Nucleic Acids Research, 2007, Vol. 35, Web Server issue ***W489–W494***
doi:10.1093/nar/gkm422

siteFiNDER|3D: a web-based tool for predicting the location of functional sites in proteins

C. Axel Innis*

Howard Hughes Medical Institute/Yale University, Department of Molecular Biophysics and Biochemistry, New Haven, CT 06520-8114, USA

Received January 30, 2007; Revised May 6, 2007; Accepted May 9, 2007

ABSTRACT

Although knowledge of a protein's functional site is a key requirement for understanding its mode of action at the molecular level, our ability to locate such sites experimentally is far exceeded by the rate at which sequence and structural information is being accumulated. siteFiNDER|3D is an online tool for the prediction of functionally important regions in proteins of known structure. At the core of the server lies the CFG analysis algorithm, which uses a moving 3D window to correlate patterns of functional/chemical group conservation in the query protein with the location of functional sites. Here, we give a general overview of the functionality offered by the siteFiNDER|3D server, along with general recommendations aimed at maximizing the accuracy and predictive value of this tool in a variety of contexts. siteFiNDER|3D can be accessed at: 'http://sage.csb.yale.edu/sitefinder3d' and requires, at a minimum, the atomic coordinates of a query protein in PDB format.

INTRODUCTION

Conserved functional group (CFG) analysis is a general method for predicting the location of functionally important regions within a protein of known structure (1). Like several other structure/sequence analysis techniques—such as evolutionary trace (ET) analysis (2,3), 3D cluster analysis (4) or ConSurf (5,6)—CFG analysis exploits the evolutionary relationships present within groups of homologous proteins to identify sites that are likely to be of functional significance. However, by using a 3D smoothing window to analyse the spatial distribution of functional group conservation, CFG analysis has been shown to succeed where low sequence diversity causes at least one other method to fail (1), making it the method of choice for the preliminary identification of protein functional sites in a structural genomics context.

In this article, we present siteFiNDER|3D, a fully integrated, web-based implementation of the CFG analysis method for functional site prediction. What follows is a brief description of the server's processing method and run-time parameters, along with a discussion of the input data required, the output generated, a comparison with other servers offering similar functionality and a set of general guidelines for effective use.

MATERIALS AND METHODS

Processing method and run parameters

The CFG analysis algorithm at the core of the siteFiNDER|3D server has been described elsewhere (1) and will not be covered in detail here. In short, CFG analysis correlates the extent and spatial distribution of functional group conservation in a query protein of known structure with the location of functionally important sites. In order to do so, it must first extract CFG clusters from a multiple sequence alignment containing the query and a number of its homologues. These clusters are defined as sets of one or more functional groups of the same type occupying equivalent positions in the alignment, with spatial coordinates assigned from the C^{β} atom of the corresponding residue in the query structure. For the purposes of this method, functional groups include chemical groups from amino acid side chains with a potential for taking part in hydrogen bonding, electrostatic or aromatic stacking interactions. Once CFG clusters have been identified and overlaid onto the query structure, a moving 3D window is used to calculate normalized functional group conservation (C_{atm}) scores for every atom in the molecule. These scores are a measure of CFG density—the local extent of functional group conservation in the structure—and regions displaying the highest C_{atm} values generally correspond to functional sites.

The CFG analysis algorithm itself is implemented in C++ (7) and features a Binary Spatial Division (BSD) tree data structure (8) for evaluating spatial relationships between atoms, residues and CFG clusters, thereby reducing significantly the complexity of such operations,

*To whom correspondence should be addressed. Tel: 203 432 5627; Email: axel.innis@yale.edu

together with the overall running time of the program. In addition, the siteFiNDER|3D server relies on third-party software to prepare the data used by the CFG analysis algorithm. When no multiple sequence alignment is provided by the user, the server accumulates homologues by performing a single BLAST (9,10) search on the non-redundant sequence database, with the E-value cut-off set to 0.001. Sequences covering <70% of the length of the query protein are discarded and redundancy is minimized using CD-HIT (11–13), thereby ensuring that the majority of sequences retained share no more than 90% sequence identity with one another. Sequences remaining after this filtering step are aligned using the ClustalW (14) program with a Blosum62 substitution matrix (15). Following the CFG analysis step, prediction results are formatted into a report that is returned to the user. This processing step makes use of the program Voidoo (16) to calculate protein and site volumes, as well as MSMS (8) and POV-Ray ('http://www.povray.org') to generate and render surface representations of the predicted sites. The various server-side scripts necessary to integrate these different tasks are written in Python ('http://www.python.org') and PHP ('http://www.php.net').

Although the siteFiNDER|3D server may be run with minimal user intervention, several parameters can be modified that affect the way in which sequence homologues are accumulated or the CFG analysis itself is performed. This includes parameters such as the BLAST E-value cut-off, the minimum percent length of the query that must be accounted for in sequences retained for the alignment or the level of sequence redundancy tolerated by CD-HIT. As far as the CFG analysis algorithm is concerned, the user can modify most of the parameters described in the original method, though doing so may lead to unpredictable results and to lower accuracy compared to the published benchmarking data (1).

Input and output data

Input data for the siteFiNDER|3D server consists, at a minimum, of a query protein with structural coordinates provided in standard PDB (17) format. In addition, the user may choose to upload a multiple sequence alignment featuring homologues of the query protein or, as mentioned previously, to allow the server to generate an alignment using sequences derived from a BLAST search. While the latter option presents the user with a quicker, more convenient alternative, it is most likely to result in a successful prediction in cases where the query protein corresponds to a well-defined evolutionary unit with a set of sequence homologues covering most of its length—as is the case with many single domain proteins, but also with multi-domain proteins that have evolved as a single unit. For more complex cases, such as multi-subunit proteins or large modular proteins with unique domain combinations, it may be necessary to perform CFG analysis on each of the isolated domains or to supply the server with a single, composite sequence alignment assembled from sets of homologues accumulated individually for each domain.

After CFG analysis has been carried out, the server generates a report detailing the results of the prediction (Figure 1). This includes a list of predicted functional sites, each consisting of one or more overlapping functional patches, delimited in space by spheres of different radii. For each predicted site, a list of all the residues whose C^{β} atom falls within the site is returned, along with the absolute and fractional volumes calculated from the set of atoms present inside that site. The latter may be used as an indicator of the usefulness of the prediction, since the majority of functional sites in proteins does not exceed 30% of the total protein volume (1). Finally, a PDB file containing all the atoms within the predicted site is available for download, together with a view of the molecule showing mapped C_{atm} scores in the region of the predicted site and the script used to generate the image with the ray-tracing program POV-Ray.

In addition to the individual descriptions of the predicted sites, the report also includes an image of the query protein sequence, with each residue coloured according to its average C_{atm} value, and a coordinate file of the query protein in PDB format, with individual C_{atm} scores mapped to the temperature factor column of the file. This gives the user the opportunity to inspect the distribution of CFG density more closely, in order to detect noisy or artefactual data arising from a sequence alignment of highly similar proteins.

Comparison with other servers

Earlier assessments of the performance of CFG analysis showed that this method is able to make reliable predictions over a wide range of sequence identities, whereas at least one other method was unable to produce meaningful output for alignments displaying >10% identity (1). In this report, we compare the performance of siteFiNDER|3D on MukB—a multi-domain protein involved in the ATP-dependent partitioning of the *Escherichia coli* chromosome during cell division (19)—with that of two other web-based services providing a similar facility: the ConSurf server (5,6) and the ET Viewer 2.0 (18) (Figure 2). In doing so, we do not wish to suggest that siteFiNDER|3D provides a better alternative overall to the use of these other servers; drawing such a conclusion would indeed require extensive benchmarking and is therefore well beyond the scope of this work. Rather, we hope that the qualitative analysis presented here will serve to highlight one of the previously demonstrated strengths of the CFG analysis method: its ability to make useful predictions with data exhibiting poor coverage of sequence space.

Our case study focuses on the 26-kDa N-terminal domain of MukB, which features a mixed α/β-fold with a central six-stranded anti-parallel β-sheet and a putative Walker A motif. The only available high-resolution structure of this domain reveals no clear structural similarity to any other known nucleotide-binding protein and suggests that the potential nucleotide-binding loop is too exposed to form a functional binding pocket. Together with additional biochemical evidence, this was used to propose a model in which the N- and C-terminal domains of MukB assemble to form an anti-parallel

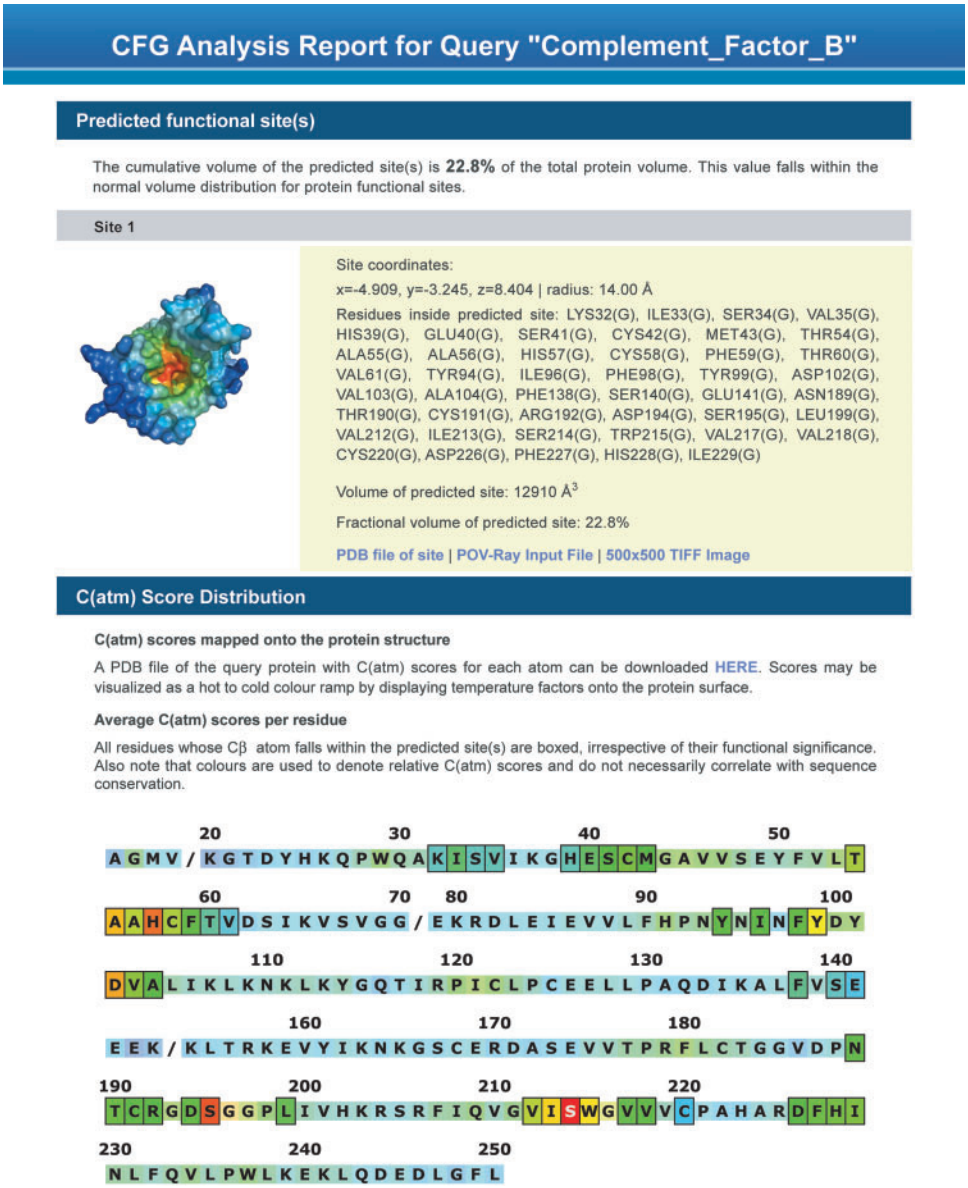

Figure 1. Typical output from the siteFiNDER|3D server, showing a successful prediction for the serine proteinase domain of Complement Factor B (PDB code: 1dle, chain G) (24). The predicted site consists of a single spherical patch that encompasses the enzyme's active site, including the catalytic triad residues His57, Asp102 and Ser195. The site accounts for 22.8% of the total protein volume and, as such, falls within the normal volume distribution for protein functional sites.

dimer, thereby leading to the formation of a complete active site (19,20).

To investigate this hypothesis further, the N-terminal domain of MukB was used as a query for siteFiNDER|3D, ConSurf and ET Viewer 2.0 and sets of sequence homologues were accumulated according to each server's particular methodology. Dataset A, derived by siteFiNDER|3D and consisting of 11 sequences with 48.5% identity, was obtained by performing a single BLAST search on the non-redundant sequence database with an E-value cut-off of 0.001, discarding all sequences <70% of the query's length, filtering for redundancy and aligning all of the remaining sequences with ClustalW (14). Dataset B, generated by ConSurf and featuring 36 sequences with 8.8% identity, was obtained by running a single BLAST search against the UniProt database (21) with an E-value cut-off of 0.001 and by aligning the resulting sequences with Muscle (22). Finally, dataset C was obtained from the ET Viewer 2.0 server and consisted of 42 sequences sharing 45.6% identity. Datasets A, B and C were each subsequently used as input to the three servers, resulting in a total of nine separate functional site predictions (Figure 2).

RESULTS AND CONCLUSIONS

The siteFiNDER|3D server was able to consistently predict a similar functional site using all three datasets and default run parameters. Indeed, the root mean square

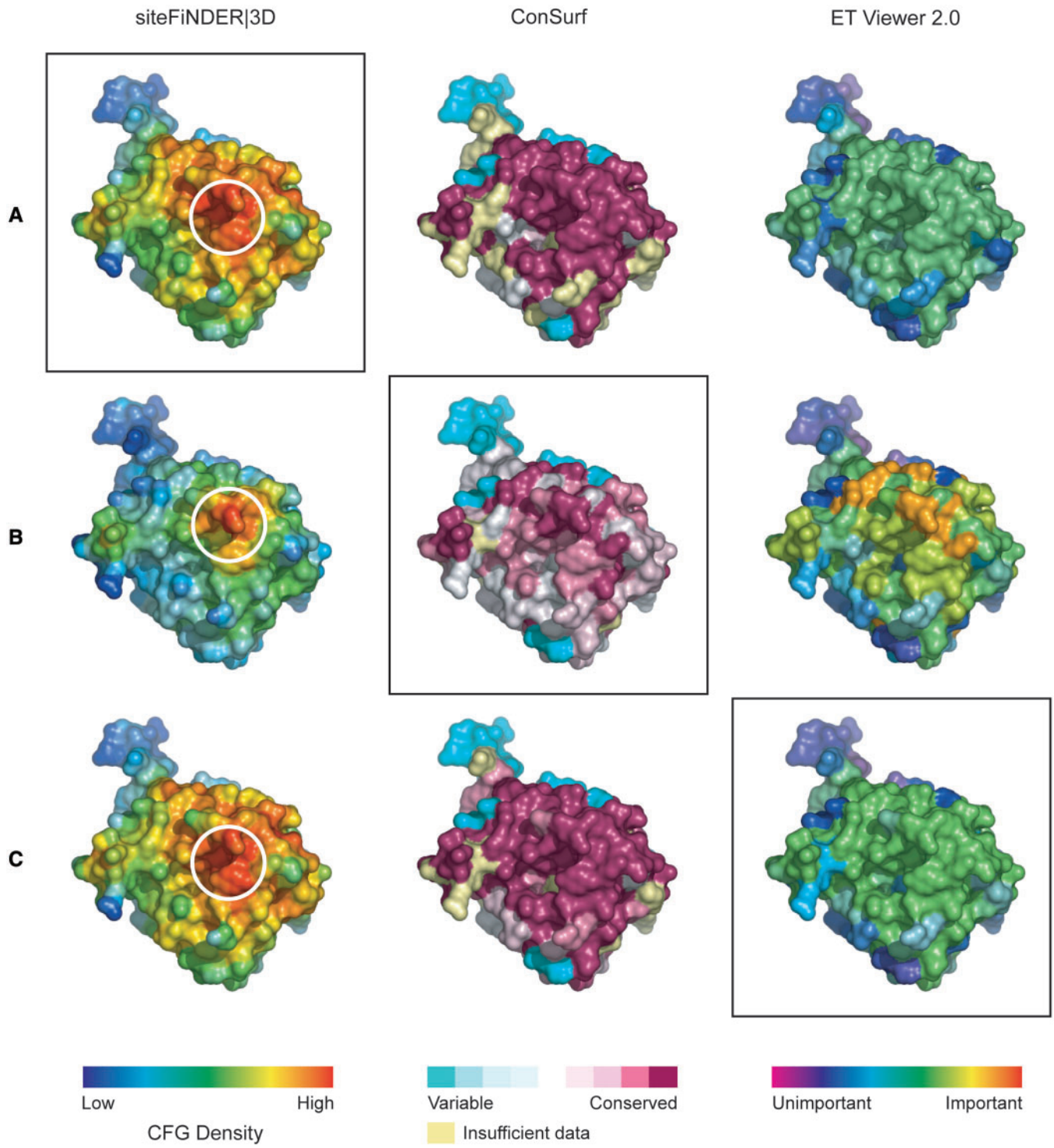

Figure 2. Comparison of the performance of siteFiNDER|3D, ConSurf and ET Viewer 2.0 on the N-terminal domain of MukB (PDB code: 1qhl, chain A), a protein involved in the partitioning of the *E.coli* chromosome (19). Scores from each method for sequence alignments obtained from the siteFiNDER|3D (**A**), ConSurf (**B**) and ET Viewer 2.0 (**C**) servers are mapped onto the surface of MukB. Results for each server and their corresponding sequence alignment are boxed. White circles are used to indicate the location and extent of the sites predicted by siteFiNDER|3D. Molecular surfaces were generated and rendered using PyMOL (25).

deviation of the centroids for these sites was 3.25 Å and their radius was 8.0 Å in all cases, with fractional volumes of 6.3%, 4.7% and 7.7% for datasets A, B and C, respectively. CFG analysis carried out for all datasets identified a region containing three of the residues belonging to the Walker-A motif of the putative G-loop ([AG](X)$_4$GK[ST])—Asn36, Lys40 and Ser41—as well as a varying number of surrounding amino acids. No

additional regions of the molecule were identified as functionally significant by this method.

To calculate conservation scores with the ConSurf server, a Bayesian method was used in conjunction with the JTT matrix for all three datasets. Dataset B gave rise to the prediction with highest specificity, with just 37 residues out of 227 (16.3%) classified as highly conserved (score of 9) and 21 residues (9.3%) as having insufficient data to calculate a meaningful score. Some of the residues predicted to be functionally important clustered around the putative G-loop and included Gly34, Asn36, Lys40 and Ser41. A few additional residues with a high degree of conservation, such as Arg 112, Glu202 or Tyr206, were also found in surrounding areas on the same face of the molecule, suggesting a possible role in the dimerization of MukB. In contrast, conservation scores calculated from datasets A and C consisted of 98 (43.2%) and 92 (40.5%) residues with a score of 9, and 54 (23.8%) and 30 (13.2%) residues considered as having insufficient data, respectively. In these cases, the ConSurf methodology offered no distinct advantage over the mapping of identical residues onto the structure and, as expected from the poor sequence diversity of the input alignment, gave rise to a prediction with very low specificity.

Results obtained from the ET Viewer 2.0 server were similar to those produced by ConSurf, with a clear, specific prediction available only for dataset B and featuring residues Gly34, Asn36, Gly37, Gly39, Lys40 and Ser41 from the Walker-A motif. Unlike the ConSurf server, however, ET Viewer 2.0 failed to make a useful prediction for its own multiple sequence alignment (dataset C), which was characterized by poor sequence diversity. An interesting feature of ET Viewer 2.0 is the ranking of predicted residues according to importance, which allows for a convenient and immediate distinction to be made between the accurate prediction for dataset B, where some of the residues were classified as relatively important, and the low-specificity predictions for datasets A and C, where residues ranged between average and unimportant.

To summarize, both ConSurf and ET Viewer 2.0 were able to predict the location of the MukB functional site accurately when the input sequence alignment provided good coverage of sequence space (dataset B), but failed to make a useful prediction when the fraction of identical residues in the input alignment was high (datasets A and C). In addition, default parameters had to be modified in both cases to obtain useful output. siteFiNDER|3D on the other hand, was capable of successfully identifying the putative nucleotide binding loop for all three datasets, thereby re-emphasising the method's ability to extract meaningful information from sub-optimal sequence data. By focusing on individual residues, however, ConSurf and ET Viewer 2.0 may be able to discern finer details than siteFiNDER|3D, such as amino acids important for the dimerization of MukB.

General considerations

Benchmarking carried out on 470 single-domain proteins belonging to 68 SCOP (23) families previously showed that CFG analysis is capable of predicting the location of functional sites correctly in ~60% of cases and partially in ~36% of cases, where a correct prediction is such that at least one of the predicted sites displays a 50% or greater volume overlap with the known functional site and a partial prediction consists of one or more sites overlapping with the known site by no more than 50% (1). For this level of reliability to be attained, however, the following guidelines should be taken into consideration:

(i) All structural domains present in the query must be accounted for in the sequence alignment. For multi-subunit proteins or proteins with unique domain combinations, sequences may need to be accumulated independently for each structural unit, aligned and reassembled into a single, composite alignment. It is crucial that each domain be equally represented in the alignment, since portions of the query with a larger number of sequence homologues are likely to introduce bias into the CFG analysis calculation.

(ii) When opting to use the BLAST feature provided by the server, different E-value (10^{-3}–10^{-5}) and length (70–90%) cut-off combinations may be used to accumulate sets of homologues of different sizes. By carrying out CFG analysis on 5–10 such sets and plotting the number of times a particular residue is found within the predicted sites, it should be possible to distinguish true hits from erroneous predictions. Indeed, correct sites should encompass clusters of residues that are predicted for the majority of the input alignments. Alternatively, building a phylogenetic tree from the initial sequence alignment and performing CFG analysis on different sequence subgroups within the tree may allow a similar cross-validation of the results to be carried out.

(iii) C_{atm} scores mapped onto the surface of the query structure should be inspected for discrepancies. Large, low scoring regions may be indicative of poor conservation, but may also be caused by incomplete coverage of the query by its homologues. Better results may therefore be obtained if the fragment for which no homologues can be identified is removed from the original query. Conversely, high C_{atm} scores found over the entire molecule typically reflect a low level of diversity in the sequence alignment, ultimately leading to lowered prediction accuracy.

(iv) If too many sites are predicted and the percentage of identical residues in the alignment is low, it is likely that the inclusion cut-off—the parameter used to determine whether an nth site is considered for inclusion into the prediction—was not assigned a sufficiently stringent value. Gradually increasing the value for this parameter should lead to fewer sites being predicted.

(v) While CFG analysis tends to be relatively resilient to errors in the sequence alignment, a manually curated alignment may enhance the accuracy of the final prediction. Any knowledge that could lead to an improved alignment, such as secondary structure

or other structural information, should therefore be taken into consideration when preparing the input data.

To conclude, it is worth pointing out that, as is often the case with sequence/structure-based functional site prediction techniques, exerting good judgment during the preparation of the input data and the analysis of the results will enhance the likelihood of success.

ACKNOWLEDGEMENTS

The author would like to thank Prof. Thomas A. Steitz for support and for providing the necessary infrastructure to host the siteFiNDER|3D Server, together with Scott Bailey, Miljan Simonović and Satwik Kamtekar for useful discussions. C.A. Innis is supported by the Howard Hughes Medical Institute. Funding to pay the Open Access publication charges for this article was provided by the Howard Hughes Medical Institute.

Conflict of interest statement. None declared.

REFERENCES

1. Innis,C.A., Anand,A.P. and Sowdhamini,R. (2004) Prediction of functional sites in proteins using conserved functional group analysis. *J. Mol. Biol.*, **337**, 1053–68.
2. Lichtarge,O., Bourne,H.R. and Cohen,F.E. (1996) An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.*, **257**, 342–358.
3. Yao,H., Kristensen,D.M., Mihalek,I., Sowa,M.E., Shaw,C., Kimmel,M., Kavraki,L. and Lichtarge,O. (2003) An accurate, sensitive, and scalable method to identify functional sites in protein structures. *J. Mol. Biol.*, **326**, 255–261.
4. Landgraf,R., Xenarios,I. and Eisenberg,D. (2001) Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. *J. Mol. Biol.*, **307**, 1487–1502.
5. Armon,A., Graur,D. and Ben-Tal,N. (2001) Consurf: an algorithmic tool for the identification of functional regions by surface mapping of phylogenetic information. *J. Mol. Biol.*, **307**, 447–463.
6. Landau,M., Mayrose,I., Rosenberg,Y., Glaser,F., Martz,E., Pupko,T. and Ben-Tal,N. (2005) ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Res.*, **33**, W299–W302.
7. Stroustrup,B. (1997) *The C++ Programming Language*, *3rd edn*. Addison-Wesley Professional.
8. Sanner,M.F., Olson,A.J. and Spehner,J.C. (1996) Reduced surface: an efficient way to compute molecular surfaces. *Biopolymers*, **38**, 305–320.
9. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
10. Ye,J., McGinnis,S. and Madden,T.L. (2006) BLAST: improvements for better sequence analysis. *Nucleic Acids Res.*, **34**, W6–W9.
11. Li,W., Jaroszewski,L. and Godzik,A. (2001) Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics*, **17**, 282–283.
12. Li,W., Jaroszewski,L. and Godzik,A. (2002) Tolerating some redundancy significantly speeds up clustering of large protein databases. *Bioinformatics*, **18**, 77–82.
13. Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.
14. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) ClustalW: improving the sensitivity of progressive multiple sequence alignment through sequence clustering, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
15. Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.
16. Kleywegt,G.J. and Jones,T.A. (1994) Detection, delineation, measurement and display of cavities in macromolecular structures. *Acta Crystallogr. D Biol. Crystallogr.*, **50**, 178–185.
17. 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.
18. Morgan,D.H., Kristensen,D.M., Mittleman,D. and Lichtarge,O. (2006) ET Viewer: An application for predicting and visualizing functional sites in protein structures. *Bioinformatics*, **22**, 2049–2050.
19. van den Ent,F., Lockhart,A., Kendrick-Jones,J. and Lowe,J. (1999) Crystal structure of the N-terminal domain of MukB: a protein involved in chromosome partitioning. *Struct. Fold. Des.*, **7**, 1181–1187.
20. Melby,T.E., Ciampaglio,C.N., Briscoe,G. and Erickson,H.P. (1998) The symmetrical structure of structural maintenance of chromosomes (SMC) and MukB proteins: long, antiparallel coiled coils, folded at a flexible hinge. *J. Cell Biol.*, **142**, 1595–1604.
21. The UniProt Consortium (2007) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **35**, D193–D197.
22. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
23. Murzin,A.G, Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
24. Jing,H., Xu,Y., Carson,M., Moore,D., Macon,K.J., Volanakis,J.E. and Narayana,S.V. (2000) New structural motifs on the chymotrypsin fold and their potential roles in complement factor B. *EMBO J.*, **19**, 164–73.
25. DeLano,W.L. (2002) *The PyMOL User's Manual* DeLano Scientific, Palo Alto, CA, USA.