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

© 2009 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
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 $C^\alpha$ 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, $1.5 \times 10^6$ fragments, each of six residues were clustered into 308 classes, each of which is a set of six bivariate Gaussian distributions for backbone $\phi$ and $\psi$ 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 $C^\alpha$ 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 $\alpha$-helices joined by some small $\beta$-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 $\beta$ 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.
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 web servers.

FUNDING
Funding for open access charge: University of Hamburg.

Conflict of interest statement. None declared

REFERENCES