DALI: Distance-matrix ALIgnment

Holm, L. and Sander, C. (1996)
Mapping the Protein Universe
Science vol. 273, 595-602.

The objective of shape comparison in DALI is to assign a one-to-one equivalence between the residues, where non-matching residues can be skipped in either chain.

This is done by finding similar patterns in distance matrices.

Constructing distance matrices (or “contact maps”) is easy;
finding maximal matching sub-matrices is hard.

Two algorithms in DALI

Scan for obvious similarities using a fast (but, in general, less accurate) algorithm, then rescan for more subtle similarities using more sophisticated (but slower) algorithms.

A) Fast heuristic 3D lookup (“hashing”)

Catches easy-to-find structural similarities.

Represent secondary structure elements by 3D line segments; match vector relationships from the query protein with a stored list; when enough matches are found with a database protein, sample a limited set of superpositions.

B) Branch-and-bound algorithm

Guaranteed to find the global optimum, but slower
(worst case: exponential number of steps).

Find the best matching sub-matrices for proteins A and B;
then recursively split the solution sub-space.

Shape comparison in DALI

(i) a suitable representation:

list of Cα atoms described by their x, y and z coordinates.

(ii) an objective function to be optimised:

accommodate the largest possible number of equivalent points within small deviations in position (typically less than 2 to 3 Å).

(iii) a comparison algorithm:

find matching sub-matrices and merge these into larger consistent blocks of agreement by removing intervening rows and columns.

(iv) appropriate decision rules:

statistical significance of comparison score (Z-score);
equivalent sets of residues (structural alignment);
3D view of the matched parts superimposed.

Problems when searching a protein structure database

(Want to perform all-against-all comparison)

Unequal representation of protein families.

Some redundancy can be eliminated by removing proteins with mutual sequence identity greater than 25%.
But many structurally similar proteins remain.

The problem of domains.

Similar sub-structures recur between several proteins.

Today we can identify sets of domains with distinct folds from resources like CATH and SCOP.

Fold recognition

The idea behind “threading”:

Imagine a wire wound into the shape of a known protein’s main chain “fold”.

Imagine next that our new sequence is represented by beads that are “threaded”, in order, onto the wire, and are pushed along the wire.

At each step, a score is calculated based on which residues are adjacent in space, which residues are buried, etc.

Repeat this process for each different known fold.

A high score indicates that the sequence is compatible with that fold.