W508–W511 *Nucleic Acids Research, 2004, Vol. 32, Web Server issue*
DOI: 10.1093/nar/gkh388

SCit: web tools for protein side chain conformation analysis

R. Gautier, A.-C. Camproux and P. Tufféry*

Equipe de Bioinformatique Génomique et Moléculaire, INSERM E346, Université Paris 7, case 7113, 2, place Jussieu, 75251 Paris cedex 05, France

Received February 13, 2004; Revised and Accepted March 24, 2004

ABSTRACT

SCit is a web server providing services for protein side chain conformation analysis and side chain positioning. Specific services use the dependence of the side chain conformations on the local backbone conformation, which is described using a structural alphabet that describes the conformation of fragments of four-residue length in a limited library of structural prototypes. Based on this concept, SCit uses sets of rotameric conformations dependent on the local backbone conformation of each protein for side chain positioning and the identification of side chains with unlikely conformations. The SCit web server is accessible at http://bioserv.rpbs.jussieu.fr/SCit.

INTRODUCTION

The analysis and the prediction of protein side chain conformations are important features of protein modelling and have been the subject of many developments. In particular, the building of side chains on-to a protein backbone, an important step in protein structure prediction using homology modelling or *ab initio* methods, has given rise to several tens of studies [see (1) and references included]. Although important, such methods remain only a subset of the tools that are necessary to investigate side chain conformations. Basic tools for the editing and comparison of side chain conformations, or simple tools listing the neighbours of a side chain in a structure, remain of interest to biologists. Another point to consider is that side chain positioning methods generally result in proposing only one conformation for each side chain. Few tools allow the exploration of alternative conformations that one particular side chain of a particular structure is likely to adopt.

SCit is a web server accepting PDB-formatted coordinate files (2) that provides a wide variety of services related to protein side chain conformations. It can calculate, plot and edit the conformations of protein side chains, but also build side chains on-to a protein backbone and provide some evaluation of the quality of the side chain conformations.

DEPENDENCE BETWEEN SIDE CHAIN AND BACKBONE CONFORMATION

The dependence of the protein side chain conformation on that of the backbone has long been described, and its accurate analysis has brought significant progress in the field of side chain positioning. Usually, the dependence of the conformation of one side chain on that of the backbone is analysed as a function of the ϕ and ψ angle values of that amino acid (3–6). Due to the number of possible combinations of the ϕ and ψ values, the extension of such dependence to fragments of several residues in length is difficult (7). We have overcome this problem by describing the protein backbone conformation using a structural alphabet. The protein backbone is considered as a series of overlapping fragments of four-residue length. This series is converted into a series of letters of a structural alphabet that was learnt using Hidden Markov models (8). This approach learns simultaneously the geometry of the letters and the local rules that govern their assembly. The size of the alphabet is currently set to 27, a size chosen as providing the most accurate description of protein structure with no over-fitting of the model parameters (A. C. Camproux, R. Gautier and P. Tuffery, submitted for publication). Given a backbone geometry, it is possible to search for the optimal series of letters of the structural alphabet describing that particular geometry using the Viterbi algorithm. Having a one-dimensional representation of the backbone conformation in a limited number of letters, we can analyse the conformations of the side chains as a function of their associated letters of the alphabet. The minimal granularity of the analysis is one letter (which corresponds to studying the conformations of each amino acid type associated with fragments of four residues). This can be generalized to studying the conformations of each amino acid type associated with a series of k letters. Presently, we use a one- or two-letter dependence (R. Gautier, A. C. Camproux and P. Tuffery, submitted for publication), and we have built, using a method described by Lovell *et al.* (9), three rotamer libraries: one backbone-independent

*To whom correspondence should be addressed. Tel: +33 1 44 27 77 33; Fax: +33 1 43 26 38 30; Email: tuffery@ebgm.jussieu.fr

The online version of this article has been published under an open access model. Users are entitled to use, reproduce, disseminate, or display the open access version of this article provided that: the original authorship is properly and fully attributed; the Journal and Oxford University Press are attributed as the original place of publication with the correct citation details given; if an article is subsequently reproduced or disseminated not in its entirety but only in part or as a derivative work this must be clearly indicated.

© 2004, the authors

rotamer library (lib0L) and two libraries obtained for dependence of one (lib1L) and two (lib2L) letters. These libraries were built from a collection of non-redundant proteins obtained using an algorithm similar to that of the culled PDB (10). Only proteins of at least 30 amino acids length, having no chain breaks, obtained by X-ray diffraction with a resolution better than 2.5 Å, were retained. In November 2003, this resulted in a collection of 2926 protein chains presenting less than 30% sequence identity. The libraries lib0L and lib1L are listed at http://bioserv.rpbs.jussieu.fr/Doc/Rotamers.html.

DETECTION OF INCORRECT SIDE CHAIN CONFORMATIONS

Besides the libraries, we have also computed the densities of the probabilities of side chain conformations on grids in steps of 10°. Analysing the variations in the density observed for the side chain conformations in the different libraries, we have set up a procedure to assess if a given side chain conformation is likely to occur for a given letter of the structural alphabet. We classify as risk level 1 side chains for which the conformation corresponds to a density greater than a threshold in lib1L, and a density less than this threshold for lib2L (typical threshold values are of the order of 0.001). We classify as risk level 2 conformations having densities less than the threshold for both lib1L and lib2L. For each type of side chain, given a set of predicted conformations, thresholds have been optimized on a collection of proteins to detect erroneous conformations. Note that this does not correspond to the detection of the off-rotameric conformations that have been described in the literature (11). We find that is possible to detect 20% of the side chains incorrectly positioned with a specificity >90%.

SIDE CHAIN POSITIONING

To position the side chains on-to a protein backbone, we presently use a procedure similar to that described by Bower *et al*. (12), but using our backbone-dependent rotamer libraries. Compared to the original method, we use a simplified approach. First, we position each side chain in its most frequent conformation with no clash with the backbone. Then we detect the side chains with clashes with other side chains and, having sorted the side chains by decreasing clash order, we attempt to solve each problem in turn, continuing to scroll through the list of conformations by decreasing occurrence frequencies. Such a procedure is sub-optimal only, since we do not attempt to optimize the energy in considering the combinations of the conformations.

IMPLEMENTATION OF THE SERVICES

The SCit tools are organized as a series of cascading common gateway interface programs (CGIs). Most of them consist of Python scripts wrapping calls to C programs to perform the calculations. The flowchart of the tools is reported in Figure 1. Briefly, the CGIs are organized around a main form that allows a PDB file to be loaded or uploaded. Once activated, SCit goes through a series of preliminary steps. (i) It allocates an identifier to the request, which is used to manage the transmission of the data through cycles of analysis and editing of the file. (ii) It performs a series of analyses to check the quality of the

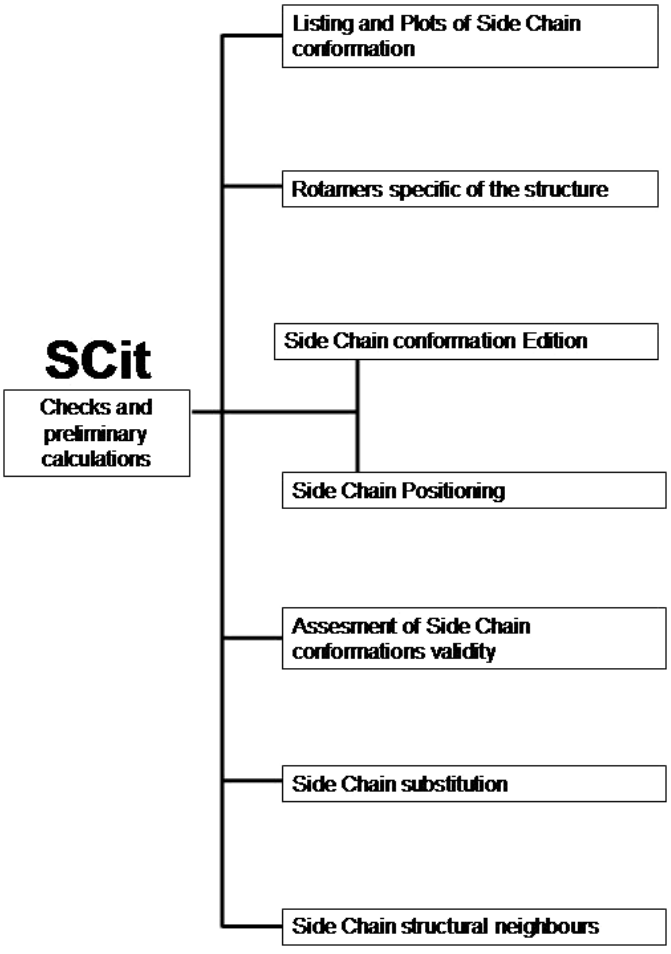

Figure 1. Flowchart of the services.

coordinates of the backbone. Such a check is necessary since the encoding of the structure using the structural alphabet is sensitive to errors in the geometry of the alpha-carbon trace of the proteins, which can result in incorrect encoding. (iii) It detects if there are missing atoms in the side chains. In such cases, the side chains are automatically rebuilt in a standard conformation corresponding to the most probable conformation of the side chain in a backbone-independent rotamer library. (iv) It performs a series of preliminary calculations, such as the compilation of the rotamer library compatible with the structure, the positioning of hydrogens and the calculation of the accessible surface area (ASA). The positioning of the hydrogens is based on the standard stereo-chemical values and is achieved using an adaptation of the procedure used by the SMD package (13). The accessible surface area is calculated using the method of T. J. Richmond (14). Filling the main form, it is also possible to specify a sequence. In this case, the file available for further work will be the result of the substitutions and side chain positioning deduced from the sequence specified.

The second-stage services provide (i) a side chain conformation listing and plot, (ii) a structure dependent rotamer listing, (iii) side chain conformation editing either at the level of chi angles or at the level of rotamers, (iv) side

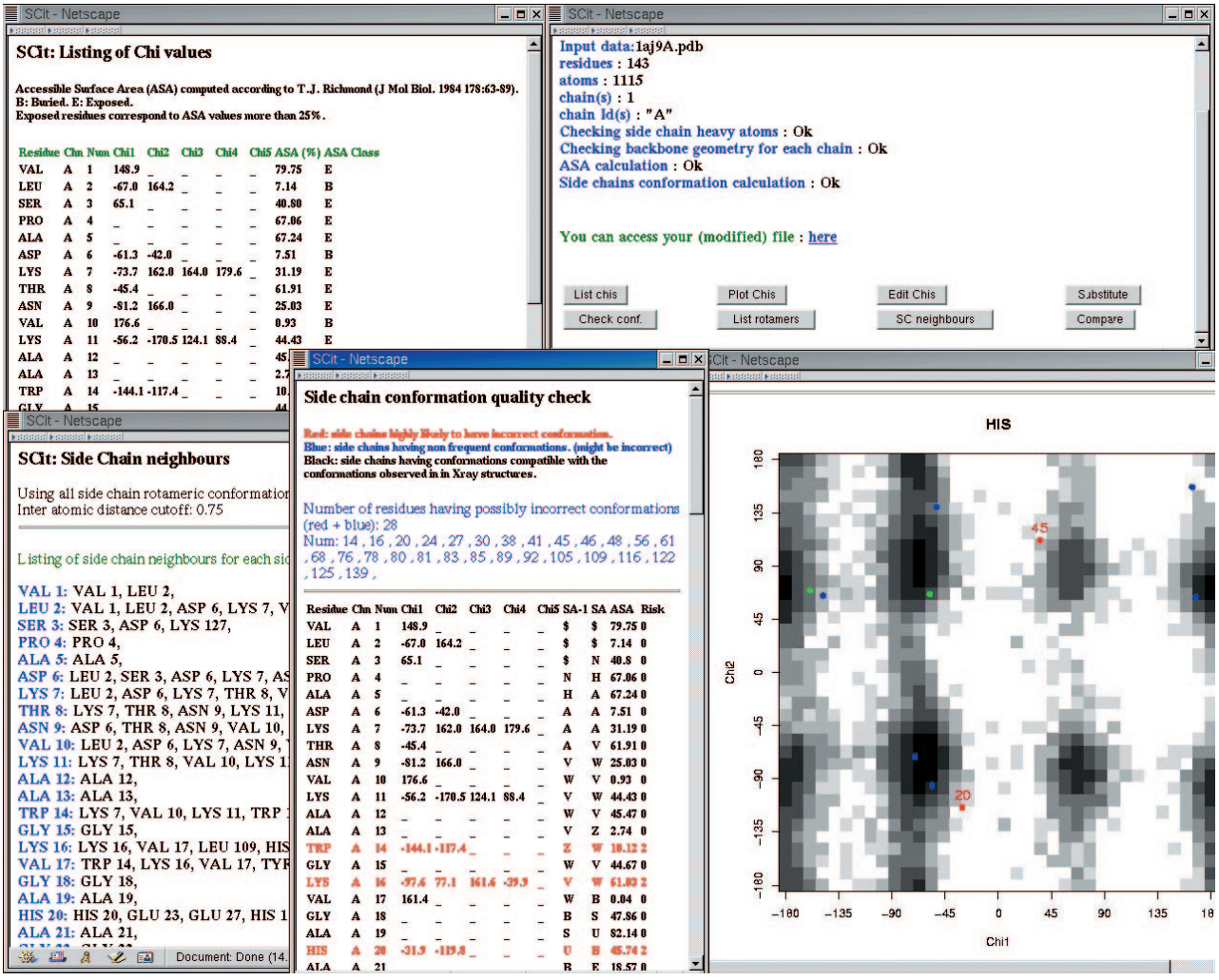

Figure 2. A screenshot montage of SCit output.

chain substitution and positioning, (v) an assesment of the quality of side chain conformations and (vi) the identification of the structural neighbours of the side chains. Neighbour identification is based on the distance between the side chain atoms (hydrogens included). Two side chains are neighbours if at least two of their atoms are in contact (i.e. their interatomic distance minus the sum of their van der Waals radii is less than a user-specified threshold—0.75 Å by default). Neighbours can be defined on the basis of the side chain current atomic coordinates or by considering all their possible rotameric conformations. In the latter case, two side chains are neighbours if at least two of their rotameric conformations are in contact. At each stage, a link is set on the current structure, so as to provide the possibility of viewing it using a visualization package using the mime facility of a web browser. Note that side chain substitutions are imposed by the user; SCit is not currently connected to any tool to assess their impact on the stability of the protein. Finally, an external service can perform the comparison of the conformations of the side chains of two structures. It can compare the conformation of equivalent side chains using for each side chain standard criteria such as the RMSd of the side chain heavy atoms coordinates or the deviation of the chi angles.

OUTPUT

The different services provide various outputs that range from the listing of chi values to plots of the dihedral angle values superimposed, for each amino acid type, to the conformations observed in a large collection of structures. For the plots, different colours represent the accessible surface area of the residues and the deviation of the conformation to the standard rotameric conformations. Finally, the results of the editing of the file are returned as a PDB-formatted file. Figure 2 provides a sample of the rich graphical and textual output from a standard SCit run.

CONCLUSIONS AND FUTURE DIRECTIONS

Work is presently in progress to make possible the analysis of the conformations given some experimental constraints such as distances resulting from NMR experiments, and to improve the algorithm for side chain positioning.

ACKNOWLEDGEMENTS

We thank J. Chomilier for suggesting the integration of our collection of tools related to side chains. This work

was supported by 'Action Conjointe CNRS-INSERM, Bioinformatique 2002'.

REFERENCES

1. Canutescu,A.A., Shelenkov,A.A. and Dunbrack,R.L.,Jr (2003) A graph-theory algorithm for rapid protein side chain prediction. *Protein Sci.*, **9**, 2001–2014.
2. 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.
3. Chakrabarti,P. and Pal,D. (1998) Main-chain conformational features at different conformations of the side-chains in proteins. *Protein Eng.*, **11**, 631–647.
4. Chakrabarti,P. and Pal,D. (2001) The interrelationships of side-chain and main-chain conformations in proteins. *Prog. Biophys. Mol. Biol.*, **76**, 1–102.
5. Dunbrack,R.L.,Jr and Karplus M. (1993) Backbone-dependent rotamer library for proteins. Application to side-chain prediction. *J. Mol. Biol.*, **230**, 543–571.
6. Dunbrack,R.L.,Jr and Cohen,F.E. (1997) Bayesian statistical analysis of protein sidechain rotamer preferences. *Protein Sci.*, **6**, 1661–1681.
7. Fetrow,J. and Berg,G. (1999) Using information theory to discover side chain rotamer classes: analysis of the effects of local backbone structure. *Pac. Symp. Biocomput.*, 278–289.
8. Camproux,A.C., Tuffery,P., Chevrolat,J.P., Boisvieux,J.F. and Hazout,S. (1999) Hidden Markov model approach for identifying the modular framework of the protein backbone. *Protein Eng.*, **12**, 1063.
9. Lovell,S.C., Word,J.M., Richardson,J.S. and Richardson,D.C. (2000) The Penultimate rotamer library. *Proteins*, **40**, 389–408.
10. Wang,G. and Dunbrack,R.L.,Jr (2003) PISCES: a protein sequence culling server. *Bioinformatics*, **19**, 1589–1591
11. Petrella,R.J. and Karplus,M. (2001) The energetics of off-rotamer protein side-chain conformations. *J. Mol. Biol.*, **312**, 1161–1175.
12. Bower,M.J., Cohen,F.E. and Dunbrack,R.L.,Jr (1997) Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool. *J. Mol. Biol.*, **267**, 1268–1282.
13. Tufféry,P., Etchebest,C., Hazout,S. and Lavery,R. (1991) A new approach to the rapid determination of protein side chain conformations. *J. Biomol. Struct. Dyn.*, **8**, 1267–1289.
14. Richmond,T.J. and Richards,F.M. (1978) Packing of alpha-helices: geometrical constraints and contact areas. *J. Mol. Biol.*, **119**, 537–553.