHHHHCPCCCEEEE…*
>P1;1mnma
solvent accessibility
TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT…*
>P1;1egwa
solvent accessibility
--TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTFTTTTTTTTTTT…*
...
```

JOY (Mizuguchi *et al.*, 1998a) is useful to annotate the alignments with the structural environments, but Ulla recognizes any environment feature definition which conforms to the format above. Paths for an environment definition file and a file containing the list of environment feature annotated alignments are given to Ulla as input:

```
$ ulla -c feature.def -l alignments.lst
```

Ulla produces three different types of substitution tables: raw counts tables, substitution probability tables and log-odds ratio tables. Heat maps also can be generated to visualize resultant substitution tables (Fig. 1b).

4 CONCLUSION

Ulla generates ESSTs from a sparse data set using entropy-based smoothing procedures. It allows us to conduct analyses of amino acid substitution patterns under various environmental restraints. The resultant ESSTs can be exploited in many ways such as binding site prediction, remote homology detection, and protein stability estimation.

Ulla is publicly available on the web site http://github.com/semin/ulla, where the code is maintained in a Git repository, and its pre-built RubyGems package can be obtained from http://rubyforge.org/projects/ulla.

ACKNOWLEDGEMENTS

We thank Juok Cho for statistical advice; Dan Bolser and Duangrudee Tanramluk for review of the manuscript; Richard Bickerton and Bernardo Ochoa for thorough beta testing.

Funding: Mogam Science Scholarship Foundation (to S.L., partial); The Wellcome Trust (to T.L.B.)

Conflict of Interest: none declared.

REFERENCES

Chelliah,V. *et al.* (2004) Distinguishing structural and functional restraints in evolution in order to identify interaction sites. *J. Mol. Biol.*, **342**, 1487–1504.

Chelliah,V. *et al.* (2005) Functional restraints on the patterns of amino acid substitutions: application to sequence-structure homology recognition. *Proteins*, **61**, 722–731.

Gong,S. and Blundell,T.L. (2008) Discarding functional residues from the substitution table improves predictions of active sites within three-dimensional structures. *PLoS Comput. Biol.*, **4**, e1000179.

Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.

Mizuguchi,K. *et al.* (1998a) JOY: protein sequence-structure representation and analysis. *Bioinformatics*, **14**, 617–623.

Mizuguchi,K. *et al.* (1998b) HOMSTRAD: a database of protein structure alignments for homologous families. *Protein Sci.*, **7**, 2469–2471.

Overington,J. *et al.* (1992) Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds. *Protein Sci.*, **1**, 216–226.

Sali,A. (1991) Modelling three-dimensional structure of proteins from their sequence of amino acid residues. PhD Thesis, University of London, London.

Sippl,M.J. (1990) Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.*, **213**, 859–883.

Topham,C.M. *et al.* (1993) Fragment ranking in modelling of protein structure. Conformationally constrained environmental amino acid substitution tables. *J. Mol. Biol.*, **229**, 194–220.