BIOINFORMATICS APPLICATIONS NOTE Vol. 20 no. 12 2004, pages 1989–1991
doi:10.1093/bioinformatics/bth190

Defining 3D residue environment in protein structures using SCORPION and FORMIGA

R. H. Higa[1], A. G. Oliveira[1,2], L. G. Horita[1,2], R. T. Miura[1,2], M. K. Inoue[1,2], P. R. Kuser[1], A. L. Mancini[1], M. E. B. Yamagishi[1], R. C. Togawa[1,2] and G. Neshich[1,*]

[1]Núcleo de Bioinformática, Centro Nacional de Pesquisa Agropecuária, Empresa Brasileira de Pesquisa Agropecuária, Campinas, SP, Brazil and [2]Laboratório de Bioinformática, Embrapa/Recursos Genéticos e Biotecnologia, Empresa Brasileira de Pesquisa Agropecuária, Brasilia, DF, Brazil

Received on September 22, 2003; revised on January 25, 2004; accepted on February 20, 2004
Advance Access publication March 25, 2004

ABSTRACT

Summary: Two web-based applications to analyze amino acids three-dimensional (3D) local environment within protein structures—SCORPION and FORMIGA—are presented. SCORPION and FORMIGA produce a graphical presentation for simple statistical data showing the frequency of residue occurrence within a given sphere (defined here as the 3D contacts). The center of that sphere is placed at the Cα and at the last heavy atom in the side chain of the selected amino acid. Further depth of detail is given in terms of a secondary structure to which the profiled amino acid belongs. Results obtained with those two applications are relevant for estimating the importance of the amino acid 3D local environment for protein folding and stability. Effectively, SCORPION and FORMIGA construct knowledge-based force fields. The difference between SCORPION and FORMIGA is in that the latter operates on protein interfaces, while the former only functions for a single protein chain. Both applications are implemented as stand-alone components of STING Millennium Suite.

Availability: http://sms.cbi.cnptia.embrapa.br/SMS, http://trantor.bioc.columbia.edu/SMS, http://mirrors.rcsb.org/SMS, http://www.es.embnet.org/SMS and http://www.ar.embnet.org/SMS. {options: Scorpion, Formiga}

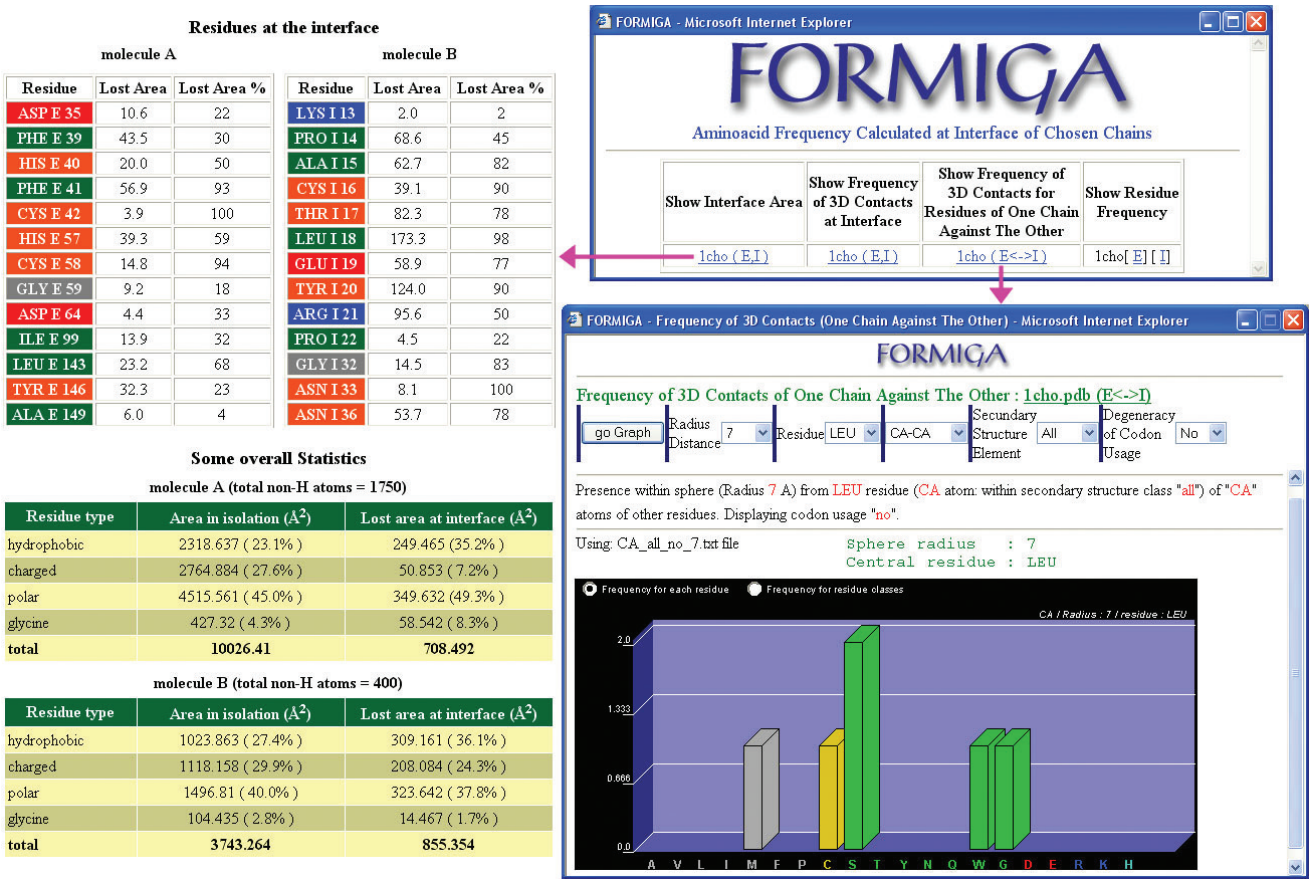
Contact: neshich@cnptia.embrapa.br

The search for coding properties hidden within the sequence of amino acids in terms of how that sequence determines its three-dimensional (3D) fold is yet without a final answer. However, some promising results have been reported (Cootes *et al*., 1998; Wilmanns and Eisenberg, 1993; Reddy *et al*., 1998; Zhang and Kim, 2000), resulting in fold prediction algorithms. Those algorithms rely mainly on knowledge-based force fields (Cootes *et al*., 1998). The purpose of the two applications described here is to help calculate the natural propensities of amino acids for a given structural environment. Special emphasis, in this respect, is given to the protein interfaces which have a fundamental importance for deciphering the mechanism responsible for the specificity in protein binding.

FORMIGA is the compilation of tools used to calculate the frequency of occurrence of amino acids at the interface formed by two or more facing chains in a protein structure described in the PDB (Berman *et al*., 2000) file(s). At the same time, those tools allow a user to visualize results in a graphically convenient and easy to interpret way. A user provides information requested on the entry web page of FORMIGA. In case of calculating the frequency of occurrence of the amino acids at the interface, it is required that the user provides: (a) the name(s) of protein(s) (PDB file name(s)) and (b) the chain names that will belong to the facing subunits. The FORMIGA algorithm operates with only two subunits that will form the interface. Those two subunits can contain one or more protein chains and those are grouped in a way that the user indicates. Once the interface and its closest vicinity are identified (by calculating the lost surface area between the two subunits upon complex formation among them), the program will count amino acids in the defined region and present them in a graphically convenient way. In the case of 3D contacts, the user may select the radius of the sphere within which the 3D contacts will be counted. In addition, the user should choose the central residue for which the contacts will be identified and counted, as well as the atom from which the distances will be measured [either alpha carbon or last heavy atom (LHA) in the amino acid side chain]. The user can also choose the secondary structure element to which the central residue belongs, and from which the 3D contacts will be counted (Fig. 1).

The user might want to calculate the 3D contacts and frequency of amino acid occurrence in an ensemble of similar proteins (say, serine proteases or alpha amylases or any other protein family of interest). Such data might help in

*To whom correspondence should be addressed.



Residues at the interface

molecule A

Residue	Lost Area	Lost Area %
ASP E 35	10.6	22
PHE E 39	43.5	30
HIS E 40	20.0	50
PHE E 41	56.9	93
CYS E 42	3.9	100
HIS E 57	39.3	59
CYS E 58	14.8	94
GLY E 59	9.2	18
ASP E 64	4.4	33
ILE E 99	13.9	32
LEU E 143	23.2	68
TYR E 146	32.3	23
ALA E 149	6.0	4

molecule B

Residue	Lost Area	Lost Area %
LYS I 13	2.0	2
PRO I 14	68.6	45
ALA I 15	62.7	82
CYS I 16	39.1	90
THR I 17	82.3	78
LEU I 18	173.3	98
GLU I 19	58.9	77
TYR I 20	124.0	90
ARG I 21	95.6	50
PRO I 22	4.5	22
GLY I 32	14.5	83
ASN I 33	8.1	100
ASN I 36	53.7	78

Some overall Statistics

molecule A (total non-H atoms = 1750)

Residue type	Area in isolation (A²)	Lost area at interface (A²)
hydrophobic	2318.637 (23.1%)	249.465 (35.2%)
charged	2764.884 (27.6%)	50.853 (7.2%)
polar	4515.561 (45.0%)	349.632 (49.3%)
glycine	427.32 (4.3%)	58.542 (8.3%)
total	**10026.41**	**708.492**

molecule B (total non-H atoms = 400)

Residue type	Area in isolation (A²)	Lost area at interface (A²)
hydrophobic	1023.863 (27.4%)	309.161 (36.1%)
charged	1118.158 (29.9%)	208.084 (24.3%)
polar	1496.81 (40.0%)	323.642 (37.8%)
glycine	104.435 (2.8%)	14.467 (1.7%)
total	**3743.264**	**855.354**

Fig. 1. Top right inset: the FORMIGA's output web page is shown for the PDB file 1cho. Partial listing of the interface forming residues (IFR) and their classification along with the report on surface accessible area is given on the left part of this figure. Bottom right inset: the user can choose the appropriate parameters for the calculation of the 3D contacts: radius distance, residue, the choice of the atom (CA-CA and LHA-LHA) from/to which the distances will be calculated. The option Secondary Structure Element will indicate to the program the restriction in terms of the conformation where the base residue should be located. The last option Degeneracy of Codon Usage is not yet available. The resulting graphical presentation indicates the type and number of residues belonging to the I chain and making 3D contact with the base residues (in this case with the LEUcines) in the E chain.

describing the observed protein family specificity for a given substrate/inhibitor, the key feature of interest in 'biological control', e.g. where the specificity of a certain enzyme for a given inhibitor has to be modified in order to achieve alternative and successful binding to a different inhibitor, resulting in a desired biological effect.

If several PDB formatted files are to be processed by either SCORPION or FORMIGA, the user will receive resulting data on each PDB file as well as on the TOTAL (sum of frequencies) over the ensemble of PDB files indicated at the input.

SCORPION AND FORMIGA DIFFERENCES

(A) SCORPION operates on *all* the residues available in each defined chain, while FORMIGA operates only on the residues localized at the interface of the two subunits; a subunit might be defined by the user as a single protein chain or the sum of several protein chains. In addition, the user might edit a PDB file and define chains according to the domain regions, effectively making it possible to study the interactions between domains.

(B) FORMIGA has an extra option 'show interface area', which SCORPION does not. This option shows tabulated data on lost surface area for all amino acids (grouped in two subunits).

SCORPION and FORMIGA are part of our Sting Millennium Suite (SMS) (Neshich *et al*., 2003). However, both programs are also available in stand-alone versions. SCORPION and FORMIGA are implemented in Perl, Fortran and Java.

The use of SCORPION and FORMIGA in conjunction with SMS Contacts and SMS IFR Contacts (Neshich *et al*., 2003) for didactic purposes has been proven to be well suited for the Bioinformatics classes, specifically to introduce various aspects of protein stability and protein-binding specificity.

ACKNOWLEDGEMENTS

This work was supported by FAPESP, Fundação de Amparo a Pesquisa do Estado de São Paulo—Project #1945/01, CNPq, Conselho Nacional de Desenvolvimento Científico e Tecnológico—Project #521093/2001-5 (NV) and FINEP, Financiadora de Estudos e Projetos—Project #01/08895-0.

REFERENCES

Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Cootes,A.P., Curmi,P.M.G., Cunningham,R., Donnelly,C. and Torda,A.E.(1998) The dependence of amino acid pair correlations on structural environment. *Proteins Struct. Func. Genet.*, **32**, 175–189.

Neshich,G., Togawa,R.C., Mancini,A.L., Kuser,P.R., Yamagishi,M.E.B., Pappas,G.,Jr., Torres,W.V., Campos,T.F., Ferreira,L.L., Luna,F.M., *et al*. (2003) STING Millennium: a Web based suite of programs for comprehensive and simultaneous analysis of protein structure and sequence. *Nucleic Acids Res.*, **31**, 3386–3392.

Reddy,B.V.B, Datta,S. and Tiwari,S. (1998) Use of propensities of amino acids to the local structural environments to understand effect of substitution mutations on protein stability. *Protein Eng.*, **11**, 1137–1145.

Wilmanns,M. and Eisenberg,D. (1993) Three-dimensional profiles from residue-pair preferences: Identification of sequences with α/β-barrel fold. *Proc. Natl Acad. Sci., USA*, **90**, 1379–1383.

Zhang,C. and Kim,S.-H. (2000) Environment-dependent residue contact energies for proteins. *Proc. Natl Acad. Sci., USA*, **97**, 2550–2555.