per se remains a distinct advantage. Many structures are still not deposited in the PDB, and for these cases the only solution is to have good memory (although the SCOP database is now extremely useful for these cases[18]).

Nevertheless, it is now more likely for a newly determined structure to adopt a familiar fold rather than a novel one.[16] Researchers who fail to use these procedures not only run the risk of missing structural similarities with their own structure, but also miss the concomitant functional and evolutionary relationships as well. There are now many programs available that have been shown to be effective and that provide the novice with an ideal introduction to the field. In addition there are a number of groups specializing in this work, who are always willing to collaborate, as well as a growing number of information sources available on the World Wide Web. With these facilities in place, we are now in an ideal position to exploit the mass of structural data anticipated in the near future.

Availability

The SCOP database of manually detected structural similarities is described in Chapter 37. The automatically classified CATH database is at http://www.biochem.ucl.ac.uk/bsm. In addition, database searches can be made using sarf2 at http://www.ncifcrf.gov/~nicka/info.html. I have additional data available at http://www.biochem.ucl.ac.uk/~swintech.

[18] A. G. Murzin, S. E. Brenner, T. Hubbard, and T. Chothia, *J. Mol. Biol.* **247,** 536 (1995).

[39] Alignment of Three-Dimensional Protein Structures: Network Server for Database Searching

By LIISA HOLM and CHRIS SANDER

Introduction

Access to computational services and biological databases over the Internet, in particular through the World Wide Web, is an increasingly important research tool for the biochemist. A major use of molecular biology databases involves searching for evolutionary links that allow transfer of functional information about one protein family to another. An increasing number of distant evolutionary relationships that are not evident by sequence comparison are being revealed by similarity of three-dimensional (3D) protein structures, both because of a rapid increase in the number of known structures and because of improved methods of detection. For example, structure comparison has revealed surprising biochemical similarities between urease and adenosine deaminase[1] and between glycogen phosphorylase and a DNA glucosyltransferase from phage T4[2] that were not detected by sequence comparison. These are just two of a long list of examples[3] to illustrate the evolutionary principle of adapting structural motifs that support particular active-site constellations to different functional roles in diverse cell types and organisms.

A large number of automated methods for protein structure comparison that use different representations of structure, different definitions of similarity, and various optimization algorithms have been developed (reviewed in Ref. 3). This chapter describes the Dali method,[4] which is a general approach for aligning a pair of proteins represented by two-dimensional matrices. The implementation of prefilters to speed up database searches has enabled us to provide Internet access using either World Wide Web software addressing http://www.embl-heidelberg.de/dali/ or electronic mail to dali@embl-heidelberg.de.

Formulation of Problem

The utility of distance matrices, also called distance plots or distance maps, in describing and comparing protein conformations has been recognized for a long time. A distance matrix is a two-dimensional (2D) representation of 3D structure. The matrix is independent of the coordinate frame and contains more than enough information to reconstruct the 3D structure, except for overall chirality, by distance geometry methods. The most commonly used variant is that containing all pairwise distances between residue centers (i.e., $C\alpha$ atoms).

Distance matrices are useful in structure comparison because similar 3D structures have similar interresidue distances. Imagine a (transparent) distance map of one protein placed on top of that of another protein and then moved vertically and horizontally. Depending on the relative displacement of the matrices, matching substructures appear as patches (submatrices) in which the difference of distances is small. Matching patches centered on the main diagonals correspond to locally similar backbone

[1] E. Jabri, M. B. Carr, R. P. Hausinger, and P. A. Karplus, *Science* **268,** 998 (1995).

[2] L. Holm and C. Sander, *EMBO J.* **14,** 1287 (1995).

[3] L. Holm and C. Sander, *Proteins* **19,** 165 (1994).

[4] L. Holm and C. Sander, *J. Mol. Biol.* **233,** 123 (1993).

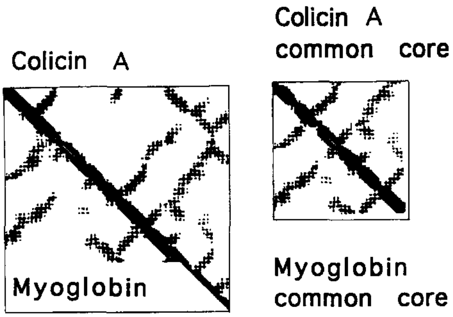

FIG. 1. Comparison of distance matrices. Structures similar in three dimensions necessarily have a similar set of intramolecular distances. The common folding pattern of colicin A and myoglobin is highlighted in the collapsed matrix (right) that brings the two sequences into register through deleting incompatible rows and columns from the full distance matrices. Formally, the Dali score [see Eq. (1)] is a weighted sum over similarities of distances between residue centers in a common core. The distance matrices depicted here have a black dot for Cα–Cα distances shorter than 12 Å. Helices show up as thick bands along the diagonal and helix pairs as black bands parallel or orthogonal to the diagonal.

conformations (i.e., secondary structures). Matches of short distances found off the main diagonals reveal similar tertiary structure contacts. The presence of a common structural motif made up of several disjoint regions of the backbone becomes visible at one glance in a pair of "collapsed" submatrices that are obtained by deleting residues with no structural equivalent in the other structure (Fig. 1). Allowing permutations in the order of rows and columns leads to detection of spatial similarities in protein structures when topological connectivities differ. An advantage of the 2D representation used in the Dali method over rigid-body 3D superimposition is that local conservation of structure is not masked by shifts in the relative positions of structural elements, for example, as a result of hinge motion of domains.

A quantitative solution to the geometrically complicated problem of comparing protein shapes requires a precise definition of similarity of protein structures. In the Dali method, the structural similarity S (Dali score) is defined as the following weighted sum:

$$S = \sum_i \sum_j \left[\left(\partial - \frac{|d_{ij}^{\mathrm{A}} - d_{ij}^{\mathrm{B}}|}{d_{ij}^{*}} \right) e^{-(d_{ij}^{*}/\mu)^2} \right] \qquad (1)$$

where the summation is over all residues i, j of the common core and d_{ij}^{*} denotes the arithmetic average of the Cα–Cα distances d_{ij}^{A} and d_{ij}^{B} in

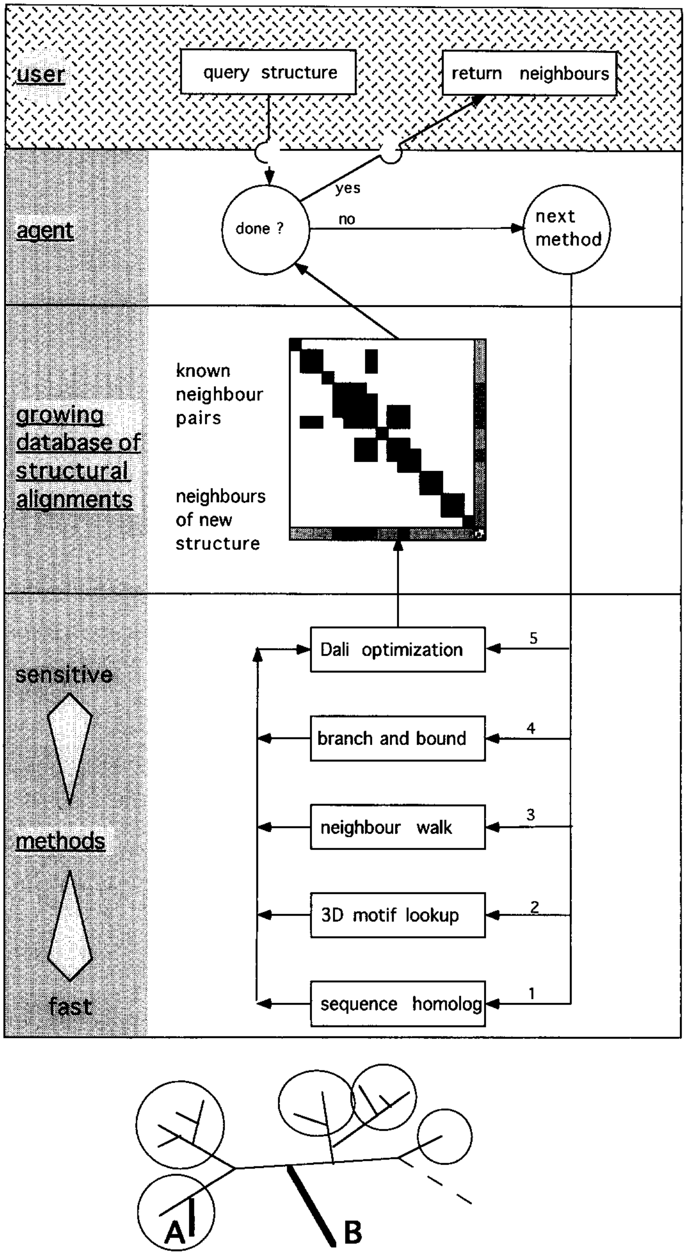

FIG. 2. System architecture. The database search system is implemented in three layers. At the top of the system is an "intelligent agent," actually a Perl script, that receives a query structure from the user and returns the list of structural neighbors of the query structure.

proteins A and B. In Eq. (1) ∂ is the threshold of similarity and is set to a relative deviation of 0.20 (20%). This means that, for example, adjacent strands in a β sheet (typical distance 4–5 Å) should match to within 1 Å, while 2–3 Å displacements are well tolerated for strand–helix or helix–helix contacts (typical distances 8–15 Å). The exponential factor downweights contributions from pairs in the long-distance range. We chose $\mu = 20$ Å calibrated on the size of a typical domain.

The set of equivalences between residues in A and B that maximizes S defines the common core of proteins A and B. Optimization of several nonoverlapping alignments in parallel leads to automatic detection of, for example, internal repeats. As similarity is quantified at the residue level, the resulting structural alignments can be directly linked to sequence and evolutionary comparisons.

Database Searches

The optimization of assignments for equivalent residue pairs looks simple graphically (Fig. 1) but is computationally hard because of the complicated combinatorics. In the original Dali method, matches are built up by combining small submatrices with similar distance patterns and using a Monte Carlo algorithm for optimization. The approach was shown to be robust and to yield accurate alignments.[4] In database searching, one is generally only interested in a few top hits so that sensitive pairwise comparison against the bulk of the database is unnecessarily costly. For the network server version of Dali, we have implemented efficient screening steps that work with approximations at the level of secondary structure elements[5] and as a result increase the speed of comparison from 5–10 min per protein pair to, in favorable cases, 5–10 min for scanning one structure against all structures in the protein database. The increase in speed is approximately 500-fold. The complete database search system is outlined in Fig. 2.

[5] L. Holm and C. Sander, *in* "Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology" (C. Rawlings, D. Clark, R. Altman, L. Hunter, T. Lengauer, and S. Wodak, eds.), pp. 179–187. AAAI Press, Menlo Park, California, 1995.

The task can be performed efficiently using stored knowledge of the structural neighborhoods of all proteins in the PDB (FSSP database) and a hierarchy of different methods (mostly FORTRAN programs). The goal is to place a query structure in the proper neighborhood of fold space, which is illustrated in the form of a tree at the bottom. Fast filters efficiently detect "trivial" similarities (branch labeled A). Only if the query structure cannot be mapped to a known neighborhood (branch labeled B) is it necessary to test its similarity to each one of the known structures in order to position properly the new structure, a more time-consuming task.

Finding all structural neighbors in a database of thousands of proteins is simplified by defining a hierarchy of neighborhoods in protein fold space. The tightest clusters of folds are formed by sequence homologs (sequence identity above 25%[6]), and each such family is represented by a single member. Remote functional homologs (proteins with sequence identity below 25%) typically are more conserved in structure than pairs of proteins that are unrelated in function and only have similar folding topology. An example is the case of the functionally related myo-, hemo-, and leghemoglobins, which form one group among three about equally distant structural classes, where the other two are phycocyanins and colicin A.[7]

The database search system exploits these observations on clusters in fold space by first trying so-called cheap and quick filters for identifying trivial hits. For example, a new globin structure or another mutant of T4 lysozyme would map to an already characterized neighborhood of fold space. Sophisticated and sensitive methods are reserved for the potentially unique structures that require charting new regions of fold space. Because the screening methods use different approximations of the proper Dali score, it is important to put all pairs, whether they come via the fast or slow route, on the same footing for consistency of the final result. This is done by passing all prealignments produced by the different methods through a Monte Carlo algorithm that optimizes complete alignments with respect to the Dali score.

Biological Meaning of Structural Similarity

We have empirically determined the background strength of similarity as a function of chain length. The statistical significance of a database hit relative to the background is reported as a *Z* score (score minus mean divided by standard deviation). In particular, the rescaling so obtained provides a general and quantitative definition of structural neighborhoods. For example, using *Z* scores raises a database match to the SH3-like domain of biotin repressor/biotin holoenzyme synthetase from rank 58 in Dali score to rank 3 in *Z* score, compensating for the effects of very different domain sizes (Table I). In reporting results from structure database searches, we list pairs of proteins or domains[8] for which the *Z* score is above 2.

It is well established that protein folds are better conserved in the course

[6] C. Sander and R. Schneider, *Proteins* **9,** 56 (1991).

[7] L. Holm and C. Sander, *FEBS Lett.* **315,** 301 (1993).

[8] L. Holm and C. Sander, *Proteins* **19,** 256 (1994).

TABLE I
IDENTIFICATION OF SIMILAR DOMAIN FOLDS

Domain description	Size of match/ size of domain, residues	Best database hit[a]	Dali score (rank[b])	*Z* score (rank[b])
N-terminal, DNA-binding	59/64	LexA repressor	258 (14)	6.7 (1)
Middle, catalytic	129/181	Seryl-tRNA synthetase	523 (1)	3.6 (6)
C-terminal, SH3-like fold	46/47	Photosystem I accessory protein (psaE)	152 (58)	4.2 (3)

[a] Structurally most similar protein in a sequence-representative set of 557 3D structures.
[b] Ranks in structure comparison against sequence-representative set of 557 3D structures are given in parentheses.

of evolution than amino acid sequences. However, the smaller and simpler a folding motif is, the more frequent its recurrence between protein families without any apparent biological connection. This raises the question under which circumstances it is justified to base inferences of, say, biochemical mechanism on structural resemblance. Indicators of common descent include conserved active-site residues, a conserved structural framework

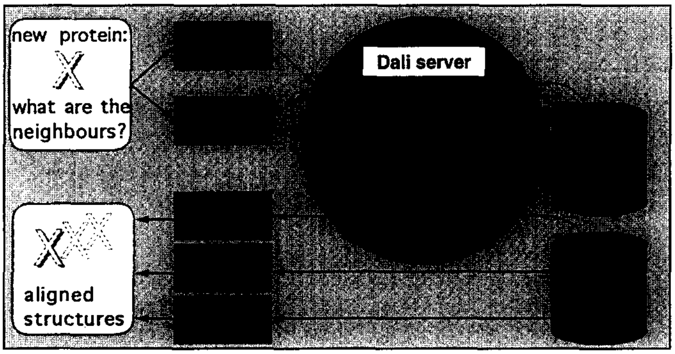

FIG. 3. Server traffic chart. The Dali server compares a query structure against the database of known structures (Protein Data Bank, PDB) and returns a list of structurally aligned neighbors in fold space. Newly solved proteins can be submitted for comparison by E-mail (dali@embl-heidelberg.de) or interactively from the WWW (home page URL http://www.embl-heidelberg.de/dali/dali.html). The results for exhaustive structure comparisons for proteins already deposited in the PDB are stored in the FSSP database of structural neighbors.[8] To get at this information over the WWW, point to the FSSP home page at URL http://www.embl-heidelberg.de/dali/fssp/. The neighbor lists in the FSSP database are updated whenever new structures are released by the PDB.

(a)

```
STRID2    Z     RMSD    LALI    LSEQ2  %IDE    PROTEIN
1kan-A 38.8    0.0     253     253     100     KANAMYCIN NUCLEOTIDYLTRANSFERASE
1kan-B 34.7    0.4     253     253     100     KANAMYCIN NUCLEOTIDYLTRANSFERASE
1bpb     5.8   4.0     106     242       9     DNA POLYMERASE BETA (BETA POLYME
1rpo     2.1   4.1      53      61       6     ROP (COLE1 REPRESSOR OF PRIMER)
```

(b)

```
1kan-A MNGPIIMTREERMKIVHEIKERILDKYGDDVKAIGVYGSLGRQTDGPYSDIEMMCVMSTEEAEFSHEW
              hhhhhhhhhhhhhhhhhs sseeeeeeeegggttt   tt  eeeeeee stt ee
1kan-B MNGPIIMTREERMKIVHEIKERILDKYGDDVKAIGVYGSLGRQTDGPYSDIEMMCVMSTEEAEFSHEW
              hhhhhhhhhhhhhhhhhhtttteeeeeebgggtss   ss    eeeees ss eee  e
1bpb   FEDFKRIPREEMLQMQDIVLNEVKKL.DPEY.IATVCGSFRRGAES..GDMDVLLTHPNFT..TKFMG
       ttggs eehhhhhhhhhhhhhhhhh  tt  eeee hhhhtt se  s eeeeee tt    seeee
1rpo   ....................................................................
```

(c)

```
swiss          FSSP
kanu_staau     1kan-A  MNGPIIMTREERMKIVHEIKERILDKYGDDVKAIGVYGSLGRQTDGPYSDIEMMCVMSTEEAEF
kanu_bacsp     1kan-A  MNGPIIMTREERMKIVHEIKERILDKYGDDVKAIGVYGSLGRQTDGPYSDIEMMCVMSTEEAEF
dpob_rat       1bpb    FEDFKRIPREEMLQMQDIVLNEVKKL~DPEY~IATVCGSFRRGAES~~GDMDVLLTHPNFT~~T
dpob_human     1bpb    FGDFKRIPREEMLQMQDIVLNEVKKV~DSEY~IATVCGSFRRGAWS~~GDMDVLLTHPSFT~~T
```

(d)

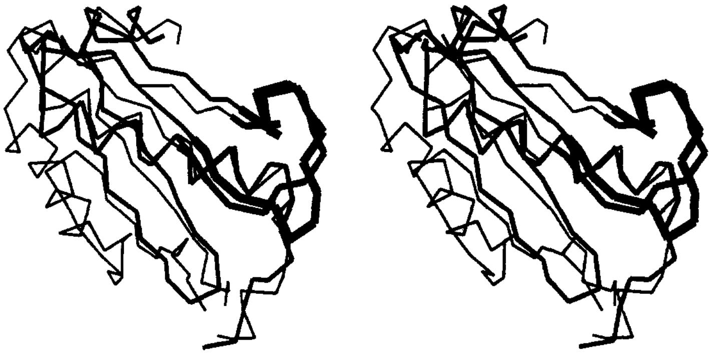

FIG. 4. Structures aligned with structures and structures aligned with sequences. (a) Ranked list of structural neighbors as a result from the database search using kanamycin nucleotidyltransferase (1kan-A) as the query structure. Hits in the database are identified by PDB code and chain (STRID2). The positional root-mean-square deviation of Cα atoms corresponding to the optimal Dali alignment (RMSD), the number of aligned residues (LALI), the length of the matched protein (LSEQ2), sequence identity among equivalenced residues (%IDE), and the name of the matched protein (PROTEIN) are given for each pair. The strongest structural similarity by *Z* score (column labeled Z) reveals a distant evolutionary connection between DNA polymerase beta (1bpb) and kanamycin nucleotidyltransferase.[10] (b) Structural alignment view loaded on the fly by the FSSP web server. The amino acid sequence and secondary structure (h, helix; e, strand; t, turn) are shown for each chain. Walking in fold space is possible by clicking underlined links. Only the structurally equivalent blocks are

around the active site, and similar biochemical function. The structure signal is usually captured by the *Z*-score ranking. The sequence signal often stands out when the structural alignment is expanded by sequence homologs from the HSSP database of sequence neighbors of structurally known proteins.[6]

Availability

The pool of known protein 3D structures is growing by hundreds of new ones each year, and systematic comparisons are needed that can keep track of all the interesting similarities. The Dali server can be used by X-ray crystallographers and nuclear magnetic resonance (NMR) spectroscopists at the last stage of structure determination to detect possible structural similarities with structures currently in the Protein Data Bank (PDB). Coordinates of new structures can be sent either by E-mail or interactively via the World Wide Web (WWW). The list of structural alignments of the query structure against all significantly similar neighbors in protein fold space is returned. The Dali search engine is also used to maintain the FSSP database[9] of precalculated structural neighbors for all structures released to the public through the Protein Data Bank (Fig. 3).

In time, the work of structural biologists will result in a complete survey of the role of protein structures in the evolution of biochemical complexity. The FSSP database of multiple structure comparisons provides a continuously updated structural classification.[9] The rich information contained in multiple structural alignments is best appreciated in graphical form. Public software (e.g., Rasmol) can be downloaded to look at superimposed 3D structure pairs and generate interactive pictures on a personal computer, and a web browser can be used to follow links from structure alignment

[9] L. Holm and C. Sander, *Nucleic Acids Res.* **22,** 3600 (1994).

shown (dots are gaps and trailing ends). Repressor of primer (ROP) (1rpo) matches to a helical domain at the C terminus and not to the N-terminal catalytic domain shown here. (c) Combining the power of structure comparison with that of multiple sequence alignment. DNA polymerase beta has been selected from the list of neighbors of kanamycin nucleotidyltransferase (1kan-A), and the structural alignment of these proteins is viewed in combination with sequence alignments from the HSSP database[6] (the parent structure is given in the FSSP column). The sequence identifiers (column labeled swiss) are linked to the SWISS-PROT database via the Sequence Retrieval System (SRS).[11] Only structurally equivalent segments are shown; excluded segments are marked by ~. (d) Structural superimposition of the Cα traces of kanamycin nucleotidyltransferase (thick lines) and DNA polymerase beta (thin lines).[10]

to sequence families, annotations of function, or literature references, facilitating a closer look at protein evolution (Fig. 4[10,11]).

Acknowledgments

We thank Antoine de Daruvar for writing the scripts managing the server traffic and Protein Data Bank updates.

[10] L. Holm and C. Sander, *Trends Biochem. Sci.* **20,** 345 (1995).
[11] T. Etzold and P. Argos, *Comput. Appli. Biosci.* **9,** 49 (1993).

[40] Converting Sequence Block Alignments into Structural Insights

By OLIVIER POCH and MARC DELARUE

Introduction

In the past decade, the number of available protein sequences has grown exponentially. In the same time, sequence analysis has become a major tool to gain insight on protein function in cellular processes from the sequence information. Many aspects of this sequence–structure–function relationship problem are not fully understood and are far from being solved by computational biology, but it is now possible to try to address problems as diverse as the localization of a protein in the cell (nucleus, organelles, membrane), the rate of protein degradation, putative posttranslational modifications, as well as some functional aspects such as binding properties (to other proteins, nucleic acids, cofactors, ions, etc.). Obviously, this list is far from being exhaustive, but it clearly highlights the fact that biologists can now expect major information about biological processes from even a single sequence.

One of the first issues to be addressed when undertaking structural and functional studies of a particular protein is to find out how many related sequences can be identified in a protein sequence database search. Even if structural data are not available for any member of a protein family, a multiple alignment will often offer much information that has to be analyzed carefully for future biological studies. The chance of hitting at least a homologous protein is constantly increasing as genome-sequencing projects proceed toward completion in different organisms (*Saccharomyces cerevisiae, Arabidopsis thaliana, Homo sapiens, Caenorhabditis elegans*). In addi-