Similarity Trees for Zinc Finger Protein Design

Dagmar Steffen **Dusan Ihracky** **Lothar Gierl**
dagmar.steffen@medizin.uni-rostock.de

Institut für Medizinische Informatik und Biometrie, Universität Rostock,
Rembrandtstr. 16/17, 18055 Rostock, Germany

Keywords: zinc finger domains, similarity, similarity trees

1 Introduction

Zinc finger domains are protein structures first identified in the Xenopus transcription factor TFIIIA. A zinc finger is composed of 20 to 30 amino-acid residues. There are two residues at both extremities of the domain which are involved in the tetrahedral coordination of a zinc atom. It has been proposed that such a domain interacts with the DNA. Many classes of zinc fingers are characterized according to the number and positions of the histidine and cysteine involved in the zinc coordination.

All biological disciplines are united by the idea that species share a common history. The use of similarity trees has a valued means in many biological problems, such as sequence alignments, protein structure and function prediction etc.

We have developed an algorithm to firstly isolate the zinc finger domains from proteins. Using the Treeprogram we generate an evolutionary tree. The architecture of the tree is based on the context-specific definition of different levels of prototyps (zinc finger families).

2 Method and Results

The comparison of sequences is based on the idea of an alignment, which defines the relation between sequences on an element to element basis. Different similarity matrices, e.g. PAM, BLOSUM, DOT and GCM matrices are used for calculating these alignments.

We first collected all available data about proteins containing zinc finger domains from databases in the internet (SwissProt, MIPS and other). Subsequently, we focused on extracting the zinc finger domains of different classes from the available protein data. These domains built up our database for further investigations.

Using specific patterns (Table 1), the domains were divided into subgroups. The domains of each subgroup now were sorted according to the similarity requirement and then arranged in a similarity tree.

Our similarity matrix contains information needed to construct the similarity tree. For the construction of the tree the following assumptions were made according to Felsenstein [1]:

- a subgroup contains n amino acid sequences, each of the same length m,
- neither insertions nor deletions occurred within a subgroup.

The root is the sequence with most direct successors. A node j is a direct successor of a node i if its number of identical amino acids is equal to $m - 1$. By iterating the search for direct successors, all sequences of a subgroup are sorted according to a descending similarity.

With our above described algorithm similarity trees for all subgroups of our database were generated. The system output is an interactive map of the trees.

Table 1: Special patterns for subgroups.

CXXCXXXXXXXXXXXXHXXH	C2H2 2,12,2
CXXCXXXXXXXXXXXXHXXXH	C2H2 2,12,3
CXXCXXXXXXXXXXXXHXXXXH	C2H2 2,12,4
CXXCXXXXXXXXXXXXHXXXXXH	C2H2 2,12,5
CXXXCXXXXXXXXXXXXHXXH	C2H2 3,12,2
CXXXCXXXXXXXXXXXXHXXXH	C2H2 3,12,3
CXXXCXXXXXXXXXXXXHXXXXH	C2H2 3,12,4
CXXXCXXXXXXXXXXXXHXXXXXH	C2H2 3,12,5
CXXXXCXXXXXXXXXXXXHXXH	C2H2 4,12,2
CXXXXCXXXXXXXXXXXXHXXXH	C2H2 4,12,3
CXXXXCXXXXXXXXXXXXHXXXXH	C2H2 4,12,4
CXXXXCXXXXXXXXXXXXHXXXXXH	C2H2 4,12,5
CXXXXXCXXXXXXXXXXXXHXXH	C2H2 5,12,2
CXXXXXCXXXXXXXXXXXXHXXXH	C2H2 5,12,3
CXXXXXCXXXXXXXXXXXXHXXXXH	C2H2 5,12,4
CXXXXXCXXXXXXXXXXXXHXXXXXH	C2H2 5,12,5
CXXCXXXXXXXXXXXXHXXC	C2HC 2,12,2
CXXCXXXXXXXXXXXXHXXXC	C2HC 2,12,3
CXXCXXXXXXXXXXXXXHXXXXXC	C2HC 2,13,5
CXXCXXXXHXXXXC	C2HC 2,4,4
CXXXXXXXCXXXXXXXXHXXXXC	C2HC 7,8,4
CXXCXXXXXXXXXXXXXCXXC	C4 2,13,2
CXXXXXCXXXXXXXXXXXXCXXXXC	C4 5,12,4

The node contain the amino acid sequence of the domain and two numbers (Fig. 1). These numbers are only used for an easier reading of the tree and represent the preceding and succeeding sequences respectively. Each node can be clicked: a window will then show background information about the corresponding protein.

Our system represents a new tool for protein designers. The amino acid sequences are sorted according to a descending similarity and the user can read this similarity easily from the tree.

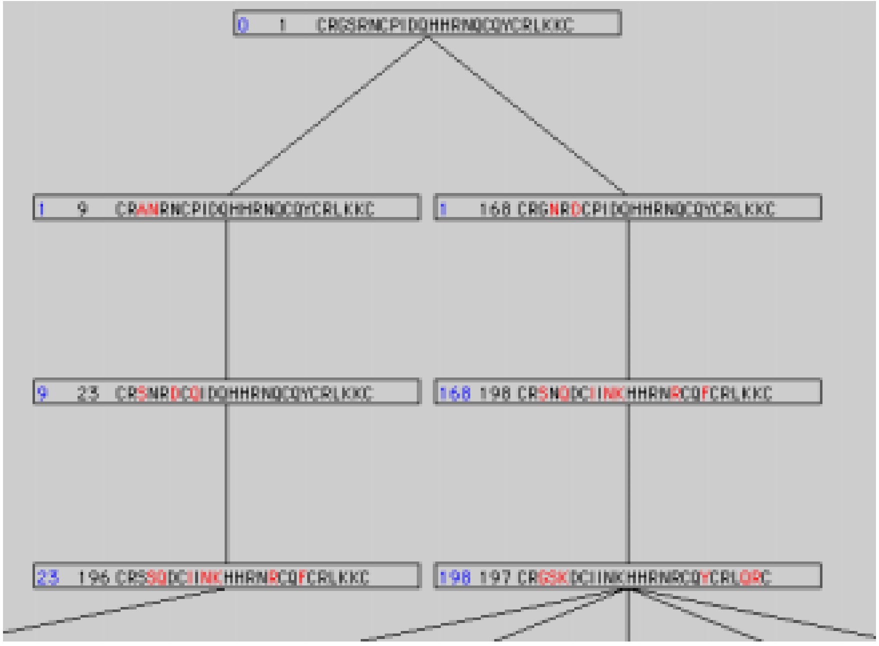
0 1 CRGSRNCPIDQHHRNQCQYCRLKKC
1 9 CRANRNCPIDQHHRNQCQYCRLKKC
1 168 CRGNRDCPIDQHHRNQCQYCRLKKC
9 23 CRSNRDCQIDQHHRNQCQYCRLKKC
168 198 CRSNQDCIINKHHRNRCQFCRLKKC
23 196 CRSSQDCIINKHHRNRCQFCRLKKC
198 197 CRGSKDCIINKHHRNRCQYCRLQRC

Figure 1: Example of a partial generated tree (C4-type zinc finger).

References

[1] Felsenstein, J., Evolutionary trees from DNA sequences, *J. Mol. Evol.*, 17(6):368–376, 1981.