Bioinformatics Advance Access published April 10, 2016

Application Note

INPS-MD: a web server to predict stability of protein variants from sequence and structure

Castrense Savojardo[1]^, Piero Fariselli[2]^, Pier Luigi Martelli[1*] and Rita Casadio[1,3]

[1]Biocomputing Group, CIG, Interdepartmental Center «Luigi Galvani» for Integrated Studies of Bioinformatics, Biophysics and Biocomplexity, University of Bologna, Italy; [2]BCA, University of Padova, Italy; [3]Interdepartmental Center «Giorgio Prodi» for Cancer Research, University of Bologna,Italy.

*To whom correspondence should be addressed.

^Authors equally contributed to the paper

Associate Editor: Prof. Anna Tramontano

Abstract

Motivation: Protein function depends on its structural stability. The effects of single point variations on protein stability can elucidate the molecular mechanisms of human diseases and help in developing new drugs. Recently, we introduced INPS, a method suited to predict the effect of variations on protein stability from protein sequence and whose performance is competitive with the available state-of-the-art tools.

Results: In this paper, we describe INPS-MD (Impact of Non synonymous variations on Protein Stability-Multi-Dimension), a web server for the prediction of protein stability changes upon single point variation from protein sequence and/or structure. Here, we complement INPS with a new predictor (INPS3D) that exploits features derived from protein 3D structure. INPS3D scores with Pearson's correlation to experimental ΔΔG values of 0.58 in cross validation and of 0.72 on a blind test set. The sequence-based INPS scores slightly lower than the structure-based INPS3D and both on the same blind test sets well compare with the state-of-the-art methods.

Availability: INPS and INPS3D are available at the same web server: http://inpsmd.biocomp.unibo.it.

Contact: gigi@biocomp.unibo.it

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

Stability of protein variants may or not be different from wild type. The information is relevant to understand the relation between protein variants and insurgence of diseases (Lu et al., 2014; Ashley, 2015). Several methods have been developed so far to predict stability change in protein variants, either based on protein sequence or structure features (Huang et al., 2007; Teng et al., 2010; Pires et al., 2014; Leimer et al., 2015; Folkman et al., 2016). Recently we developed INPS (Fariselli et al., 2015), a sequence-based method devised to prediction of protein stability change (ΔΔG) upon single-point variations, well comparing with the state-of-the-art methods. Here we present a web server, INPS-MD that includes the sequence-based INPS, and a new method INPS3D, exploiting descriptors derived from the protein 3D structure. Benchmark results, performed on two experimentally derived datasets of variations, show that the structure-based INPS-3D performs better than the sequence-based INPS when predicting the change in protein stability upon variation. INPS-MD is available at http://inpsmd.biocomp.unibo.it.

2 INPS3D

INPS (Fariselli et al., 2015) is based on descriptors extracted from the protein sequence. Briefly, seven features are used to encode a single-point mutation: (i) the substitution score w →m derived from the Blosum62 matrix (Henikoff and Henikoff, 1992); (ii) the Kyte-Doolittle hydrophobicity (Kyte and Doolittle, 1982) scores of native and changed residues (2 descriptors); (iii) the mutability index of the native residue (Dayhoff et al., 1978); (iv) the molecular weights of native and changed residues (2 descriptors); (v) the difference in the alignment score between the native and variant sequences and a HMM, encoding evolutionary information of the wild type sequence (Fariselli et al., 2015). Sequence descriptors are mapped to ΔΔG values using a Support Vector Regression (Chang et al., 2011) with a Radial Basis Function (RBF) ker-

© The Author (2016). Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

Downloaded from http://bioinformatics.oxfordjournals.org/ at University of California, San Diego on April 12, 2016

nel. In INPS3D, the set of INPS descriptors includes also features derived from protein 3D structure. In particular, the following two additional structure-based descriptors are considered: (i) one descriptor corresponding to the Relative Solvent Accessibility (RSA) of the native residue. The absolute solvent accessibility is computed from the PDB file using the DSSP method (Kabsch and Sander, 1983) and then normalized as previously described (Rost and Sander, 1994); (ii) another descriptor encodes for the local Energy Difference (ED) between native and mutated protein structures. The energy is computed by means of a contact potential previously described (Bastolla et al., 2001). Given the set N of neighbors of the changed residue (two residue are considered in contact if the minimal distance between all atoms is <5Å), the contact energy difference is scored as:

$$\sum_{r \in N} P(r, w) - P(r, m) \tag{1}$$

where w and m are respectively the native and the mutated residues, and P is the contact potential defined over pair of residues (Bastolla et al., 2001).

2.1 Web server usage

INPS and INPS3D are included in INPS-MD, available at http://inpsmd.biocomp.unibo.it. Here, the user can select the sequence-based INPS or the structure-based INPS3D, when the protein structure is available. In both cases, the user must provide input files and parameters through the input submission form: 1) the query protein in the form of a single FASTA file, in the case of INPS or a valid PDB file in the case of INPS3D. In the latter case, the target PDB chain ID must be specified; 2) a single file containing the list of point variations relative to the sequence or PDB chain. Upon submission, the server provides the user with a universal job identifier that can be thereafter used to retrieve results. For each variation listed in the input file, the server computes protein descriptors and performs ΔΔG predictions. Upon job completion, results can be visualized online or downloaded in plain-text format.

Table 1. Performance of INPS and INPS3D

Method	S2648 dataset Corr / SE	Blind test Corr / SE	P53 blind test Corr / SE
INPS	0.52 / 1.26	0.68 / 1.26	0.69 / 1.45
INPS+ED	0.52 / 1.25	0.68 / 1.26	0.71 / 1.45
INPS+RSA	0.54 / 1.24	0.70 / 1.18	0.74 / 1.36
INPS3D	0.58 / 1.20	0.72 / 1.15	0.76 / 1.35

Corr=Pearson's correlation coefficient between predicted and experimental DDG values. SE=standard error in DDG prediction (kcal/mol); ED=Energy Difference; RSA=Relative Solvent accessibility.

3 Results and conclusion

INPS and INPS3D are benchmarked on three different sets previously released (Pires et al., 2014): (i) the S2648 dataset, comprising 2648 single-point mutations in 132 proteins derived from ProTherm (Kumar et al., 2006), (ii) a subset of S2648 used as blind test set comprising 351 variations in 60 proteins and (iii) a dataset of 42 variations within the DNA binding domain of the tumor suppressor protein P53. Table 1 lists results obtained for INPS and INPS3D with a cross validation procedure on the first set and adopting the remaining two as blind test sets. INPS3D outperforms INPS, achieving a Pearson's correlation 7 percentage points higher than INPS and a standard error of 0.1 kcal/mol smaller in ΔΔG prediction. The relative contribution of the two structural features indicates that RSA is more informative than ED. Their combination leads to a further improvement in the performance (Pearson's correlation is 0.58). INPS and INPS3D predictions strongly correlate (Fig 1S) and well compare with the state-of- the-art predictors when tested on the same datasets (see Table 1S in Supplementary Material). INPS3D input includes information derived from multiple sequence alignment and its performance is dependent on the number of aligned sequences. INPS3D is less sensitive than INPS, and even when the number of aligned sequences is less than 100, its Pearson's correlation is still 0.5 (Fig. 2S). S2648 dataset contains some redundancy (44 proteins out of the 132 share more than 25% identity). In the most stringent per-protein split (Pires et al. 2014), 16 pairs of proteins have sequence similarity between the training and testing sets. However, INPS and INPS3D seem unaffected by this redundancy, since no differences in the cross-validation performances are detected when the sequence similarity is removed. These new cross-validation sets, together with the BLAST outputs, are now available as new material in the web server page. INPS-MD is a new web server that integrates information from sequence and structure to predict protein stability perturbation upon residue variations and allows the prediction in multi-dimensions.

Funding

Funding: PRIN 2010-2011 project 20108XYHJS (to P.L.M.) (Italian MIUR); COST BMBS Action TD1101 and Action BM1405 (European Union RTD Framework Program, to R.C); PON projects PON01_02249 and PAN Lab PONa3_00166 (Italian Miur to R.C. and P.L.M.); FARB UNIBO 2012 (to R.C.)

Conflict of Interest: none declared.

References

Ashley, E.A. (2015) The precision medicine initiative: a new national effort, *JAMA*, 313, 2119-2120.

Bastolla, U. *et al.* (2001) How to guarantee optimal stability for most representative structures in the protein data bank, *Proteins*, **44**, 79-96.

Chang, C.C. *et al.* (2011) LIBSVM: a library for support vector machine. *ACM Transactions on Intelligent Systems and Technology*, **2**(27), 1-27.

Dayhoff, M.O. *et al.* (1978) A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure*, **5**(3).

Fariselli, P. *et al.* (2015) INPS: predicting the impact od non-synonymous variations on protein stability from sequence. *Bioinformatics*, **31**, 2816-2821.

Folkman L., *et al.* (2016) EASE-MM: Sequence-based prediction of mutation-induced stability changes with feature-based multiple models. *J. Mol. Bio.*, Jan 22.

Henikoff, S. and Henikoff, J.G. (1992) Amino acid substitution matrices from protein blocks. *PNAS*, **89**, 10915-10919.

Huang,L.T. et al. (2007) iPTREE-STAB: interpretable decision tree based method for predicting protein stability changes upon mutations. Bioinformatics, 23, 1292–1293.

Kabsch, W. and Sander, C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, *Biopolymers*, **22**, 2577-2637.

Kyte, J. and Doolittle, R.F. (1982) A simple method for displaying the hydrophatic character of a protein. *J. Mol. Bio.*,**157**, 105-132.

Kumar,M.S. et al. (2006) Protherm and pronit: thermodynamic databases for proteins and protein-nucleic acid interactions. Nucleic Acids Res., **34**, D204-D206.

Laimer, J et al., (2016) MAESTROweb: a web server for structure based protein stability prediction. Bioinformatics, 2016 Jan 6.

Lu *et al.* (2014) Personalized medicine and human genetic diversity. Cold Spring Harb *Perspect. Med.*, 4, a008581.

Pires *et al.* (2014) mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics*. 30:335-42.

Rost, B. and Sander, C. (1993) Conservation and prediction of solvent accessibility in protein families, *Proteins*, **20**, 216-226.

Teng, S. et al. (2010) Sequence feature-based prediction of protein stability changes upon amino acid substitutions. BMC Genomics, 11, 5.

Downloaded from http://bioinformatics.oxfordjournals.org/ at University of California, San Diego on April 12, 2016