Syst Synth Biol (2014) 8:337–341
DOI 10.1007/s11693-014-9156-4

METHODOLOGY ARTICLE

bPE toolkit: toolkit for computational protein engineering

Gaurav Jerath · Prakash Kishore Hazam · Vibin Ramakrishnan

Received: 28 July 2014 / Revised: 7 October 2014 / Accepted: 8 October 2014 / Published online: 19 October 2014

Abstract We present a computational toolkit consisting of five utility tools, for performing basic operations on a protein structure file in PDB format. The toolkit consists of five different programs which can be integrated as part of a pipeline for computational protein structure characterization or as a standalone analysis package. The programs include tools for chirality check for amino acids (ProChiral), contact map generation (CoMa), data redundancy (DaRe), hydrogen bond potential energy (HyPE) and electrostatic interaction energy (EsInE). All programs in the toolkit can be accessed and downloaded through the following link: http://www.iitg.ac.in/bpetoolkit/.

Keywords Toolkit · Protein engineering · Protein design · Protein structure

Introduction

Recently, experimentalists have started using computational methods extensively in formulating and fine tuning their research plans. Most of these groups employed in protein engineering research have shared their interest in using certain basic tools supporting their research, yet not part of a fairly large program suite which would be difficult for them to use independently without the help of a computational biologist. Additionally, we would like to have few independent tools that can be incorporated in a pipeline of various operations with protein structure files, such that they can create their own computational work flow. To accomplish these objectives, we present a protein engineering toolkit, consisting of five programs that can perform basic operations in a protein structure file.

G. Jerath · P. K. Hazam · V. Ramakrishnan (✉)
Department of Biotechnology, Indian Institute of Technology, Guwahati 781039, India
e-mail: vibin@iitg.ernet.in; vibin@iitg.ac.in

Computational methods and their validation

The theoretical background of computational programs and their validation is described as follows. The science and logic behind their design has also been briefly illustrated.

ProChiral: chirality check program

Chirality is a property of a molecule which doesn't have a super-imposable mirror image of itself. All amino acids, except glycine are chiral with the CA atom being the chiral center. The chirality of an amino acid can be determined by checking the orientation of the side chain with respect to the protein/peptide backbone, i.e. the orientation of the CA atom with respect to the backbone. This can be done by calculating the improper dihedral angle formed by the normals from the planes involving N-CA-C and C-CA-CB. A positive value for the angle signifies the Levorotatory form of the amino acid whereas, a negative value depicts the Dextrorotatory form. The chirality program gives the orientation of the amino acid residue side chain as Dextro or Levo-rotatory.

ProChiral was tested using the existing protein data bank (PDB) structure 1GRM.pdb, which has a mixture of L and D-amino acids. The program was also applied to other PDB structure files which are homo-chiral in nature. Furthermore, few in-house generated hetero-chiral PDB files were also tested with the program. The program successfully identified the chiral orientation of the CA atom with 100 % accuracy for all the structures tested. Additionally, the structures were

Table 1 a) The prediction of handedness for 1GRM.pdb by the HandleCheck and ProChiral. b) The prediction results for in-house generated sequences with different stereochemistry using HandleCheck and ProChiral

a)													
1GRM sequence	V	G	A	L	A	V	V	V	W	L	W	L	W
Stereochemical sequence	D		L	D	L	D	L	D	L	D	L	D	L
HandleCheck	D		L	D	L	D	L	D	L	D	L	D	L
Pro Chiral	D		L	D	L	D	L	D	L	D	L	D	L

b)										
Amino Acid Sequence	R	R	R	R	R	R	R	R	R	R
Stereochemical sequence	L	L	D	L	D	D	L	L	D	D
HandleCheck	L	L	D	L	D	D	L	L	D	D
Pro Chiral	L	L	D	L	D	D	L	L	D	D
Amino Acid Sequence	T	R	P	A	N	N	A	P	R	T
Stereochemical sequence	L	L	L	L	L	D	D	D	D	D
HandleCheck	L	L	L	L	L	D	D	D	D	D
Pro Chiral	L	L	L	L	L	D	D	D	D	D
Amino Acid Sequence	A	A	A	A	A	A	A	A	A	A
Stereochemical sequence	D	D	D	D	D	L	L	L	L	L
HandleCheck	D	D	D	D	D	L	L	L	L	L
Pro Chiral	D	D	D	D	D	L	L	L	L	L
Amino Acid Sequence	R	K	1	L	W	W	W	K	K	R
Stereochemical sequence	D	L	D	L	L	D	L	L	D	L
HandleCheck	D	L	D	L	L	D	L	L	D	L
Pro Chiral	D	L	D	L	L	D	L	L	D	L
Amino Acid Sequence	C	C	S	S	Q	Q	W	W	Y	Y
Stereochemical sequence	L	D	L	D	L	D	L	D	L	D
HandleCheck	L	D	L	D	L	D	L	D	L	D
Pro Chiral	L	D	L	D	L	D	L	D	L	D

also verified with the HandleCheck program available in the WHAT IF web-server (Rodriguez et al. 1998). The time taken by ProChiral for correctly identifying the chirality of the above mentioned structures was 0.01 s/structure. A detailed description of results is shown in Table 1.

CoMa: contact map program

Contact Maps are 2D graphical representations of a 3D protein structure that describe the spatial arrangements of amino acid residues with respect to each other. The two most common forms of such maps are based on either the CA–CA distance or the CB–CB (CA in case of glycine) distance and the cutoff varies from 6 to 12 Å for reporting a contact. The contact map can be analyzed for the identification of parallel, anti-parallel beta sheets and alpha helices. The utility can be extrapolated to the identification and/or prediction of various protein folds (Hamilton and Huber 2008; Vendruscolo et al. 1999). The contact map program reports a contact on the basis of the calculated distance between two CA atoms in a protein structure. The program is designed to have a maximum threshold of eight angstroms between two representative atoms of each residue for reporting a contact. Up to two immediate sequential neighbors (i, i ± 2) have been exempted from the contact list.

CoMa has two output modes, one produces a .xpm file for every structure that can be viewed in any image viewer or may further be converted to a .xps file via xpm2ps command available in GROMACS (Berendsen et al. 1995; Pronk et al. 2013) suite. The other format is the text output format, which enlists all the contacts for every residue against its number. The program was evaluated using the contact map generated by a similar program supplied with the CMView (Vehlow et al. 2011) program. The program successfully generated the contact map for ten PDB structures and all maps were identical with the maps generated by CMView. CoMa successfully mapped all the contacts for the said PDB structures at a rate of 0.03 s/structure. Each structure had sequence length in the range of 150–200 residues. The contact maps for 1AOE generated by both the programs are shown in Fig. 1.

DaRe: data redundancy and homology program

The foreseen utility behind the development of this program is to avoid duplication of protein structures or

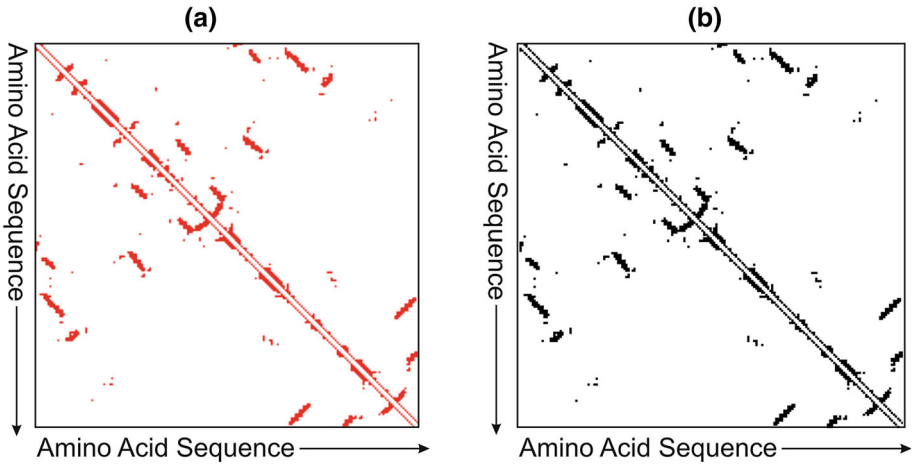

Fig. 1 The contact maps generated for 1AOE.pdb by **a** CMView program and **b** CoMa

Table 2 The list of PDB IDs for the human variant of of Dihydrofolate reductase enzyme used for validation of DaRe

PDB IDs				
1BOZ	1DHF	2DHF	1DLR	1DLS
1DRF	1HFP	1HFQ	1HFR	1KMS
1KMV	1MVS	1MVT	1OHJ	1OHK
1PD8	1PD9	1PDB	1S3U	1S3V
1S3W	1U72	1YHO	2C2S	2C2T

sequences while deducing analytical conclusions after performing computational or statistical methods. The program uses the global alignment algorithm (Needleman and Wunsch 1970) for determination of sequence identity among all pairs of sequences from the structure files present in the working directory. Thus, the program can also be utilized for checking the percentage sequence identity among a given set of structures without the additional requirement of downloading a separate sequence file from (Berman et al. 2000) PDB database or elsewhere. This is a utility designed to solve one of the biggest problems during handling of biological data, viz. data redundancy. This program truncates the redundant PDB structure files from a folder to a truncated PDB file folder, while keeping the largest sequence and other non-redundant files in the original folder. The redundancy is checked by calculating the percent identity within the sequences of structure files. The program first extracts the sequence from the PDB file by either reading the SEQRES entries or on its failure through the coordinates' list. Next, the sequence identity is calculated among all the files available in the folder which is also returned in the output. Finally, the structure with maximum sequence length among the redundant entries(more than 50 % sequence identity) are retained in the original folder while all the redundant files are moved in a folder of truncated PDBs and their list is also printed as output.

DaRe was evaluated by downloading 25 structure files of human variant of Dihydrofolate reductase enzyme from the PDB along with the PDB structure files for the same enzyme in other organisms. The human variants of the structure had more than 90 % similarity and the one with the largest sequence length was kept while the other smaller sequences were moved to the redundant PDB folder. For the above mentioned dataset, DaRe consumed 0.3 s to complete the redundancy check. The list of PDB IDs used can be seen in Table 2.

HyPE: hydrogen-bond potential energy program

Hydrogen Bonding is arguably one of the most important non-bonding interactions in bio-molecules. HyPE is designed to calculate the potential of hydrogen bonds formed between the N–H and C = O atoms in the backbone of a protein structure. The calculation of this potential is in accordance with the 12–10 model (Gordon et al. 1999). The program will run on a protein data file, with all hydrogen atom positions provided explicitly. The parameters followed for the prediction of a hydrogen bond are: (1) Maximum Cutoff distance of 0.35 nm between the donor and acceptor atoms, (2) The maximum angle cutoff of 30° for the acceptor–donor-hydrogen angle as illustrated in Fig. 2. The equation used for the calculation of the potential energy as described by Gordon et al. (Gordon et al. 1999) is as follows:

$$E_{HB} = D_0 \left[5 \cdot \left(\frac{R_0}{R} \right)^{12} - 6 \cdot \left(\frac{R_0}{R} \right)^{10} \right] cos^2\theta cos^2\Phi$$

where R_0 is the equilibrium distance, D_0 is the well depth and R is the interatomic distance between the donor and acceptor heavy atoms. The angles θ and Φ refer to the hydrogen–acceptor–base angle (where the base is the atom covalently attached to the acceptor) and the angle between

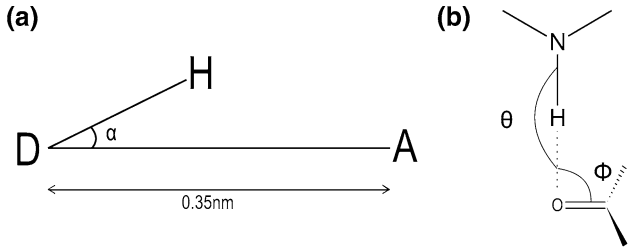

Fig. 2 **a** The figure illustrates the parameters followed for Hydrogen bond prediction, where α represents the hydrogen-donor–acceptor angle. **b** Angles θ and Φ are defined for the calculation of hydrogen bond interaction energy

Table 3 a) The coefficient of correlation (R) for the number of hydrogen bonds in the eight structures in a MD trajectory of each of the PDB structures with the total Hydrogen bond potential energy as calculated by HyPE, b) The coefficient of correlation (R) for the total Coulomb's energy for the eight structures in a MD trajectory of each of the PDB structures with the electrostatic interaction energy as calculated by EsInE

PDB ID	a) HyPE	b) EsInE
1DF7	−0.95	0.99
1JUV	−0.94	0.99
1VDR	−0.94	0.98
3DFR	−0.96	0.99
4DFR	−0.96	0.99
8DFR	−0.91	0.99

the normals of the planes defined by the six atoms attached to the two centers, respectively.

HyPE has been validated with the help of the g_hbond command of GROMACS, which was used to extract the values for the number of hydrogen bonds in eight different frames extracted during the energy minimization procedure of six protein structures. The number of bonds was correlated with the values of the potential energy for these frames from HyPE. The correlation values deviated from the bond number distribution as no bond had been pre-fixed with a defined energy and thus showed the sensitivity of program to take account of the distance and the orientation of the bond in consideration. The numbers of bonds predicted by the program were consistent for all the structures with the ones predicted by the GROMACS package. Calculations done by HyPE took an average 0.3 s per structure for the chosen set of six PDB structure files. The correlation coefficients for the structures used are shown in the Table 3a.

EslnE: electrostatic interaction energy

The electrostatic energy plays a vital role while considering the functional properties in protein design. These interactions have significance in the specificity of protein folding and functional interactions, rather than stability. The electrostatic interaction energy between two charges can be calculated as a derivative of Coulomb's law (Gordon et al. 1999) as

$$E_{EI} = 322.0637 \times \frac{(q_i \cdot q_j)}{\varepsilon \cdot R_{ij}}$$

where, q_i and q_j are the charges separated by a distance R_{ij} in a medium with dielectric constant ε.

The program has been coded to calculate the electrostatic interactions in a protein structure for a user defined dielectric constant. The program gives the total interaction energy for a protein structure as output.

EsInE has been verified by calculating the correlation between the interaction energy calculated by it and the total Coulomb's energy from the g_energy command of GROMACS. The values were calculated for eight frames extracted during the energy minimization procedure, repeated for six protein structures. EsInE took an average 2.69 s to calculate electrostatic interaction energy for each structure. The values of the correlation coefficients between the total Coulomb's energy (GROMACS) and the electrostatic interaction energy (EsInE) along with the PDB IDs of the structures are given in Table 3b.

Conclusions

Most of the programs described above, are available in various web based computational resources or as part of a fairly large program suite. We tried to have these programs under one source, so that it may be helpful for experimental scientists working in the area of protein engineering and design. Moreover, quite significantly, these tools may be included in an automated pipeline of tools along with other UNIX based programs. All users can, and in fact are requested to test the authenticity and utility by downloading these tools from our website. All programs are fully downloadable from the website http://www.iitg.ac.in/bpetoolkit/.

References

Berendsen HJC, van der Spoel D, van Drunen R (1995) GROMACS: a message-passing parallel molecular dynamics implementation. Comput Phys Commun 91:43–56

Berman HM, Westbrook J, Feng Z et al (2000) The protein data bank. Nucleic Acids Res 28:235–242

Gordon DB, Marshall SA, Mayot SL (1999) Energy functions for protein design. Curr Opin Struct Biol 9:509–513

Hamilton N, Huber T (2008) An introduction to protein contact prediction. Methods Mol Biol 453:87–104. doi:10.1007/978-1-60327-429-6_3

Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48:443–453

Pronk S, Pall S, Schulz R et al (2013) GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. Bioinformatics 29:845–854. doi:10.1093/bioinformatics/btt055

Rodriguez R, Chinea G, Lopez N, Pons T, Vriend G (1998) Homology modeling, model and software evaluation: three related resources. Bioinformatics 14:523–528

Vehlow C, Stehr H, Winkelmann M, Duarte JM, Petzold L, Dinse J, Lappe M (2011) CMView: interactive contact map visualization and analysis. Bioinformatics 27:1573–1574

Vendruscolo M, Najmanovich R, Domany E (1999) Protein Folding in Contact Map Space. Phys Rev Lett 82:656–659