W480–W484 *Nucleic Acids Research, 2009, Vol. 37, Web Server issue* *Published online 22 May 2009*
doi:10.1093/nar/gkp431

The SALAMI protein structure search server

Thomas Margraf*, Gundolf Schenk and Andrew E. Torda

Centre for Bioinformatics, University of Hamburg, Bundesstr. 43, 20146 Hamburg, Germany

Received January 31, 2009; Revised April 30, 2009; Accepted May 11, 2009

ABSTRACT

Protein structures often show similarities to another which would not be seen at the sequence level. Given the coordinates of a protein chain, the SALAMI server at www.zbh.uni-hamburg.de/salami will search the protein data bank and return a set of similar structures without using sequence information. The results page lists the related proteins, details of the sequence and structure similarity and implied sequence alignments. Via a simple structure viewer, one can view superpositions of query and library structures and finally download superimposed coordinates. The alignment method is very tolerant of large gaps and insertions, and tends to produce slightly longer alignments than other similar programs.

INTRODUCTION

Purpose of SALAMI

Sequence similarity is the classic measure for finding related proteins and the starting point for assigning function, building phylogenies and protein modelling. Sequence similarity will not, however, be enough to detect remote relationships. For this, one needs methods that detect pure structural similarity. Given the coordinates of a protein chain, the SALAMI server will search the protein data bank (1), for similar chains, calculate structural alignments and generate a list of structurally related proteins. In some sense, structure is preserved more than sequence during evolution (2) so even within a family of related proteins, there may be members with no significant sequence similarity to another (3–8). This means that questions of function or phylogenetic relations will often only be answerable given structural relationships (9). Furthermore, there is the question of alignment quality. In the case of weak sequence similarity, the alignment implied by a structural superposition should be more reliable and more useful for problems such as predicting functional sites.

Structure comparison

Aligning protein structures is a fundamentally NP-complete problem when one allows for arbitrary gaps and insertions (10). This means that all methods rely on some approximations and there will always be trade-offs between quality and speed. Furthermore, the problem is not perfectly defined since there may be no unique ideal alignment (11,12) and there is not even a single definition of alignment quality. One could argue that a good alignment minimizes differences in Cartesian space, but one could also say that a good method will find the corresponding residues despite large coordinate shifts due to hinge-bending or domain motions. For someone working on structure determination, it may be very useful if a method can recognize structural similarities when faced with the irregularities of an initial NMR-derived structure or unrefined crystallographic coordinates. Finally, programs will differ because they have been tuned to different goals. Some authors prefer shorter alignments of very similar regions, whereas some prefer longer alignments including regions of greater variation.

Because the alignment problem is difficult and not even well defined, there is a large variety of approaches and using *n* different programs may give *n* different structural alignments (13–43). There are, however, some common ideas. Some methods try to build a crude seed alignment which can be extended or iteratively improved (17,30). Some methods assign descriptors to sites which can be aligned using methods similar to those in sequence alignment. These descriptors, of course, come in many forms ranging from distance matrices to textbook secondary structure or fragment-based alphabets (18,33,44).

SALAMI also attaches descriptors to sites, but they are fuzzy or probabilistic. This means that there are no predefined thresholds and no requirement that a fragment be seen as helix, sheet or coil. Instead, fragments are compared with each other using a continuous estimate of similarity.

Although there is a large number of methods for structural alignment, relatively few are fast enough to search a large library of structures (21,22,24,25,33). The SALAMI server is fast enough to search the protein data bank for medium-sized proteins in 10–20 min using a single CPU.

*To whom correspondence should be addressed. Tel: +49 40 42838 7341; Fax: +49 40 42838 7332; Email: margraf@zbh.uni-hamburg.de

MATERIALS AND METHODS

Input data and library

The server takes the coordinates of a protein chain in PDB format and an email address for sending results to. The only adjustable parameter is the number of aligned structures to return.

Output of the web server

The server sends a rather minimal mail message as its result. It contains only a link to a temporary web page (lifetime 1 week) containing a list of candidate structurally related proteins. Selecting a candidate brings up a view of the superposition using Jmol (http://jmol.org) by E. Willighagen *et al*. (requires Java plugin). In another pane, the implied sequence alignment is shown, the superimposed coordinates can be downloaded and a list of more proteins with 90% or more sequence similarity to the candidate is given.

Each alignment is evaluated by scoring functions such as the alignment length, root mean squared difference (*rmsd*) of C^{α} atoms of aligned residues, a *z*-score calculated from a distribution of random alternative alignments (45), Smith and Waterman alignment scores (46) and a quality score based on the fraction of distance matrices which are similar between the query and aligned protein (45,47). This measure is used for the initial sorting of the list, but one can select a ranking by any of the other scores.

Processing method

Our method is a specialization of a very general technique which has been described in detail (13). Briefly, 1.5×10^6 fragments, each of six residues were clustered into 308 classes, each of which is a set of six bivariate Gaussian distributions for backbone ϕ and ψ angles. The more populated classes are recognizable as classic secondary structure, while the less populated classes are simply pieces of common protein motifs. Given a query fragment, one can calculate its probability of being in each of the classes, resulting in a long list (vector) of probabilities. A typical fragment may have a probability near 1.0 of being in some class, but even an unusual fragment will have some characteristic pattern of probabilities. Any two fragments can be compared by taking the dot product of these probability vectors which leads to the final alignment method as previously described (13). A similarity matrix is built based on all overlapping fragments from each protein. The scores associated with a residue come from all the fragments which it is a part of, so for fragments of length $k = 6$, a residue is sensitive to an environment of $2k - 1 = 11$ residues. The residue alignment can be read out from a conventional dynamic programming calculation (46,48) and superpositions are computed based on the aligned C^{α} atoms (49).

The method is fast since probabilities associated with databank proteins are precalculated and updated weekly. The similarity score has no hard thresholds, so the method fares well even when faced with slightly unusual structures. We give an example of this property below. Technically, it is interesting to note that the *rmsd* in Cartesian coordinates is never used during the alignment, so the method will find similarities even when confronted with domain or hinge-bending movements.

The server does not search all proteins in the protein data bank, but rather a subset of $<2 \times 10^5$ is chosen so that no two chains have >90% sequence identity (50).

RESULTS

Precision of search results

Results from the structure similarity servers usually differ from another in two main ways. First, the length of alignments is rarely the same from two different programs. Second, there is some concept of sensitivity. For some query, related proteins should be ranked higher than unrelated proteins. There is, however, often no correct answer when relationships are weak. Rather than debate this, we have simply taken SCOP (4–7) as a reference. It is also rather easy to find query proteins which suit a particular method. Rather than try to be objective, we give an example which suits SALAMI, one where all methods perform well and one where SALAMI performs poorly.

Figures 1–3 show plots of the precision of SALAMI, DALI (51) and VAST (52). We considered up to 100 related proteins from each server for each query and filtered out all chains which were not classified by SCOP. Chains which contained a domain in the same superfamily as a domain in the query chain were considered to be true positives. The remaining chains were regarded as false positives. The plots show the fraction of true positives at each rank.

First, Figure 1 shows the results using 1WOT as a query. This protein clearly suits SALAMI. VAST finds the four closest relatives. DALI, however has more interesting behaviour with a large number of false positives near the middle of the list. The structure has three α-helices joined by some small β-strands. In SCOP, it is placed in the Nucleotidyltransferase superfamily. There is, however, a set of proteins in the KH-domain superfamilies with a similar fold which can be superimposed surprisingly well. They are declared to be unrelated in SCOP, but they score well in DALI.

Figure 2 shows all the three methods performing equally well for 1QLW from the superfamily of alpha/beta hydrolases. Here, all results are in near perfect agreement with the SCOP classification. Only the SALAMI server includes a few false positives towards the end of the list.

Finally, Figure 3 shows the results with 1WK2 from the PUA domain-like superfamily as the query. This does not suit the SALAMI server. It is a mostly β protein, but more than 30 of its 121 residues are missing. The correct relatives are pushed down the ranking by unrelated proteins. DALI and VAST still perform well here because their similarity scores are much more influenced by spatial distances to elements which are not necessarily close in sequence.

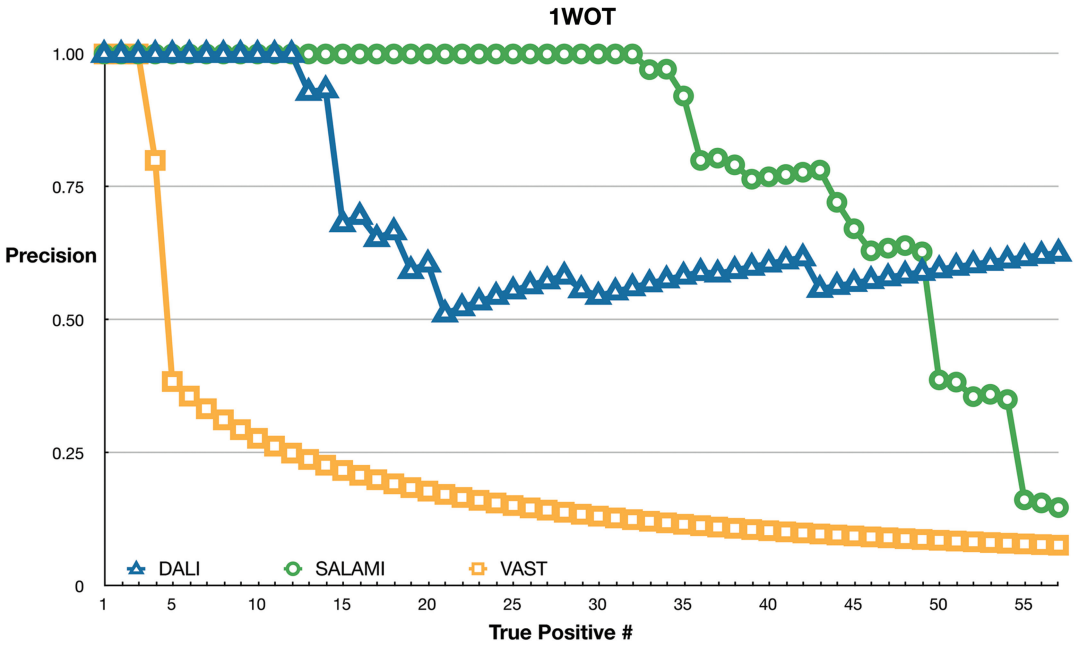

Figure 1. Sensitivity of servers using 1WOT as a query. For each rank on *x*-axis, each point shows the number of true positives divided by the rank. Servers (DALI, VAST and SALAMI) are marked as shown in the key. Lines joining the points have no meaning and only serve to guide the eye.

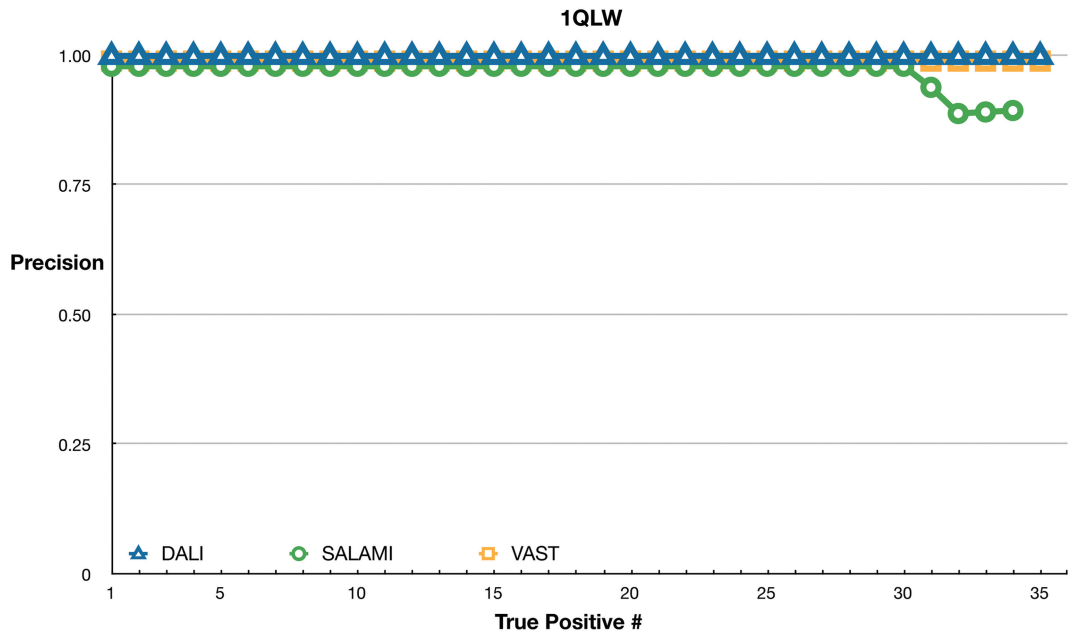

Figure 2. Sensitivity of servers using 1QLW as a query. Markers and servers as in Figure 1.

DISCUSSION AND CONCLUSION

The few results are certainly no benchmark. They are, however, clear examples of the ways different methods will work well with different query structures.

SALAMI has the disadvantage that it relies on chain connectivity and can be confused by broken structures. This means it may not be very useful for the broken skeletons that one can encounter in crystallographic structures with initial phasing. SALAMI has the advantage that it relies on chain connectivity and has no problem finding similarities when there are hinge-bending or domain motions. The graduated similarity measures mean that poor quality structures and deviations from regular geometry are well treated (13).

The methodology here has another interesting property. The graduated measure of similarity leads to a scoring

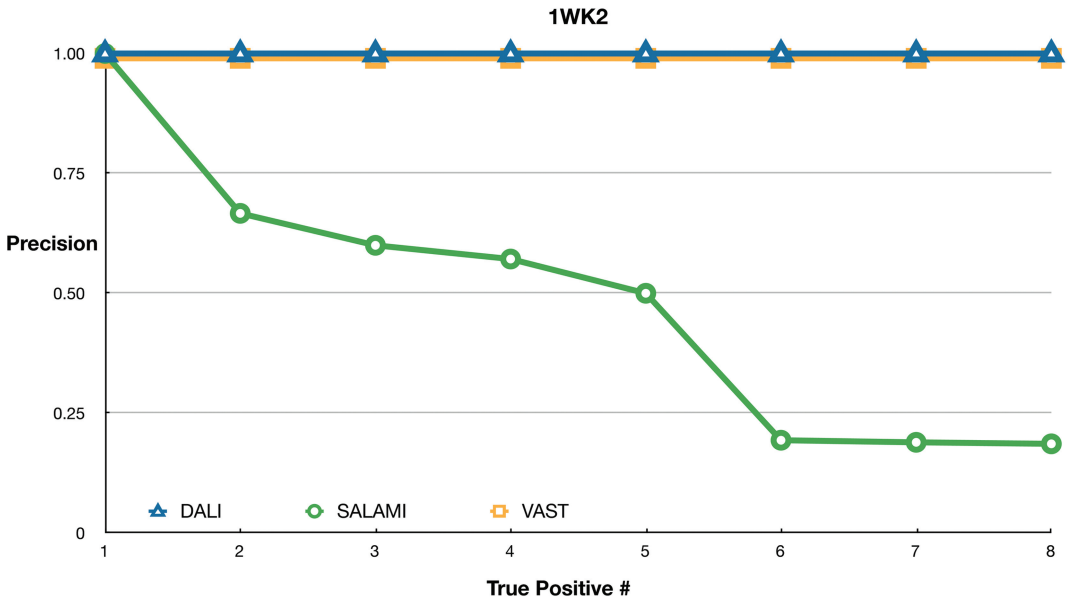

Figure 3. Sensitivity of servers using 1WK2 as a query. Markers and servers as in Figure 1.

function which is reliable and applies to any kind of structural unit. The use of a dynamic programming method then guarantees that the alignments are optimal within this scoring function. This, together with the good results for difficult structures and the flexible interface make it a valuable alternative to existing webservers.

FUNDING

Funding for open access charge: University of Hamburg.

Conflict of interest statement. None declared

REFERENCES

1. 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.
2. Holm,L. and Sander,C. (1996) Mapping the protein universe. *Science*, **273**, 595–602.
3. Holm,L. and Sander,C. (1996) The FSSP database: fold classification based on structure structure alignment of proteins. *Nucleic Acids Res.*, **24**, 206–209.
4. Andreeva,A., Howorth,D., Chandonia,J.-M., Brenner,S.E., Hubbard,T.J.P., Chothia,C. and Murzin,A.G. (2008) Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.*, **36**, D419–D425.
5. Andreeva,A., Howorth,D., Brenner,S.E., Hubbard,T.J.P., Chothia,C. and Murzin,A.G. (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res.*, **32**, D226–D229.
6. Conte,L.L., Brenner,S.E., Hubbard,T.J.P., Chothia,C. and Murzin,A.G. (2002) SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res.*, **30**, 264–267.
7. 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.
8. Cuff,A.L., Sillitoe,I., Lewis,T., Redfern,O.C., Garratt,R., Thornton,J. and Orengo,C.A. (2009) The CATH classification revisited–architectures reviewed and new ways to characterize structural divergence in superfamilies. *Nucleic Acids Res.*, **37**, D310–D314.
9. Scheeff,E. and Bourne,P. (2005) Structural evolution of the protein kinase-like superfamily. *PLoS Comp. Biol.*, **1**, E49.
10. Eidhammer,I., Jonassen,I. and Taylor,W. (2000) Structure comparison and structure patterns. *J. Comput. Biol.*, **7**, 685–716.
11. Feng,Z.K. and Sippl,M.J. (1996) Optimum superimposition of protein structures: ambiguities and implications. *Fold Des.*, **1**, 123–132.
12. Godzik,A. (1996) The structural alignment between two proteins: is there a unique answer? *Protein Sci.*, **5**, 1325–1338.
13. Schenk,G., Margraf,T. and Torda,A.E. (2008) Protein sequence and structure alignments within one framework. *Algorithms Mol. Biol.*, **3**, 4.
14. Mosca,R., Brannetti,B. and Schneider,T. (2008) Alignment of protein structures in the presence of domain motions. *BMC Bioinformatics*, **9**, 352.
15. Mosca,R. and Schneider,T. (2008) RAPIDO: a web server for the alignment of protein structures in the presence of conformational changes. *Nucleic Acids Res.*, **36**, W42–W46.
16. Zuker,M. and Somorjai,R. (1989) The alignment of protein structures in three dimensions. *Bull. Math. Biol.*, **51**, 55–78.
17. Russell,R.B. and Barton,G.J. (1992) Multiple protein sequence alignment from tertiary structure comparison. *Proteins*, **14**, 309–323.
18. Holm,L. and Sander,C. (1993) Protein Structure Comparison by Alignment of Distance Matrices. *J. Mol. Biol.*, **233**, 123–138.
19. Subbiah,S., Laurents,D. and Levitt,M. (1993) Structural similarity of DNA-binding domains of bacteriophage repressors and the globin core. *Curr. Biol.*, **3**, 141–148.
20. Alexandrov,N.N. (1996) SARFing the PDB. *Protein Eng.*, **9**, 727–732.
21. Gibrat,J.F., Madej,T. and Bryant,S.H. (1996) Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.*, **6**, 377–385.
22. Orengo,C.A. and Taylor,W.R. (1996) SSAP: sequential structure alignment program for protein structure comparison. *Methods Enzymol.*, **266**, 617–635.
23. Suyama,M., Matsuo,Y. and Nishikawa,K. (1997) Comparison of protein structures using 3D profile alignment. *J. Mol. Evol.*, **44**, S163–S173.
24. Shindyalov,I. and Bourne,P. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.*, **11**, 739–747.
25. Holm,L. and Park,J. (2000) DaliLite workbench for protein structure comparison. *Bioinformatics*, **16**, 566–567.

26. Jung,J. and Lee,B. (2000) Protein structure alignment using environmental profiles. *Protein Eng.*, **13**, 535–543.
27. Lackner,P., Koppensteiner,W.A., Sippl,M.J. and Domingues,F.S. (2000) ProSup: a refined tool for protein structure alignment. *Protein Eng.*, **13**, 745–752.
28. Ortiz,A.R., Strauss,C.E.M. and Olmea,O. (2002) MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci.*, **11**, 2606–2621.
29. Shatsky,M., Nussinov,R. and Wolfson,H.J. (2002) Flexible protein alignment and hinge detection. *Proteins*, **48**, 242–256.
30. Blankenbecler,R., Ohlsson,M., Peterson,C. and Ringnér,M. (2003) Matching protein structures with fuzzy alignments. *Proc. Natl Acad. Sci. USA*, **100**, 11936–11940.
31. Kawabata,T. (2003) MATRAS: a program for protein 3D structure comparison. *Nucleic Acids Res.*, **31**, 3367–3369.
32. Ilyin,V.A., Abyzov,A. and Leslin,C.M. (2004) Structural alignment of proteins by a novel TOPOFIT method, as a superimposition of common volumes at a topomax point. *Protein Sci.*, **13**, 1865–1874.
33. Krissinel,E. and Henrick,K. (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr. D Biol. Crystallogr.*, **D60**, 2256–2268.
34. Ochagavia,M.E. and Wodak,S.J. (2004) Progressive combinatorial algorithm for multiple structural alignments: application to distantly related proteins. *Proteins*, **55**, 436–454.
35. Shapiro,J. and Brutlag,D. (2004) FoldMiner and LOCK 2: protein structure comparison and motif discovery on the web. *Nucleic Acids Res.*, **32**, W536–W541.
36. Carpentier,M., Brouillet,S. and Pothier,J. (2005) YAKUSA: a fast structural database scanning method. *Proteins*, **61**, 137–151.
37. Chen,Y. and Crippen,G.M. (2005) A novel approach to structural alignment using realistic structural and environmental information. *Protein Sci.*, **14**, 2935–2946.
38. Zhang,Y. and Skolnick,J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.*, **33**, 2302–2309.
39. Zhu,J.H. and Weng,Z.P. (2005) FAST: a novel protein structure alignment algorithm. *Proteins*, **58**, 618–627.
40. Konagurthu,A.S., Whisstock,J.C., Stuckey,P.J. and Lesk,A.M. (2006) MUSTANG: a multiple structural alignment algorithm. *Proteins*, **64**, 559–574.
41. Lisewski,A.M. and Lichtarge,O. (2006) Rapid detection of similarity in protein structure and function through contact metric distances. *Nucleic Acids Res.*, **34**, E152.
42. Taubig,H., Buchner,A. and Griebsch,J. (2006) PAST: fast structure-based searching in the PDB. *Nucleic Acids Res.*, **34**, W20–W23.
43. Oldfield,T.J. (2007) CAALIGN: a program for pairwise and multiple protein-structure alignment. *Acta Crystallogr. D Biol. Crystallogr.*, **63**, 514–525.
44. Tyagi,M., Gowri,V.S., Srinivasan,N., deBrevern,A.G. and Offmann,B. (2006) A substitution matrix for structural alphabet based on structural alignment of homologous proteins and its applications. *Proteins*, **65**, 32–39.
45. Torda,A.E., Procter,J.B. and Huber,T. (2004) Wurst: a protein threading server with a structural scoring function, sequence profiles and optimised substitution matrices. *Nucleic Acids Res.*, **32**, W532–W535.
46. Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.
47. Russell,A.J. and Torda,A.E. (2002) Protein sequence threading: averaging over structures. *Proteins*, **47**, 496–505.
48. Gotoh,O. (1982) An improved algorithm for matching biological sequences. *J. Mol. Biol.*, **162**, 705–708.
49. Diamond,R. (1988) A note on the rotational superposition problem. *Acta Cryst.*, **A44**, 211–216.
50. Li,W.Z., Jaroszewski,L. and Godzik,A. (2001) Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics*, **17**, 282–283.
51. Holm,L., Kääriäinen,S., Rosenström,P. and Schenkel,A. (2008) Searching protein structure databases with DaliLite v.3. *Bioinformatics*, **24**, 2780–2781.
52. Gibrat,J.-F., Madej,T. and Bryant,S. (1996) Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.*, **6**, 377–385.