BIOINFORMATICS APPLICATIONS NOTE

Vol. 24 no. 23 2008, pages 2780–2781
doi:10.1093/bioinformatics/btn507

Structural bioinformatics

Searching protein structure databases with DaliLite v.3

L. Holm[1,2,*], S. Kääriäinen[2], P. Rosenström[2] and A. Schenkel[2]

[1]Department of Biological and Environmental Sciences and [2]Institute of Biotechnology, P.O.Box 56 (Viikinkaari 5), 00014 University of Helsinki, Finland

Received on June 10, 2008; revised on August 14, 2008; accepted on September 22, 2008

Advance Access publication September 25, 2008

Associate Editor: Thomas Lengauer

ABSTRACT

The Red Queen said, ‘It takes all the running you can do, to keep in the same place.’ Lewis Carrol

Motivation: Newly solved protein structures are routinely scanned against structures already in the Protein Data Bank (PDB) using Internet servers. In favourable cases, comparing 3D structures may reveal biologically interesting similarities that are not detectable by comparing sequences. The number of known structures continues to grow exponentially. Sensitive—thorough but slow—search algorithms are challenged to deliver results in a reasonable time, as there are now more structures in the PDB than seconds in a day. The brute-force solution would be to distribute the individual comparisons on a massively parallel computer. A frugal solution, as implemented in the Dali server, is to reduce the total computational cost by pruning search space using prior knowledge about the distribution of structures in fold space. This note reports paradigm revisions that enable maintaining such a knowledge base up-to-date on a PC.

Availability: The Dali server for protein structure database searching at http://ekhidna.biocenter.helsinki.fi/dali_server is running DaliLite v.3. The software can be downloaded for academic use from http://ekhidna.biocenter.helsinki.fi/dali_lite/downloads/v3.

Contact: liisa.holm@helsinki.fi

1 INTRODUCTION

Comparative analyses of protein sequences and structures are a cornerstone of bioinformatics. When sequence and structure similarities have an evolutionary origin, it is often possible to infer similarities in the biological functions of the proteins, which would be difficult to predict directly. Structure comparisons have a longer look-back time than sequence comparison and have led to the identification of many ‘super-families’ of distantly related proteins.

Many measures have been proposed to quantify structural similarity. The Dali method uses a weighted sum of similarities of intra-molecular distances, which correlates with expert classifications in the sense that the structures of homologous proteins typically get higher similarity scores than the structures of evolutionarily unrelated proteins (Sierk and Pearson, 2004). This property is useful to a biologist using structure comparison to learn more about her query protein: the biologically informative neighbours are found at the top of the match list with relatively few false leads.

The Dali method has been used to systematically scan new structures against the Protein Data Bank (PDB) for some 15 years (Holm and Sander, 1994). The overall strategy is to screen the structure database with many different methods, starting with fast but unreliable ones and ending with the most sensitive but slow methods. This ensures that no significant similarity is missed. The search space is pruned between methods; if a strong match has been found, then subsequent methods only compare the query structure to the neighbours of the strong match. This strategy requires that all the neighbours of the known structures are precomputed in all versus all fashion within a representative subset of structures. The size of the structure set has grown by two decades since the system was introduced, and all versus all comparison is a quadratic problem in the number of structures. Recently, the paradigm of all versus all comparisons became untenable when the weekly PDB updates began to take more than a week to process.

DaliLite is a standalone package of the Dali algorithm. The first release of DaliLite (Holm and Park, 2000) contained all the functionality of the Dali server at EBI except the site-specific, complicated database update protocol. The main DaliLite program is a wrapper that calls a variety of methods for protein structure comparison. New workflows can thus be easily implemented by ‘rewiring the regulatory logic’ but keeping the basic algorithms unchanged. In DaliLite v.3, we introduce new options for database searching (DaliLite –quick) and database updates (DaliLite –update). The new protocols improve server throughput and vastly simplify the updates, making the complete system portable.

The key change from earlier is that we abandon the all versus all matrix of similarities in favour of a connected graph of similarities. The nodes of the graph represent protein structures and edges represent structural alignments. Whereas before each representative structure was directly linked to all its structurally similar neighbours, we now require only that there is a path of continuous structural similarity through the graph. The structural neighbours of a query structure are collected by walks through the graph. Not only need the graph be less densely connected than the all versus all matrix, thus saving computational effort, but also there is the added benefit that the incremental updates of the structural similarity graph and the choice of structural representatives are completely decoupled.

2 METHODS

2.1 PDB clustering

The PDB is highly redundant. The structures of some proteins and their mutants have been determined in various conditions, though the structures

[*]To whom correspondence should be addressed.

Table 1. Comparison of DaliLite v.3, the SSM server and SCOP

Query	Dali −q time (min)	Dali −q P	Dali −q Z AUC	Dali −l time (min)	Dali −l Z AUC	SSM Q AUC	SCOP T	SCOP class
1c52	14	241	**0.885**	297	0.822	0.674	219	a.3.1.
1sfxA	12	401	**0.757**	357	0.733	0.386	331	a.4.5.
1azu	14	529	0.967	369	**0.969**	0.516	289	b.6.1.
1wk2A	7	49	0.800	201	**0.966**	0.197	60	b.122.1
3be7A	51	542	**0.990**	1861	0.968	0.843	135	c.1.9.
1qlwA	30	414	**0.908**	1211	0.897	0.576	456	c.69.1.
2baa	18	726	**0.693**	486	0.354	0.025	979	d.2.1
1wotA	13	500	0.558	251	**0.642**	0.009	179	d.218.1

AUC, area under the curve of reliability (TP/P) versus coverage (TP/T), where Ts (‘true’) are members of the same SCOP (Murzin *et al*., 1995) superfamily as the query structure, Ps (‘positive’) are the top *n* matches from the ranked list for varying *n* and TPs (‘true positive’) are the intersection of sets T and P. Only PDB entries classified in SCOP v.1.73 were evaluated and each PDB entry was counted once. SSM server (http://www.ebi.ac.uk/msd-srv/ssm/cgi-bin/ssmserver; Krissinel and Henrick, 2004) parameters were set to 10 : 10 and highest precision. The DaliLite search was performed using the –list (Dali –l) or –quick option (Dali –q with MAX_HITS = 1000) and reporting matches with $Z > 2$.

remain the ‘same’ for classification purposes. We use a representative subset at 90% sequence identity level (PDB90), derived from the current set of PDB sequences using CD-HIT (Li and Godzik, 2006). The PDB contains over 100 000 structures (chains), which is reduced to about 20 000 PDB90 representatives. Further clustering of similar folds at lower levels of sequence identity was not cost effective.

2.2 Structural similarity graph

The structural similarity graph and alignment data are stored in a relational database (MySQL). The graph is updated incrementally. If a new structure has strong similarity to structures already in the graph, one edge is sufficient to connect the new structure to the graph in the proper neighbourhood. If there is no strong match, we compare the new structure to all existing structures and add edges for all significant similarities.

Similarity is measured by Dali Z-scores. ‘Significant similarities’ have a Z-score above 2; they usually correspond to similar folds. ‘Strong matches’ have sequence identity above 20% or a Z-score above a cutoff that depends on the size of the query protein. The Z-score cutoff was empirically set to $n/10 - 4$, where n is the number of residues in the query structure. We additionally require that the complete structure is covered by structural alignments; a segment of the query structure longer than 80 residues without any structural matches always disqualifies a strong match.

2.3 Database searching

The database search option DaliLite –quick compares a query structure to all structures in the PDB, as organized in the structural similarity graph. To initiate a transitive search of structures in the graph, the query structure must be attached to *some* structural neighbours. Fast feature filters are often successful in finding near neighbours. We currently use sequence comparison by Blast, GTG sequence motifs (Heger *et al*., 2007) and secondary structure triplets to rank the structures in PDB90. We convert the feature filter scores to Z-scores in order to combine the ranked lists. The top 100 structures are compared using the normal Dali procedures. If a strong match is found, we move to the next step (transitive alignment). Otherwise, the query structure is compared against all 20 000 structures in PDB90.

The entry points connect the query structure to one or more structures in the structural similarity graph. These are direct (first shell) neighbours of the query. Structures in the second shell are compared in batches of 100, selecting those with the strongest connections first. Connection strength is the lesser Z-score along the path from query to the first neighbour to the second neighbour. The transitive alignment (via first neighbour) between the query structure and second neighbour is used as starting point for refinement, skipping the costly alignment optimization from scratch. The expansion is repeated until the connection strength drops below a Z-score cutoff of 2, or a maximum number of matches have been reported (default: MAX_HITS = 500).

3 RESULTS

The utility of a protein structure database search method (i.e. similarity measure and optimization algorithm) must depend on its ability to report back ‘interesting’ matches. As an illustration, we chose query and target structures representing diverse superfamilies from the four main structural classes in SCOP: cytochromes c and winged helix DNA-binding domains from the all-alpha class, cupredoxins and PUA-like domains from the all-beta class, metallo-dependent hydrolases and alpha/beta hydrolases from the alpha/beta class, and lysozyme-likes and nucleotidyltransferases from the alpha + beta class (Table 1). Match lists were evaluated using the AUC, where the maximum value of one indicates perfect sensitivity and selectivity. Compared to optimizing the alignment from scratch (DaliLite –list), the new transitive search mode (DaliLite –quick) is about 30 times faster, without affecting AUC much (we removed all pre-existing edges from the query structures to the structural similarity graph). Compared to the SSM server’s Q-score, the higher AUC values in Table 1 indicate superior discrimination of homologous proteins from unrelated proteins by Dali’s Z-score.

In conclusion, Dali remains a useful tool for structural bioinformatics. The Dali server has been running DaliLite –quick for a number of months now, with a throughput of 50 user queries—a mixture of redundant and unique structures—per day per CPU.

Funding: Academy of Finland (grants #109849 and #1105210).

Conflict of Interest: none declared.

REFERENCES

Heger,A. *et al*. (2007) The global trace graph, a novel paradigm for searching protein sequence databases. *Bioinformatics*, **23**, 2361–2367.

Holm,L. and Park,J. (2000) DaliLite workbench for protein structure comparison. *Bioinformatics*, **16**, 566–567.

Holm,L. and Sander,C. (1994) Searching protein structure databases has come of age. *Proteins*, **19**, 165–173.

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.

Murzin,A.G. *et al*. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol*., **247**, 536–540.

Krissinel,E. and Henrick,K. (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Cryst*., **D60**, 2256–2268.

Sierk,M.L. and Pearson,W.R. (2004) Sensitivity and selectivity in protein structure comparison. *Protein Sci*., **13**, 773–785.