Relation patterns and their Automatic Discovery in Biosequences

Alvis Brazma (abra@cclu.lv)

Institute of Mathematics and Computer Science, University of Latvia

<u>Inge Jonassen</u>, Ingvar Eidhammer ([inge,ingvar]@ii.uib.no)

Department of Informatics, University of Bergen, Norway

Esko Ukkonen (ukkonen@cs.helsinki.fi)

Department of Computer Science, University of Helsinki, Finland

Introduction

Several authors have documented the presence of correlations in protein sequences (Klinger and Brutlag 1994; Clarke 1995). These are believed to be caused by contact potentials and compensating changes.

We wanted to investigate whether it is possible to express this kind of regularities using PROSITE-like patterns, because this would give a very simple and easily interpreted representation.

The database PROSITE uses a special notation (language) for writing deterministic patterns. We decided to extend this language in order to allow for expression of dependencies between residues. We call patterns written in this extended language ***relation patterns*.**

- For pairs of ambiguous positions we can restrict the pairs of amino-acids to be accepted in a match, either by listing the allowed pairs or by listing the pairs that are forbidden.

Correlation

Two residues are correlated if the occuring amino-acids vary together.

```
LVSIGAEEIVDGNVKMTLGMIWTII
LVSIGAEEIVDGNVKMTLGMIWTII
LVSIGAEEIVDGNLKMTLGMIWTII
LVSIGAEEIVDGNVKMTLGMIWTII
LVSIGAEEIVDGNVKMTLGMIWTII
LVGIGAEELVDKNLKMTLGMIWTII
LVSIGAEEIVDGNLKMTLGMIWTII
LVSIDSKAIVDGNLKLILGLIWTLI
LVNIGSSDIVDGNHKLTLGLIWNII
LVNIGSTDIVDGNHKLTLGLIWNII
LVNIGSTDIVDGNHKLTLGLIWNII
VVNIGSQDLQEGKPHLVLGLLWQII
```

Relation pattern describing this alignment:

$\texttt{[ILMV]-x-[GNS]-[ILMV]-[ADEGH]-[AS]}_1\texttt{-x-[ADEG]-[ILMV]-x-[ADEG]-x(4)-[ILMV]}_2\texttt{-x-L-}$
$\texttt{[AGS]-[ILMV]-[ILMV]-W-x-[ILMV]-[ILMV]; (1,2,\{}\,A \leftrightarrow L\texttt{ , }S \leftrightarrow M\,\texttt{\})}$

A sequence matches this pattern if it matches the PROSITE pattern part, and additionally if `[AS]` is matched by A, then `[ILMV]` should be matched by L, or if `[AS]` is matched by S then `[ILMV]` should be matched by M.

Fitness measure

From the minimum description length (MDL) principle (Rissanen 78) it follows that the fitness of a pattern *P* with respect to a set of sequences *S* can be expressed as

minimum description length for *S*

minus

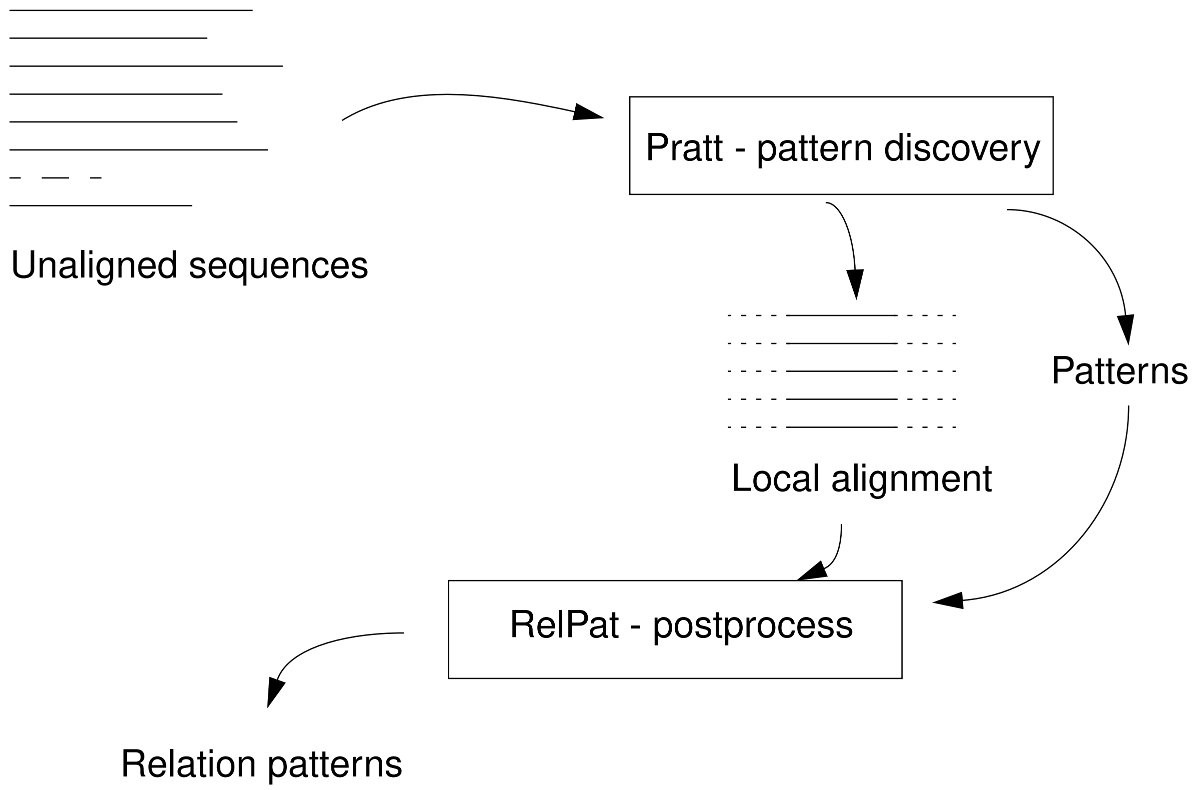
minimum description length for *P* and *S* coded using *P*

Brazma et al reported at ISMB96 a fitness measure for unions of PROSITE type patterns derived using the MDL principle. The fitness measure is generalised to a measure for relation patterns:

Relations added to the pattern makes the pattern description longer, but some bits may be saved when coding the sequences using the pattern if the relation is non-trivial (not all-to-all).

If adding a relation to a pattern increases its fitness it can (under certain conditions) be shown that this means the two related columns have mutual information > 0.

Discovery of relation patterns



Pratt

Pratt (Jonassen et al 1995; Jonassen 1996) is a pattern discovery tool that

- allows the user to specify a class of patterns to be found, by giving the amount of ambiguity, length and flexibility of wildcard regions etc., and
- then finds the highest scoring patterns in this class that matches at least a user-given minimum number of sequences.

Pratt allows for flexible gaps and ambiguous pattern positions. For example, it is able to find the pattern

```
C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H
```

conserved in a family of zinc finger sequences (ZINC_FINGER_C2H2 in PROSITE).

Experiment with the PROSITE database

We used the PROSITE families to test our algorithms. We did two experiments:

1. using the patterns given in the database
2. using Pratt to find patterns from the unaligned sequences belonging to each family

Next the patterns were extended to relation patterns when this increased their fitness.

tables go here

Example:

The sequences in the 1433_1 family from PROSITE were input to Pratt which discovered (amongst other) the pattern

```
E-x-[DR]-[NST]-L-x(1,3)-S-x(0,2)-Y-K-[DN]-[SV]-x-[GL]-
x(2)-[QR]-[AL]-x(2)-[DR]-x-[IL]-[ST]-x(2)-[ET]-[QS]-[DEK]
```

matching all 42 sequences.

RelPat analyses the alignment of the matching segments and finds that it can be extended to the relation pattern:

$\texttt{E-x-[DR]-[NST]-L-x(1,3)-S-x(0,2)-Y-K-[DN]-[SV]-x-[GL]-}$
$\texttt{x(2)-[QR]-[AL]}_1\texttt{-x(2)-[DR]-x-[IL]-[ST]-x(2)-[ET]-[QS]}_2\texttt{-}$
$\texttt{[DEK]; (1,2,}\{A \leftrightarrow Q,\ L \leftrightarrow S\}\texttt{)}$

which matches 38 of the sequences.

Note that we allow for *exceptions*: the relation patterns need not match all the sequences.

Correlations in the Homeodomain family

Clarke (1995) analysed the family of homeodomain sequences, and used a measure of mutual information in order to find correlated residues. The most significant correlation reported was one between residues 19 and 30. We analysed the same set of sequences using our method to see if we could automatically discover this correlation.

alignment 1enh
from HSSP

Pattern

Analyse pairs of columns to find relations to be added to the pattern to achieve maximum fitness

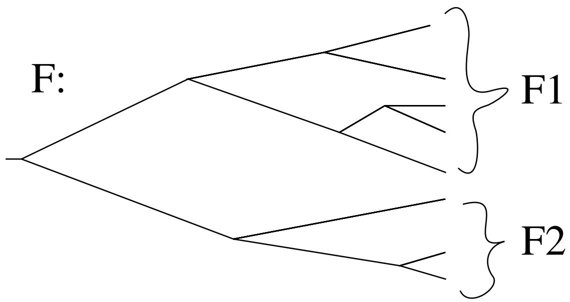
With a variety of parameter values, the pair 19,30 was always among the ones giving the highest fitness

Example of the relations found for this pair: $\{A \leftrightarrow E, A \leftrightarrow T, E \leftrightarrow A, E \leftrightarrow K, A \leftrightarrow R, K \leftrightarrow E, R \leftrightarrow E, Y \leftrightarrow D\}$ which is very close to the data given by Clarke in his table 2. Using charge indexing for the relation, we find the relation $\{pos \leftrightarrow neg, neg \leftrightarrow pos\}$ for the pair 19,30.

Relation patterns and subfamilies

Our motivation was to make a method for finding correlations in protein sequences that would be caused by structural or functional constraints. Based on the results we have obtained so far, it seems that the highest scoring relation patterns found using our method in many cases define subfamily relationships. A method is needed to automatically detect this kind of situation.

For instance, if the proteins are related in this way,



The complete family may share one pattern P, and internally F1 and F2 may share strong patterns P1 and P2 which are specialisations of P. A relation pattern P;R may be found for F. We did one experiment where we for each relation pattern P;R analysed the subfamilies defined by the relation trying to find one pattern for each subfamily. In many cases the patterns found for each subfamily was much stronger than that for the complete family.

Conclusions:

Our experiments have shown that we can identify the correlations reported by other authors as resulting from functinal constraints of the residues. This has demonstrated the applicability of our pattern language as well as the fact that our method for automatic discovery of such patterns is able to handle the “real world” cases. We have also shown that there are many families in PROSITE that can be described by relation patterns in a sensible way. Since the relation pattern language is an extension of the language used in the PROSITE database, and like PROSITE pattern language is very simple and transparent means for describing the sequence properties, it may be useful in some cases to describe PROSITE families by relation patterns.

The experiments also show that there is a number of problems still to be solved to make an autmatic method for finding pairs of residues related by compenstating mutations. The reason for this is that there are also other possible explanation for relations giving compresssion (for instance subfamilies and skewed amino-acid ditributions), and there is a need for methods which can decide between these cases. This problem seems to be common also to other methods for finding correlations between residues. One possibility is to use a phylogenetic tree in this analysis. Another alternative is to use the MDL principle to approach this problem - by defining alternative description schemes, each corresponding to one explanation model, and evaluting these in a similar way to what we did for the relation patterns and unions of patterns in this work.

WWW Pointers:

More details of experiments and experimental results will be given on the page http://www.ii.uib.no/~inge/papers/relation/

Information on Pratt and downloadable source code is found on http://www.ii.uib.no/~inge/Pratt.html

Related papers are found on http://www.ii.uib.no/~inge/papers/

References:

Bairoch A, Bucher P, Hofmann K. 1996. The PROSITE database, its status in 1995. Nucleic Acids Res. 24 (1), 189-196.

Brazma A, Jonassen I, Ukkonen E, Vilo J. 1996. Discovering patterns and subfamilies in biosequences. Proc. ISMB96, 34-43.

Clarke N D. 1995. Covariantion of resicues in the homeodomain sequence family. Protein Science 4, 2269-2278.

Jonassen I, Collins J F, Higgins D G. 1995. Finding flexible patterns in unaligned protein sequences. Protein Science 4 (8), 1587-1595.

Jonassen I. 1996. Efficient discovery of conserved patterns using a pattern graph. Dept. of Informatics, University of Bergen. Report 118.

Klinger T M, Brulag D L. 1994. Discovering Structural Correlations in alpha-Helices. Protein Science 3, 1847-1857.

Rissanen J. 1978. Modeling by the shortest data description. Automatica-J.IFAC 14, 465-471.

Summary

We have defined a new pattern language that extends that used in PROSITE to allow for description of dependencies between residues -- patterns in this language are called relation patterns.

A fitness measure for how good a relation pattern fits a set of sequences, was defined using the minimum description length (MDL) principle.

A heuristic algorithm for finding relation patterns for a set of unaligned sequences was developed.

Experiments were done to see if extending the patterns in PROSITE to include relations increases the fitness of the patterns. A test for increased selectivity may be done.

An earlier reported strong correlation in the faimly of homeodomain proteins was autmoatically rediscovered.

Future research could include analysis that is more sensitive to subfamily relationships and sets of very similar sequences which bias the patterns and their scores today.