T U M

I N S T I T U T F Ü R I N F O R M A T I K

Experimental Prototype for automated Beta Sheet Pairing based on NOESY spectra

Michael Riss, Murray Coles, Horst Kessler, Alois Knoll

TUM-I0506
Mai 05

T E C H N I S C H E U N I V E R S I T Ä T M Ü N C H E N

TUM-INFO-05-I0506-0/1.-FI

Druck: Institut für Informatik der
Technischen Universität München

1 Abstract

A test prototype has been developed in cooperation with Institute of Chemistry (Prof. Kessler) and the Max-Planck-Institute Tübingen (Dr. Murray Coles) to determinate the alignment of beta-strands within the beta-sheets of a protein from NMR data. Alignment of beta-strands can usually be performed by a human expert, but the process is time consuming and errors are possible. The prototype was intended to automate the tasks involved and avoid errors by carrying out a thorough search of possible alignments. The prototype takes chemical shifts and NOESY peak-lists as input. The location of individual beta-strands must be determined in advance and is also supplied as input. As output it produces several charts plotting the different beta-sheet alignments against their consistency with the observed NOESY peaks. We conclude that the alignment of beta-sheets purely based on chemical shifts and NOESY peaks is possible, but the signal-to-noise ratio is surprisingly small.

2 Background

The focus of our cooperation lies on assisting the human expert in manual analysis of NMR spectra and to automate subtasks. It has become clear that a computer must use a different approache to NMR analysis than a human. While humans are good at integrating small amounts of data from many information sources, a computer is best when processing large amounts of data from few information sources. In NMR analysis the human expert in fact uses many information sources: the NMR spectra themselves, genetic knowledge, knowledge about similar proteins and something called "expert knowledge". The later is experience as to when the other information sources provide correct data and in when they do not. Currently it is not possible to generally teach a computer such "expert knowledge". To do so, considerable advances in the field of machine learning and AI are necessary. For our attempts to automate NMR analysis steps we therefore have to rely on clever architectures, NMR-specific pseudo expert knowledge (statistical data) and the ability of the computer to process far more data than a human expert.
This prototype was written to test whether it is possible to align the beta-strands within the beta-sheets of a protein solely based on measured NOESY data. This task in normally performed "by hand" by a human expert, but the process is time consuming and subject to potential errors, as no thorough search of all pairing possibilities is made.

3 Principle

The idea of the prototype is simple. An assumption on the pairing of two beta-strands is made and the expected pattern of cross-peaks in the spectra for this pairing is generated. Comparing the expected pattern with the peak-list of the measured spectra, the correctness of the assumption can be tested. The better the correspondence of the predicted peaks with the real peaks, the higher the score of the pairing.

4 Example - KdpBN

The prototype was tested with a test protein, KdpBN [3]. This was done in two steps, the first step was to generate a synthetic peak-list from the known structure of KdpBN and use it as input for our prototype. An example plot can be seen in figure 1. In the second step we used peak-lists from real spectra

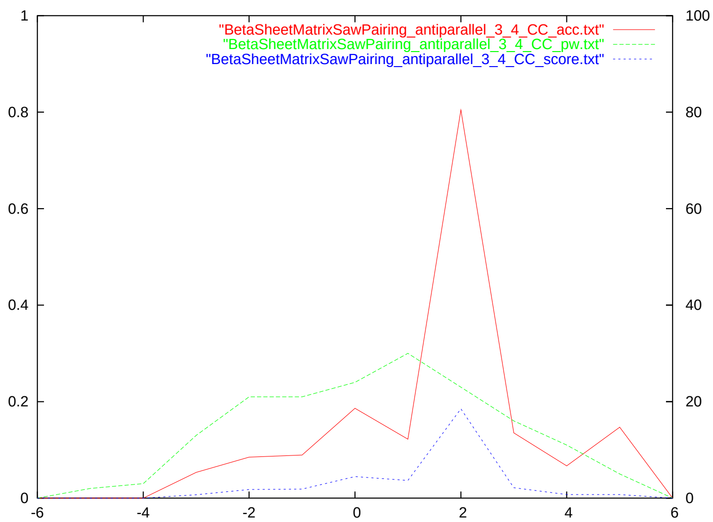

Figure 1: Example of a Beta-sheet-Alignment-Plot with synthetic data

and repeated the analysis, the corresponding plot is shown in figure 2. The first thing that is apparent from the plots is the maximum at offset +2; the program considers this offset the most likely for a pairing between the two beta-strands (and in fact this is the right offset).
The different colored lines have the following meaning:

- RED - the ratio between found peaks and expected peaks (left scale)
- GREEN - the number of processed peaks for this offset (right scale)
- BLUE - the score, computed as the product of RED and GREEN

The RED and the BLUE lines are the more significant. The RED line marks offsets with a very strong signal, but is numerically unstable when the number of processed peaks is low. In these cases the BLUE line - which Visualizes the product between RED and GREEN - is more reliable. Typically high peaks of the RED line are found first and then verified with the BLUE line. The following printout shows which residues were used to compute the scores for the different offsets in figure 1. For the maximum at offset +2 the residue 397 was paired with residue 431, 396 with 432, 395 with 433, 394 with 434 and 393 with 435, i.e. an anti-parallel beta-sheet.
In an anti-parallel beta-sheet there are two alignments possible for each residue pair. Either the two H^α or the two H^N face each other. Here 397 and 431 are in the H^α-H^α alignment (CC), 396 and 432 in H^N-H^N alignment (NN), etc. For a full overview of all plots from KdpBN see figures 7 - 10 in the Appendix.

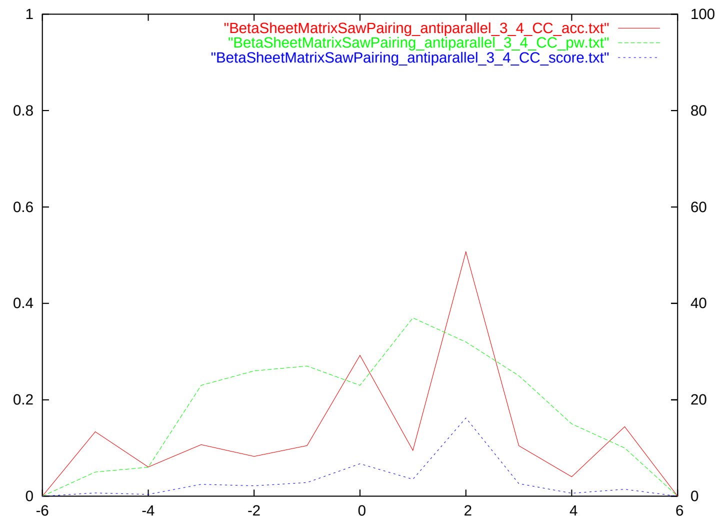

Figure 2: Example of a Beta-sheet-Alignment-Plot with real data

```
-6  CC: 391/429
-5  CC: 392/429 NN: 391/430
-4  CC: 393/429 NN: 392/430 CC: 391/431
-3  CC: 394/429 NN: 393/430 CC: 392/431 NN: 391/432
-2  CC: 395/429 NN: 394/430 CC: 393/431 NN: 392/432 CC: 391/433
-1  CC: 396/429 NN: 395/430 CC: 394/431 NN: 393/432 CC: 392/433 NN: 391/434
 0  CC: 397/429 NN: 396/430 CC: 395/431 NN: 394/432 CC: 393/433 NN: 392/434 CC: 391/435
+1  CC: 397/430 NN: 396/431 CC: 395/432 NN: 394/433 CC: 393/434 NN: 392/435
+2  CC: 397/431 NN: 396/432 CC: 395/433 NN: 394/434 CC: 393/435
+3  CC: 397/432 NN: 396/433 CC: 395/434 NN: 394/435
+4  CC: 397/433 NN: 396/434 CC: 395/435
+5  CC: 397/434 NN: 396/435
+6  CC: 397/435
```

Figure 3: Contact traces for each offset

5 Implementation

The prototype was implemented in a mixture of C and C++. As input it uses:

- a list with the residue sequences which form beta-strands
- a list with the chemical shifts
- several peak-lists: currently we use these 3D spectra: CCH, CNH, HCH, HNH and NNH [2]
- several parameter files describing the distortion zones in the spectra (diagonal and H_2O lines), peaks in these areas are ignored

As output the prototype produces several score files in a format suitable for input in gnuplot. Each score file covers all possible pairings between two beta-strands. Besides the score files the program produces description files that explain the offsets in more detail, as in figure 3.

5.1 Matrix

The central data structures for beta-strand pairing are matrices. For each pair of beta strands a matrix is created. Its size is determined by the lengths of the beta strands, e.g. if the beta strands have a length of 4 and 6 residues, the matrix has the dimensionality of 4x6. Each cell in the matrix corresponds to the combination of two residues of the two strands.
In the first pass a score for each cell is computed. This is done by predicting the peaks which would result if the two residues make contact, and comparing them with the peaks in the given peak-list. The more of the predicted peaks are found in the peak-list, the higher the score for this cell.
In the second pass we sum up diagonal traces in the matrix. Figure 4 displays the principle for the anti-diagonal case. Each trace corresponds to a specific

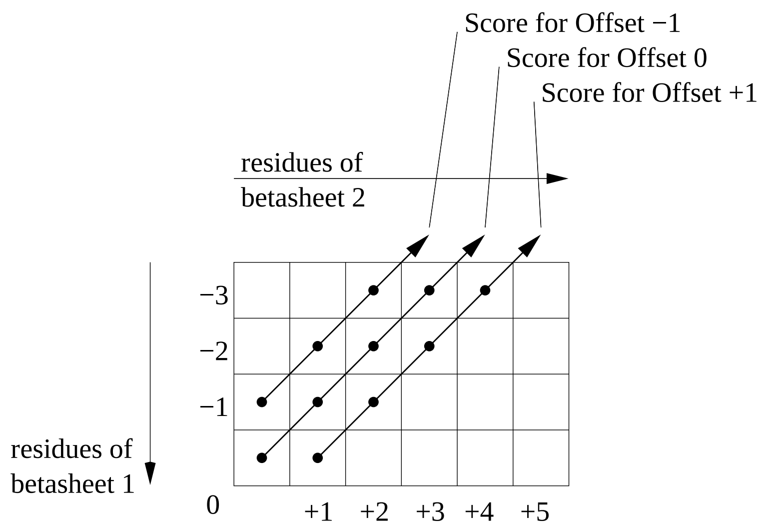

Figure 4: Residue matrix and anti-diagonal traces

beta-strand alignment. The computed score for the trace is the final score for the corresponding alignment offset.

5.2 Scores

A score consists of a data structure with two fields: accuracy and peak-weight. Accuracy tells us how well the predicted alignment matches the real data in the peak lists; i.e. it is a measure of quality. Peak-weight tells us how many peaks were used to compute the Accuracy parameter; it is a measure of quantity. When computing a high-level score, such as a cell score or a trace score, several low-level scores have to be "summed". This is not done as a simple sum; rather a high-level score A is calculated from the two low-level scores B and C using the formula:

$$\begin{aligned} A.acc &= \frac{B.acc * B.peakw + C.acc + C.peakw}{B.peakw + C.peakw} \\ A.peakw &= B.peakw + C.peakw \end{aligned}$$

5.3 Which peaks to use?

Theoretically all possible atom-atom combinations between two residues could be used in the analysis, and a search for the corresponding peaks in the peak lists performed. In practice, however, only peaks between certain atoms of the residues involved are significant. For our beta-sheet search we look for the following combinations:

- H^N - H^N, directly opposite
- H^α - H^α, directly opposite
- H^β - H^β, directly opposite
- H^N - H^β, directly opposite
- H^N - H^α, diagonally opposite

Each peak search translates into a score. When searching for peaks tolerance margins are applied, by default these are 0.03 ppm in H-dimensions and 0.3 ppm in C- and N-dimensions. The diagram for accuracy assignment in the H-dimension is displayed in figure 5. If the peak is found exactly at the expected location the accuracy is 1.0. If the peak is outside the tolerance margins then the accuracy is 0.0. "Near misses" get accuracy scores from 1.0 to 0.5 according to a gauss-shaped function. The peak-weight part of the score is currently set to

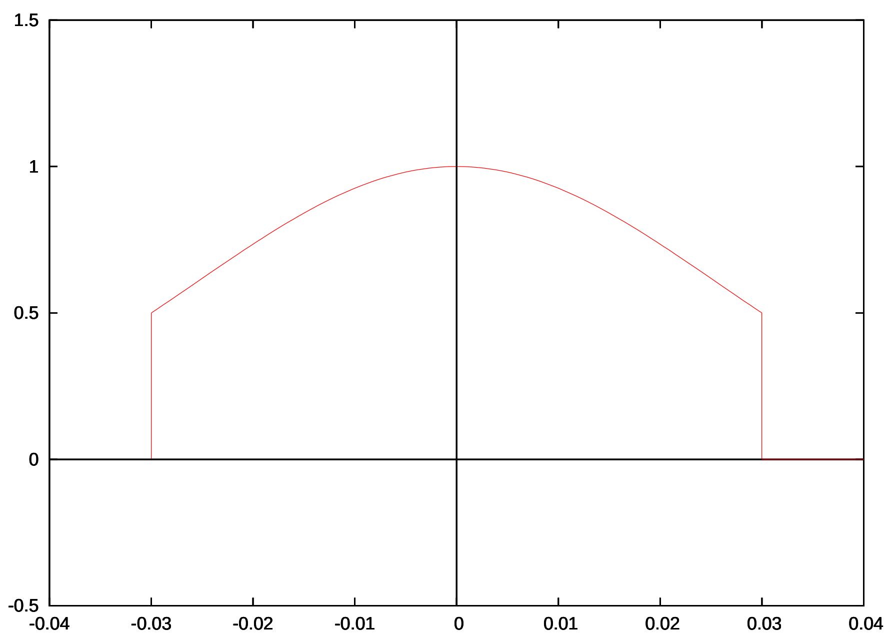

Figure 5: accuracy dependent on ppm drift, 0.03 ppm margins

1.0 for each peak. Peaks expected in the distortion zones (diagonals, H_2O-Lines) are ignored.

5.4 Conformation ranking

The scores reflect the probability of the different pairings. Ideally the correct pairs will have the highest scores. This was tested for KdpBN by sorting all scores and examining the pairings with the highest scores. In this case the five highest scoring pairings were in fact the correct ones. Figure 6 shows a plot of the ten highest score values. The first five scores correspond to the correct pairs,

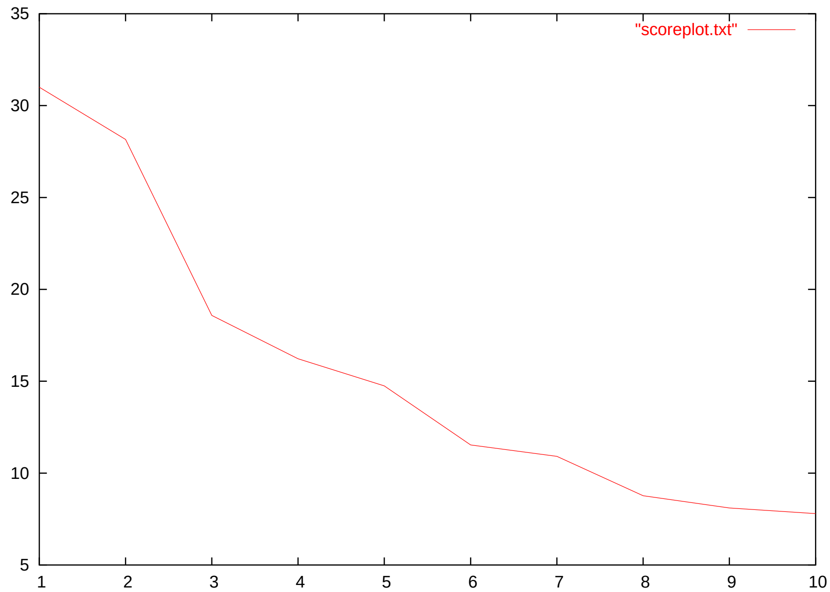

Figure 6: The 10 highest scores for KdpBN, only the first 5 are correct

the next five scores are scores for wrong pairs due to noise. The figure shows that there is no significant score drop after the five correct pairings, so it's not possible to use the score to distinguish between correct pairs and false pairs, this still has to be done by a human expert. Additionally it shows that the signal-noise-ratio is just sufficient. Considering that the spectra of the KdpBN test protein were very good, It is possible that wrong pairings might get higher scores than correct ones when using spectra of lower quality.

5.5 Problem: Beta Bulges

While our prototype detects the correct beta-strand pairings in most cases, it can't cope with beta-bulges. Due to our fixed scheme of summing up scores along diagonal traces the algorithm is unable to adapt to the changing offset at a bulge. One simple but inefficient solution could be to try out all bulge possibilities at each position and pick the one which fits best. But apart from the vastly increased computational effort this approach might have, stability problems due to its increased sensitivity to noise could arise. This problem is also minimised by the fact that beta-strands containing bulges will often be recognized as two separate strands, which will then be aligned individually into their correct locations.

6 Conclusion

The prototype proved that it is possible to determine the right beta-sheet alignment from NOESY-Spectra under good conditions. Despite the good quality spectra available for KdpBN, the signal-noise-ratio is surprisingly small. This is all the more surprising when it is considered that a human expert can usually successfully perform this task, although much more slowly and without the thorough search of possible pairings performed by the program. Either the human expert is using considerable knowledge from other sources, or the prototype is failing to make full use of the available data. One such extra information source is the complementary nature of the spectra; a peak in one spectrum should have equivalent peaks in other spectra. The prototype does not make use of such information in its present form. Inclusion of this type of data, plus improvements to the scoring function should improve the signal-to-noise ratio achieved. Efforts in this direction are currently underway. A user interface to aid the expert user in interpreting the data is also planned.

A KdpBN - all plots

On the following pages are the full plates of plots for the protein KdpBN. (figures 7 - 10). The first two plates show the plots for synthetic data and the next two plates the plots for real data.
The plates are arranged in the order of the beta-sheets used. The first plot shows the pairing of beta-sheet 0 with 1, the second shows 0 with 2, etc.
The correct beta-sheet pairings for KdpBN are 0-5, 1-2, 2-3, 3-4 and 4-5.

References

[1] Tammo Diercks, Murray Coles & Horst Kessler:
An efficient strategy for assignment of cross-peaks in 3D heteronuclear NOESY experiments
Journal of Biomolecular NMR, **15**: 177-180, 1999

[2] Melina Haupt, Mark Bramkamp, Murray Coles, K. Altendorf & Horst Kessler:
Inter-domain motions of the N-domain of the KdpFABC complex, a P-type ATPase, are not driven by ATP-induced conformational changes Journal of Molecular Biology, **342**: 1547-1558, 2004

[3] M. Haupt, M. Bramkamp, M. Coles, K. Altendorf, H. Kessler:
The Solution Structure Of The Nucleotide Binding Domain Of Kdpb
Protein Database **1X6K**, 11-Aug-2004

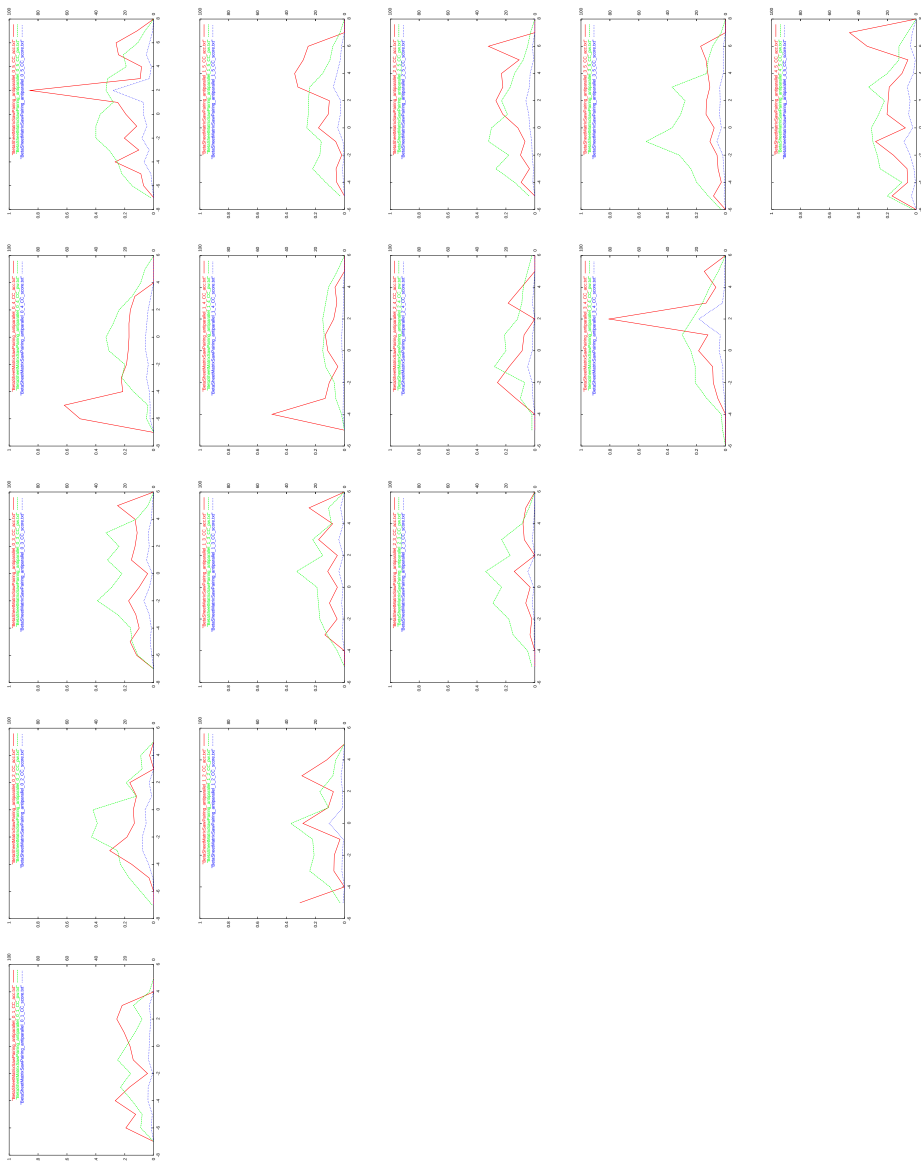

Figure 7: synthetic data, all matches starting in CC conformation

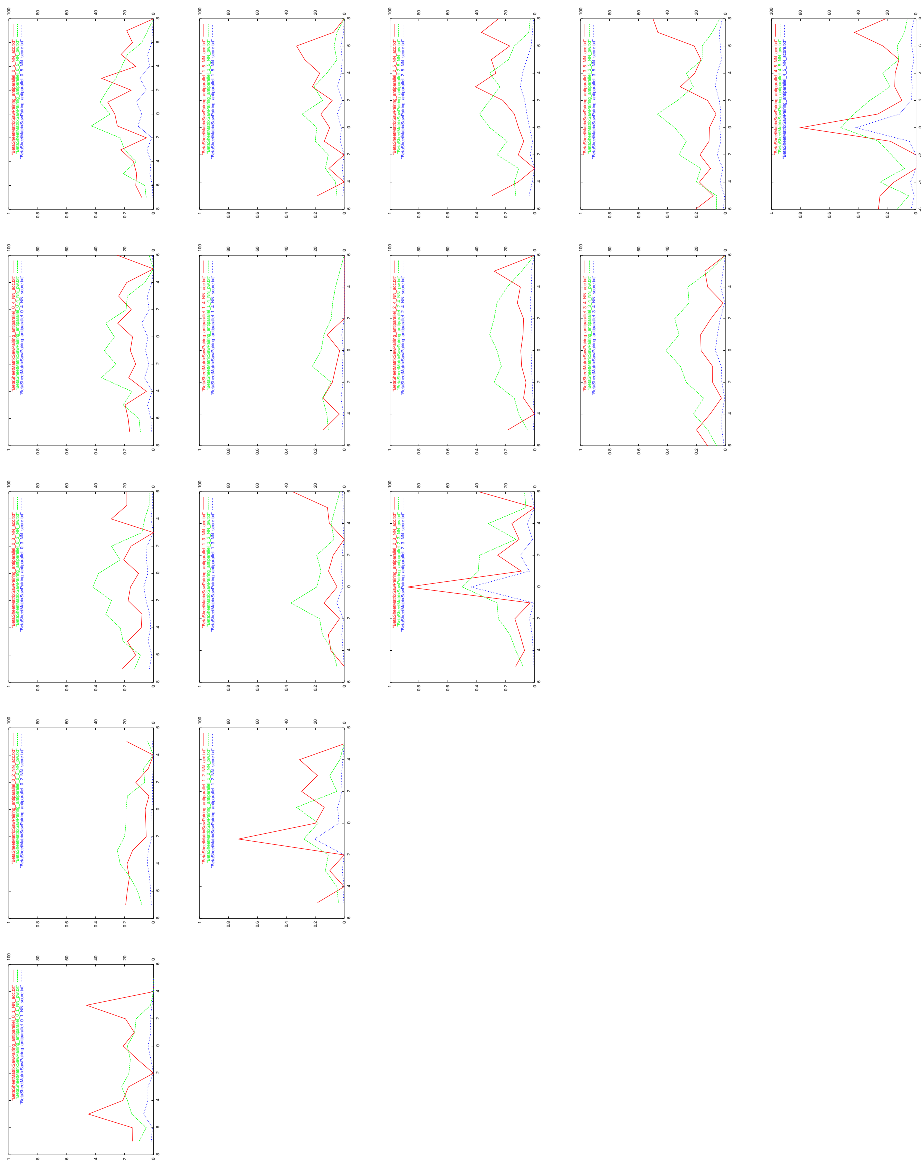

Figure 8: synthetic data, all matches starting in NN conformation

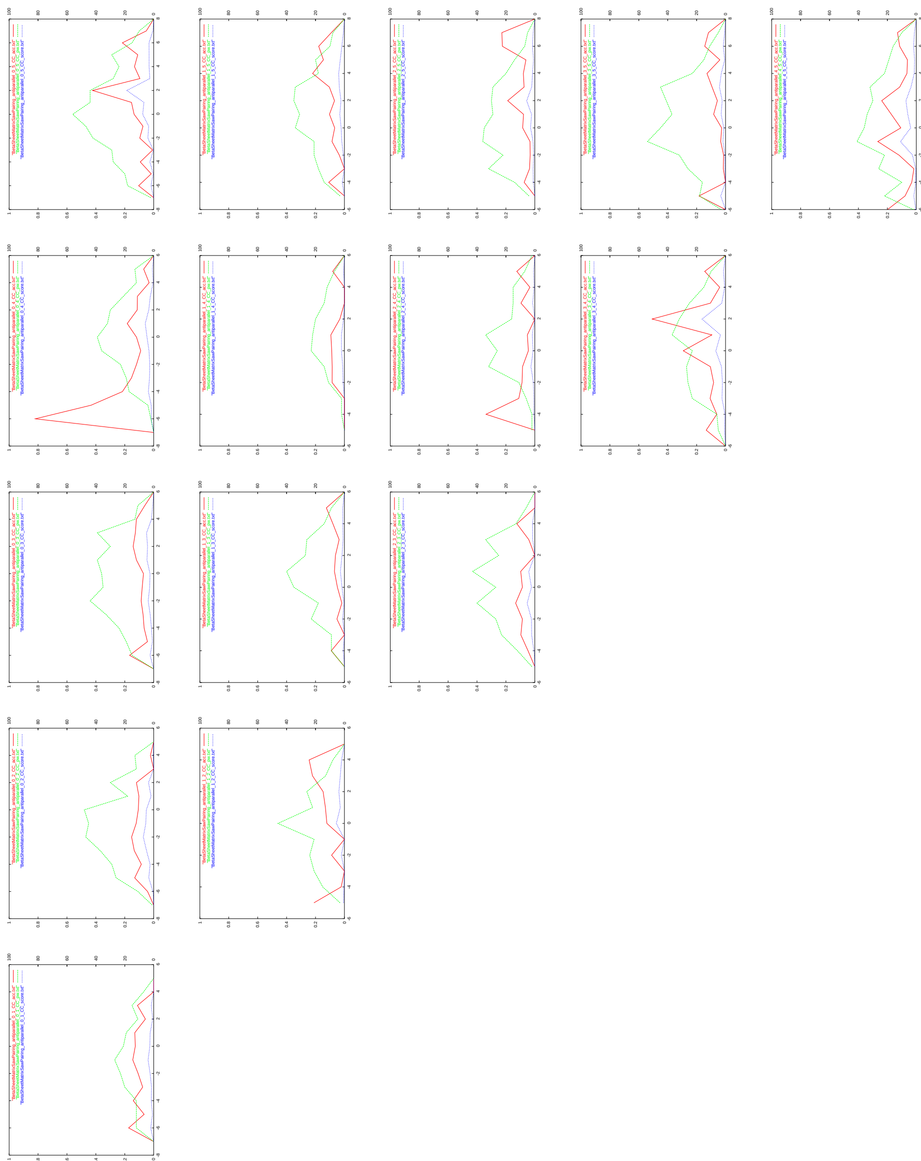

Figure 9: real data, all matches starting in CC conformation

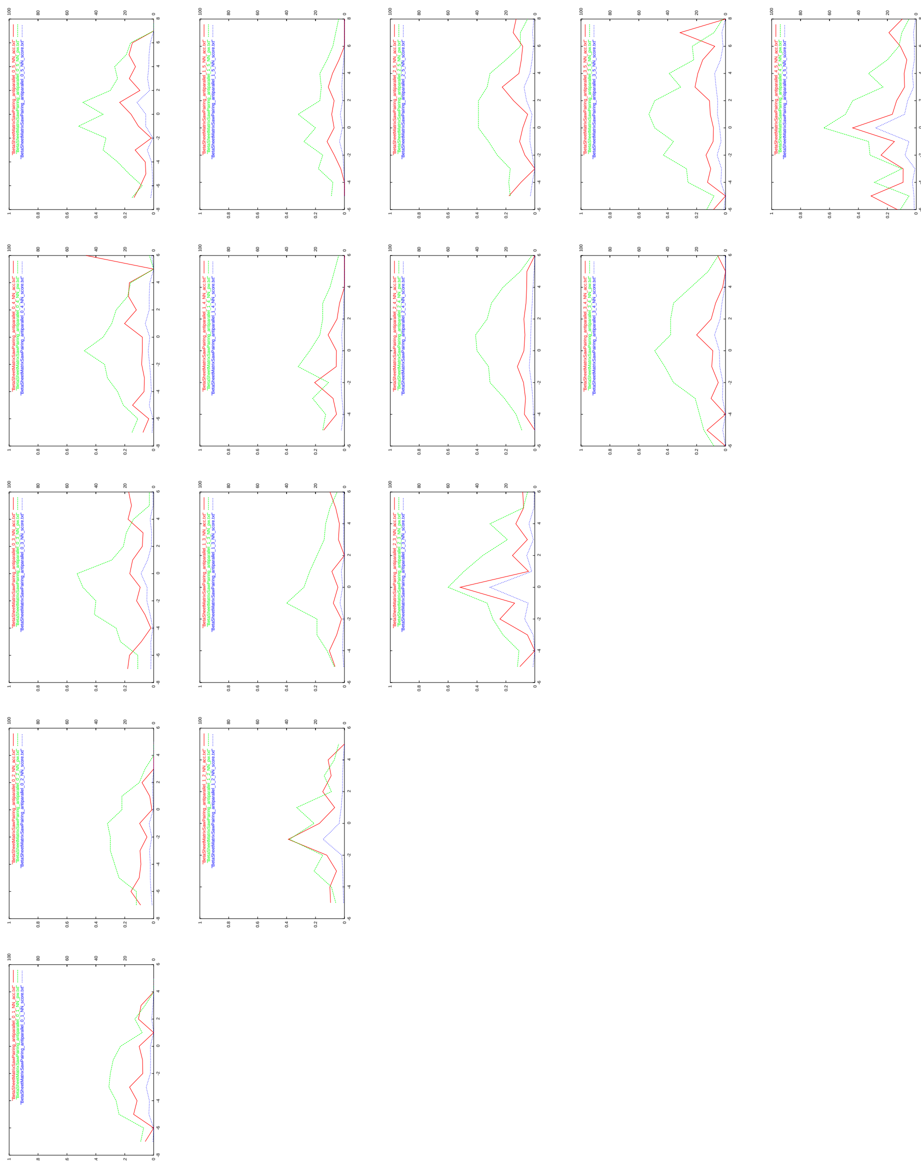

Figure 10: real data, all matches starting in NN conformation